A New Paradigm in Brain Inspired Lifelong Machine Learning for Data Intensive Environments

B. K. Uvindu Rashmika Nawaratne
BSc. Eng. (Honours) (University of Moratuwa)

Research Centre for Data Analytics and Cognition
La Trobe Business School
College of Arts, Social Sciences and Commerce
La Trobe University, Victoria, Australia

A thesis submitted in total fulfilment of the requirements for the degree of
Doctor of Philosophy

December 2020

A journey of thousand miles starts with a single first step, and none of these would be possible if not for the two people who taught me that very first step,

To my dear parents,
Gamini and Manori

For always standing by me during the hurdles of life by being my companion,

To my loving wife,
Achini

Statement of Authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgement in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

B. K. Uvnidu Rashmika Nawaratne
07 December 2020

Acknowledgements

It is with my heartfelt gratitude that I appreciate the two pillars of strength to complete this scholarly pursuit. My supervisors, Prof. Damminda Alahakoon and A/Prof. Daswin De Silva.

Being my principal supervisor, Prof. Damminda Alahakoon guided me into venturing into novel avenues in self-organizing learning and artificial intelligence. It was his knowledge, advice, and support that continually inspired me to think innovatively and advance my research in many directions. Thank you, this would not have been possible if not for your constant guidance and encouragement.

My co-supervisor, A/Prof. Daswin De Silva, has been a great support by encouraging me to pursue great heights in research as well as in my academic career. I admire and appreciate your time and efforts in advising me and guiding me throughout my Ph.D. journey especially with conferences, publications and academic career. Thank you very much for believing in me.

Furthermore, during my Ph.D. career, I was fortunate to collaborate with distinguished researchers from diverse backgrounds at La Trobe University. I got the opportunity to work with Prof. Leeanne Carey (Health Sciences), Prof. Michael Kingsley (Sports Sciences) and Dr. Su Ngyen (AI, Energy), where I was able to translate my research outcomes into practical applications and on-going research. This has certainly given me the confidence of using my research in real-time applications as well as to gain experience working with industry and academic collaborators. Thank you for your rewarding opportunities.

During the four years as a research scholar, I met with many colleagues at the Research Centre for Data Analytics and Cognition at the La Trobe Business School. I cherish and appreciate the fond memories of us working together as a team and thank you for being wonderful companions in my research career at La Trobe University.

Further, I would like to pay my gratitude to La Trobe University for supporting my research by providing me with a La Trobe Full Fee Research Scholarship and a La Trobe University Postgraduate Scholarship, which aided me to successfully complete my research as well to be funded for overseas travel for research conferences.

Last but not least, I must applaud the ‘unsung heroes’ of my academic success so far. The privilege of education came through the blood and sweat of taxpayers in my home country. It is my utmost duty to appreciate each hard-earned penny contributed towards my education as an engineer and a scholar. My final salute to all the tax-payers in Sri Lanka.

Abstract

Non-deterministic data intensive environments are characterised by the increasing complexity of structures and strictures that motivate and mandate individuals, communities, systems and societies. Although the human mind comprehends this complexity, its computational disposition exacts an intelligence that is artificial. Adaptations of conventional Artificial Intelligence (AI), which was designed for deterministic settings, typically employs an asymmetric learning paradigm where historic data is used to train AI to solve current and future problems, and is inadequate to address this challenge. The primary limitations can be summarised as training with pre-labelled data, catastrophic forgetting and the stability plasticity dilemma. To address these limitations, a new symmetric development is necessitated that can not only learn from past but able to cater to the challenges presented by non-stationary continuous data streams.

This thesis proposes lifelong learning machines as a paradigm shift from conventional AI. It is inspired by the structure of the human brain, where neocortex and hippocampus complementarily facilitate lifelong learning and preservation of memory, as well as the function of the human brain, where invariant representation, transience, sequential and recurrent memory formation facilitate stability and plasticity of the biological memory. Lifelong learning machines are conceptualised as a cognitive model that effectuates continuous memory acquisition and knowledge augmentation. This model is substantiated by the design and development of two new algorithms for lifelong learning, RTGSOM and LifeNet. The algorithms are demonstrated and evaluated across several non-deterministic data intensive environments, including smart cities and digital health. This empirical evidence base circumstantiates lifelong learning machines in addressing the emerging challenges of non-deterministic data intensive environments.

Table of contents

List of figures

List of tables

Publications originated from the thesis

The concepts, new algorithms and experiments described in this thesis were published in different peer-reviewed journals and conference proceedings.

First author journal articles based on this thesis

1. Nawaratne, R., Kahawala, S., Nguyen, S., & De Silva, D. (2020). A Generative Latent Space Approach for Real-time Road Surveillance in Smart Cities. IEEE Transactions on Industrial Informatics.

2. Nawaratne, R., Adikari, A., Alahakoon, D., De Silva, D., & Chilamkurti, N. (2020). Recurrent Self-Structuring Machine Learning for Video Processing using Multi-Stream Hierarchical Growing Self-Organizing Maps. Multimedia Tools and Applications.

3. Nawaratne, R., Alahakoon, D., De Silva, D., O'Halloran, P. D., Montoye, A. H., Staley, K., ... & Kingsley, M. I. (2020). Deep Learning to Predict Energy Expenditure and Activity Intensity in Free Living Conditions using Wrist-specific Accelerometry. Journal of Sports Sciences, 1-8.

4. Nawaratne, R., Alahakoon, D., De Silva, D., Kumara, H., & Yu, X. (2019). Hierarchical Two-Stream Growing Self-Organizing Maps with Transience for Human Activity Recognition. IEEE Transactions on Industrial Informatics, 16(12), 7756 - 7764.

5. Nawaratne, R., Alahakoon, D., De Silva, D., & Yu, X. (2019). Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance. IEEE Transactions on Industrial Informatics, 16(1), 393-402.

6. Nawaratne, R., Alahakoon, D., De Silva, D., Chhetri, P., & Chilamkurti, N. (2018). Self-evolving intelligent algorithms for facilitating data interoperability in IoT environments. Future Generation Computer Systems, 86, 421-432.

Other journal articles based on this thesis

7. Alahakoon, D., Nawaratne, R., Xu, Y., De Silva, D., Sivarajah, U., & Gupta, B. (2020). Self-building artificial intelligence and machine learning to empower big data analytics in smart cities. Information Systems Frontiers, 1-20.

8. Bandaragoda, T., Adikari, A., Nawaratne, R., Nallaperuma, D., Luhach, A. K., Kempitiya, T., ... & Chilamkurti, N. (2020). Artificial intelligence based commuter behaviour profiling framework using Internet of things for real-time decision-making. Neural Computing and Applications, 1-15.

9. Nallaperuma, D., Nawaratne, R., Bandaragoda, T., Adikari, A., Nguyen, S., Kempitiya, T., ... & Pothuhera, D. (2019). Online incremental machine learning platform for big data-driven smart traffic management. IEEE Transactions on Intelligent Transportation Systems, 20(12), 4679-4690.

10. Hettiarachchi, P., Nawaratne, Alahakoon, D., De Silva, D., & Chilamkurti, N. (2021). Rain Streak Removal for Single Images Using Conditional Generative Adversarial Networks. Applied Sciences, 2021, 11(5), 2214.

11. Gunawardena, P., Amila, O., Sudarshana, H., Nawaratne, R., Luhach, A. K., Alahakoon, D., ... & De Silva, D. (2020). Real-time automated video highlight generation with dual-stream hierarchical growing self-organizing maps. Journal of Real-Time Image Processing, 1-19.

12. Kingsley, M. I., Nawaratne, R., O'Halloran, P. D., Montoye, A. H., Alahakoon, D., De Silva, D., ... & Nicholson, M. (2019). Wrist-specific accelerometry methods for estimating free-living physical activity. Journal of science and medicine in sport, 22(6), 677-683.

Conference papers and pre-print articles

13. Nawaratne, R., Alahakoon, D., De Silva, D., & Yu, X. (2019, July). HT-GSOM: dynamic self-organizing map with transience for human activity recognition. In 2019 IEEE 17th International Conference on Industrial Informatics (INDIN) (Vol. 1, pp. 270-273). IEEE.

14. Nawaratne, R., Bandaragoda, T., Adikari, A., Alahakoon, D., De Silva, D., & Yu, X. (2017). Incremental knowledge acquisition and self-learning for autonomous

video surveillance. In IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society (pp. 4790-4795). IEEE.

15. Adikari, A., Nawaratne, R., De Silva, D., Carey, D., Walsh, A., Baum, C., Davis, S., Donnan, G., Alahakoon, D., & Carey, L.. "Is mild really mild?": Patient profiling using Artificial Intelligence. (Unpublished)

16. Madhavi, I., Chamishka, S., Nawaratne, R., Nanayakkara, V., Alahakoon, D., & De Silva, D. (2020, September). A Deep Learning Approach for Work Related Stress Detection from Audio Streams in Cyber Physical Environments. In 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (Vol. 1, pp. 929-936). IEEE.

17. Gunawardena, P., Sudarshana, H., Amila, O., Nawaratne, R., Alahakoon, D., Perera, A. S., & Chitraranjan, C. (2019, September). Interest-Oriented Video Summarization with Keyframe Extraction. In 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer) (Vol. 250, pp. 1-8). IEEE.

18. Rathnayaka, P., Abeysinghe, S., Samarajeewa, C., Manchanayake, I., Walpola, M. J., Nawaratne, R., ... & Alahakoon, D. (2019). Gated Recurrent Neural Network Approach for Multilabel Emotion Detection in Microblogs. arXiv preprint arXiv:1907.07653.

19. Rathnayaka, P., Jayasundara, V., Nawaratne, R., De Silva, D., Ranasinghe, W., & Alahakoon, D. (2019). Kidney Tumor Detection using Attention based U-Net.

20. Abeysinghe, S., Manchanayake, I., Samarajeewa, C., Rathnayaka, P., Walpola, M. J., Nawaratne, R., ... & Alahakoon, D. (2018, September). Enhancing Decision Making Capacity in Tourism Domain Using Social Media Analytics. In 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer) (pp. 369-375). IEEE.

Chapter 1

Introduction

1.1 Background

We have constructed cities that extend for miles across land and built skyscrapers that rise up to the clouds. Roads have filled up with vehicles, ships and submarines sail and drive across the seas, while airplanes and satellites fill up the atmosphere and beyond. While continuing to pursue domination of the world and its natural environment, humans seem to have set trajectories to conquer even the solar system.

It took over 2 million years from the supposed inception of the human race to reach this point, however, the progress has snowballed since the first industrial revolution of the 18^{th} century. In less than 250 years, we have transformed and elevated from animal-drawn carts to self-driving cars; from messenger pigeons to voice command based online-communication platforms using intelligent applications such as Siri, Alexa or Google Assistant (Stiehler, 2018). Industrial environments have transitioned towards autonomous machinery, cyber-physical systems and energy-efficient layouts. As a result, the urban environments have become densely populated with multi-level buildings, increased vehicular, pedestrian traffic and crowd movements.

In this new digital age, an important development has arisen to revolutionize the earth's long-held human-centric status quo, the transcendence of *Artificial Intelligence* (AI), that is identified as the fourth revolution in an era driven by automation and connectivity (Butt, 2019). The vertical and horizontal expansion of modern cities, assets and area utilisation in both industrial and urban environments have eventuated an exponential increase in the deployment of intelligent mechanisms powered by AI.

During the first half of the 20^{th} century, AI became known to the general population as a scientific fantasy through science fiction in the form of Robots, i.e., machines that resemble human beings and able to replicate certain human movements and functions automatically.

For instance, the concept of AI Robots was introduced with a humanoid robot intent on taking over the titular mega-city by inciting chaos in the movie *Metropolis* in 1927, then the "heartless" Tin man from the 1939 fantasy movie *The Wizard of Oz* (Anyoha, 2017). By the mid 20th century, a generation of scientists, mathematicians and philosophers assimilated their minds on the advancement of AI. One such person, Alan Turing, explored the mathematical possibility of AI, founding a logical framework to use available information as well as reasoning capabilities in order to solve problems and make decisions (Turing, 2004). In 1930s, Alan Turing provided a formalisation to the concepts of computation algorithms with the introduction of *Turing Machine*, that can be considered as an early model of a general purpose computer (Turing, 1936).

Compared to the dawdling pace of the materialization of AI since mid 20th century, the latter part of 20th century to early 21st century saw a revolution in the process of materialization of AI in many facets; communication, transportation, healthcare, finance, education, industrial automation, and day to day human activities. Machine learning is the branch of AI focused on building applications that learn from data and improve their accuracy over time without being programmed to do so, which has excelled in upholding the success of AI in the recent past. In machine learning, machines are 'trained' to find patterns and features in massive amounts of data in order to make decisions and predictions based on new data (Education, 2020). Many organizations have realized the potential of AI and Machine Learning in modern applications and as a result, around 2,000 start-ups globally have ventured into integrating AI and Machine Learning as a core part of their business model (Stiehler, 2018). Furthermore, recent highlights such as Google's AlphaGo (Granter *et al.*, 2017) defeating the Go world champion and Baidu's personal assistant Duer accepting orders at KFC restaurants in China (Tan Min, 2016) demonstrate promising progress of technological disruption in foreseeable future.

The advancement of the Internet of Things (IoT) has further proliferated the rise of AI by enhancing the ability to exploit a granular level of real-time data (Osuwa *et al.*, 2017). Designed to provide a dense degree of connectivity across various devices connected to the Internet and allowing these devices to exchange information with each other, IoT sensors cut across many areas of modern-day living, thus offering the ability to measure, infer and understand environmental indicators across delicate ecology, natural resources and urban environments. For instance, Tesla electric cars are equipped with 21 sensors only to generate an overview perception of the immediate surrounding; 8 video cameras (three forward looking, two side looking, two side-mounted rear looking and one rear), 12 ultrasonic sensors, and 1 front radar. A front facing camera is responsible for Autosteer and in keeping the car in the lane (Heath, 2018). The radar is used to determine relative

speed to objects in front in order to control speed for the Traffic Aware Cruise Control. The front-facing cameras can also recognize objects such as vehicles, pedestrians and road signs. The ultrasonic sensors are used to handle the blind spot and control auto lane change. They also control the Summon and Auto-Park features (Heath, 2018). In addition, there are a number of IoT sensors that function to open the doors of a garage before the person arrives home, control the temperature, and to provide a framework whereby the driver can design an own app and use this app to check the battery status and control the speed of the car from anywhere (Wongthongtham *et al.*, 2017).

The proliferation of these devices in a communicating-actuating network has created the IoT, wherein, sensors and actuators blend seamlessly with the environment around us (Gubbi *et al.*, 2013). Along this steady ascent of development, Gartner Inc. forecasted that, by 2020, the devices connected in IoT will grow up to 20.4 billion units from 8.4 billion which was in use worldwide by 2017 (Gartner, 2017). Sensors from a large number of devices simultaneously and continuously generate a huge amount of data, often referred to as the *Big Data* (Ahmed *et al.*, 2017). Handling this vast *volume* of structured and unstructured data in different *varieties* (e.g., numbers, text, audio, images, video) that are being generated in a high *velocity* (i.e., real-time) imposes significant challenges when time, resources and processing capabilities are constrained (Tole *et al.*, 2013). The highly dynamic data acquisition technologies have challenged the way humans interact and operate systems with the help of AI, indicating an essential need for a new breed of AI. This new AI requires the capacity to augment the ability of current AI systems for the optimal utilization of the Big Data technologies, as such the Big Data to derive insights leading to decision, recommendations, and actions.

In the past, the digital representation of the natural world was mostly stationary and in silos; thus the traditional AI algorithms were designed specifically to address such discrete data representations. With the proliferation of Big Data and IoT, data sensing and capturing technologies have made significant advances, by generating large, multi-modal, multi-source, dense, high frequent and non-stationary datasets (Tole *et al.*, 2013). This has resulted in the natural environment being represented and reflected in the digital world in a higher resolution. Today, with multiple data capturing technologies (e.g., IoT, CCTV, wearables) as well as human interaction, opinions, moods and emotions capturing have led to creating a more holistic digital representation of events and situations from diverse data sources. As the changes in such situations are captured over time, the new digital world provides a closer representation of the natural world and it is dynamics than ever.

Traditional AI has been perfected to wrangle with discrete and stationary data modalities, that has resulted in an asymmetric learning paradigm where the historic data being used to

solve current tasks. This asymmetric learning paradigm is also known as isolated learning because it does not consider any other related and background information or take in to account any current knowledge (Hong *et al.*, 2020). The fundamental problem with this isolated learning paradigm is that it does not accumulate knowledge continuously throughout its lifetime to use it in future situations. This is in contrast to the human learning, where humans never learn in isolation but always accumulate knowledge throughout its lifetime and use it to guide future learning and problem solving. Thereby, whenever we encounter a new situation or problem, we may notice that many aspects of it are not really new because we have seen them in the past in some other contexts (Chen and Liu, 2016). Without the ability to accumulate knowledge, an AI system typically needs a large number of labelled past training examples in order to learn effectively.

Labeling of training data is often done manually, which is highly labour-intensive and time-consuming. The world is widely complex with many possible tasks, making it almost impossible to label a large number of examples for every possible task for an AI algorithm to learn. Adding further complications, the environment changes constantly, and any labeling thus needs to be done frequently and regularly to be useful, making it a daunting task for humans. In today's digital environment, where IoT is geared to generate non-stationary and high-frequent continuous data volumes, such practice becomes intractable. This would deem inefficient and unrealistic in most of the real-world scenarios, where the streaming data might disappear after a given period of time and may not allowed to be stored at all due to storage or privacy constraints (Aljundi, 2019). This has generated the need for a symmetric learning paradigm where the AI system is not only geared to learn from past data to solve the current task but to harness the non-stationary and continuous data streams to acquire knowledge for past and future tasks.

In contrast, human knowledge acquisition process is quite different. Humans accumulate and maintain the knowledge learned from previous tasks and use it seamlessly in learning new tasks and solving new problems. Over time we learn more and more, and become increasingly knowledgeable, and more effective at learning. *Continuous Lifelong learning* (CLL) in AI aims to mimic this human learning process and capability, where the AI systems are designed to learn from continuous streams of data adapting to the external environment and associated with different tasks with the goal of augmenting the acquired knowledge for problem solving and future learning (Chen and Liu, 2018; Aljundi, 2019). CLL is also known as lifelong machine learning, continual learning and continuous learning, bearing a resemblance to lifelong learning of humans whose learning system is perfected in harvesting non-stationary and continuous data streams to acquire knowledge to perform past and future tasks. CLL stands for the smooth update of AI systems taking into account different tasks

and data distributions while still being able to re-use and retain useful knowledge and skills learnt previously. CLL is a learning paradigm that focuses on a higher and realistic time-scale where data and tasks become available in real-time and the access to previous data are limited.

Early approaches for CLL consisted of memory systems that stored previous data and regularly replay these previous data interleaved with samples drawn from new data (Robins, 1993; Rebuffi *et al.*, 2017). A major drawback of storing previous data throughout the lifetime of learning models is that they require explicit storage, leading to larger memory requirements. In addition, for connectionist models, due to the limited fixed number of neural resources, special mechanism need be used to consolidate knowledge from being overwritten and to maintain the same model performance level for different data distributions (Parisi *et al.*, 2019). Generally, in a lifelong learning scenario, the number of tasks to be performed is not known at the beginning, and as such constraining the learning model with a pre-defined amount of neural resources may compress the knowledge leading to gradual degradation of performance. Thereby, allocation of additional neural resources for new knowledge in connectionist models have been attempted in recent work to address the fixed neural resource limitation (Parisi *et al.*, 2016; Rusu *et al.*, 2016). For instance, additional neurons are added to a neural network architecture in subsequent learning steps when the model is exposed to new data. However, addition of neurons continuously may lead to scalability issues when the neural architecture becomes extremely large requiring increased computational efforts.

In contrast, humans are exposed to a dynamic world with a multitude of experience, and incrementally the human will acquire knowledge about the environment, from birth. At birth, the neuronal connections in biological brain begins at a relatively limited capacity and incrementally develop the capacity with age (Shatz, 1992). Thereby, advancing the development of CLL computation mechanisms using inspiration from biological brain with neural mechanisms with ability to incrementally adapt structure has the potential to bring improved performances (Parisi *et al.*, 2016).

In recent work, supervised and unsupervised machine learning prospects have been utilised to propose CLL architectures where, supervised learning prospects infer a learning from a set of labelled training examples and unsupervised learning aims to find hidden structures in unllabeled data. Supervised neural network based CLL architectures have been proposed in which complementary memory modules were utilized (Kemker and Kanan, 2017; Rebuffi *et al.*, 2017). A major limitation with such supervised learning methods is that the need for large volumes of labelled data, i.e., the learning algorithm is trained using a labeled dataset where each record is labeled with a known outcome. With data being frequently generated in high volumes, in most practical scenarios finding labeled data is unrealistic and generating such labeled data is time-consuming and expensive. In this light, AI systems

of the future are expected to incorporate higher degrees of unsupervised learning in order to generate value from unlabeled data. This type of unsupervised learning is quite natural because things around us are closely related and interconnected. Knowledge learned about particular subjects can help us understand and learn some other subjects. For example, we humans do not need 1,000 images of each animal in the planet as an AI algorithm would need in order to build an accurate classifier to classify animals.

In summary, traditional AI algorithms have been perfected to wrangle with discrete and stationary datasets that represent the natural environment in silos, which has resulted in an asymmetric learning paradigm where the historic data being used to solve current tasks. The advancement of IoT and Big Data have resulted in the natural environment being more holistically represented digitally and changes in the environment are continuously captured over time, which has created the need for a symmetric learning paradigm where the AI system should not only geared to learn from past data to solve the current task but to harvest the non-stationary and continuous data streams to accumulate knowledge for past and future tasks. In contrast, human knowledge acquisition process function differently by accumulating and maintaining the knowledge learned from previous tasks and using it seamlessly in learning new tasks and solving new problems. Inspired by this continuous lifelong learning nature of humans, this research aims to design and develop a new conceptual, theoretical and algorithmic foundation for AI systems to learn from continuous streams of data adapting to the external environment and associate with different tasks with the goal of augmenting the acquired knowledge for problem solving and future learning.

On this premise, this thesis intends to address limitations of current AI on the fronts of (1) representing the continuously changing natural world in the form of a digital representation informed by multitude of data sources, (2) capacitating AI systems with continuous lifelong learning, and (3) utilizing a higher degree of unsupervised learning in the AI systems, through the inspiration from philosophical and human neurophysiological findings.

1.2 Motivation

The rise of IoT enabled sensing technologies and Big Data have led to natural environment being more holistically represented digitally and changes in the environment continuously captured over time. Thereby, events, situations and behaviours of the environment are presented in a digital form for AI systems to function, in which the scientific community has made significant progress during the last decade to improve AI to adapt this new digital environment. Nevertheless, the current AI systems still have limitations to represent, adapt to and operate in the continuously evolving natural environments. In contrast, humans excel at

learning continuously while constantly adapting to and exploiting new information, making appropriate decisions on the basis of sensorimotor experiences learned throughout their lifespan (Bremner *et al.*, 2012). This ability to continuously acquire information and refine knowledge over sustained periods of time is regulated by a rich set of neurophysiological formations and cognitive functions in the biological brain that together contribute to the development of our perceptual and motor skills (Lewkowicz, 2014; Parisi *et al.*, 2019). Thereby, this thesis finds a distinctive opportunity for proposing inspiration from the neurophysiological formations and cognitive functions in the biological brain to advance current AI systems to not only be geared to learn from past data to solve current tasks but to harvest the non-stationary and continuous data streams to continuously acquire knowledge to solve future problems.

Nature is unpredictable if not indeterminate. The scientific community has demonstrated the complexity of nature in terms of chaos theory, the role of consciousness, free will in human behaviour and implications of quantum mechanisms (Boccaletti *et al.*, 2000). To perceive and function in this natural environment, human's biological sensors and brain collaboratively constitute a holistic representation through the information it has conceived over the lifetime and self-structure based on the new information and experience gathered by the human (Hawkins and Blakeslee, 2007). In contrast, the current AI systems have been built for pre-identified and well-defined problems, and they have limited capabilities to represent such a holistic view of the environment and continuously update their knowledge based on the novel experiences they acquire due to their asymmetric learning mechanisms (Nawaratne *et al.*, 2018). To overcome these limitations, AI systems should be able to develop representations of the environment that resembles the volatility and evolution of the external natural environment. Thus, pre-defining the structure of this representation could saturate the knowledge at a very early phase making it unstable to acquire new information while preserving previously learnt knowledge. Therefore, looking beyond traditional computational approaches, this thesis proposes an exploration of the ability of *Self-Structuring*, i.e., ability to update the knowledge representation structure over time by going beyond the traditional practise of only updating the weights representing learnt knowledge within a pre-fixed structure, as essential to lay the foundation for the new AI systems. We advocate that self-structuring is not only of pivotal importance for the development of neural structures with better internal representation but also helps to bootstrap the emergence of cognitive abilities encoded into a latent representation (Lungarella and Sporns, 2005). Accordingly, this thesis is motivated to bring further advancements to AI systems, in order to function in the natural environments by capacitating them to generate a holistic representation of the environment, inspired by how humans perceive nature and learn from it.

One of the main obstacles in AI algorithms towards continuously adapting to the environment is *catastrophic forgetting*, which results in severe disruption to existing knowledge during the acquisition of new knowledge (French, 1999). Catastrophic forgetting typically leads to an abrupt performance decrease or, in the worst case, to the past knowledge being completely overwritten by the new. While humans can gradually forget past knowledge, a complete loss of previous knowledge rarely occur unless due to a failure in biological organs or functions (Aljundi, 2019). Thereby, it is deemed justified to assume that for an AI mechanism to succeed in preserving knowledge, it can be inspired by the mechanism of human knowledge acquisition (Said and Masud, 2013).

Further to that, most existing AI systems that are designed to perform in the natural environment are based on supervised machine learning techniques that are based on examples in labelled training datasets. In contrast, humans have the ability to self-learn via exposure and observation to augment supervised learning. Incorporating a higher degree of such unsupervised self-learning, i.e., ability to recognize patterns, learn from data, and become more intelligent over time, in an AI system can lead to a realistic performance in natural environments. This thesis is motivated by the self-learn capability of humans, thus propose unsupervised self-learning for more biologically plausible AI that can successfully operate in natural environments.

As proposed by Kiritsis (2011), the advancement of AI should be followed through the inheritance of the essential characteristics and inspiration from the highly evolved human perception system. Accordingly, this thesis is inspired by these remarkable abilities of humans, with their unprecedented capabilities to adapt to diverse environmental conditions due to the evolution for over 6 million years through simple lifestyles to complex ones, under unpredictable environmental conditions and circumstances. This thesis is further inspired by the benefits to be gained by addressing limitations of current computational and AI approaches in representing, adapting and performing in the continuously evolving natural environments. Thereby, we are motivated to develop AI systems that can continuously acquire and augment knowledge from the natural environments thereby be used to derive insights that can be transformed into actions and recommendations that are useful for humans and society.

1.3 Aims and Objectives

The aim of this thesis is to conceptualise, design and develop an AI system for data-intensive digital environments based on neurophysiological findings, behavioral studies and natural phenomena. We intend to bring inspiration from underlying neural mechanisms of the

biological perception system, specifically, the human visual perceptive system, and the neural mechanisms of the biological brain.

First, the thesis investigates the limitations and opportunities introduced by the new digital environments augmented by IoT and Big Data alongside the existing AI techniques to perform in such environments. Next, we intend to study the biological counterpart, human brain, to understand how nature has provided means to address the identified limitations in intelligence based biological systems. Thereby, the thesis explores the structural, functional and behavioural facets of the biological sensors and brain in order to understand the ability of humans to demonstrate complex behaviours, skills whilst having a memory a symmetric learning mechanism that can constitute a holistic representation of environment through the information it has conceived over the lifetime, self-structure based on the new information and experiences, and continuously learn and adapt to evolving natural environments. The goal of the first phase is to develop an overarching conceptual framework to enable AI systems to achieve continuous learning capability through the sensing of natural environments.

Second, the thesis aspires to advance new self-structuring AI approaches by proposing theoretical and algorithmic formulations to; (i) represent the continuously changing natural world in the form of a digital representation informed by multitude of data sources, (ii) continuously acquire knowledge and adapt behaviour, and (iii) utilize a higher degree of unsupervised learning and self-learning.

Third, the thesis transforms the novel self-structuring AI algorithms proposed in this thesis to practical innovations in industrial settings to demonstrate and validate the contributions and benefits. The applicability of the proposed lifelong machine learning architectures are demonstrated in two real-world application areas: (i) intelligent surveillance in a smart city context, and (ii) intelligent memory articulation for human behavioral studies. The first case-study focuses on Smart Cities that endeavour to deliver safe, sustainable, effective asset utilization and service provision, amidst rapid urbanization. The second case-study focuses on Neuroscience and Mental Health providing a comprehensive view of stroke rehabilitation experiences through patient profiling and trajectory analysis using the proposed AI techniques, thereby contributing towards uplifting the quality of life and rehabilitation of stroke survivors.

1.4 Research Questions

Based on the aim and objectives, the research questions of this study are framed as follows:

1. How can computational continual lifelong learning enable the natural world to be represented digitally, making use of continuous streams of data from a variety of digital

sensors? What aspects of the structural and functional facets of neurophysiological studies can be used as a foundation premise to develop techniques for computational continual lifelong learning in data intensive digital environments?

- How has the new digital world, made up of Big Data and IoT, transformed AI systems in perceiving natural environments, acquiring and updating knowledge for past and future tasks?
- What are the core structural components and functional mechanisms in the human neurophysiological system that support the continual lifelong learning in humans?
- How can these neurophysiological facets of humans be used to inspire artificial representation of natural world in data intensive digital environments?
- How can such neurophysiological inspiration be used to combine the features of big data and digital environment to form an overarching conceptual framework?

2. What are the computational and machine learning constituents of continuous lifelong learning for materializing the proposed conceptual framework?

   - What are the computational and machine learning foundations for representation learning in the digital world that have been proven through both neurophysiological and ecological studies?
   - What are the structural and algorithmic limitations in current AI for achieving continuous lifelong learning and what fundamental architectural changes will address these limitations?
   - How can the knowledge embedded in computational models preserve stability and plasticity when introduced to continuous data streams?
   - With multiple facets and characteristics of data being captured to represent actions, events and situations, how can a comprehensive representation be developed for digital environments?
   - What neurophysiological theories enable the development of a computationally plausible memory formulation to achieve continuous lifelong learning?

3. How can such a self-structuring AI based continuous lifelong learning architecture with memory be developed in to technology platforms to advance AI systems in perpetual data intensive environments, such as national security, smart cities, and digital health?

In summary, the thesis aims to conceptualise, design and develop a lifelong machine learning memory formulation, inspired by neurophysiological findings and natural phenomenon, and validate the proposed memory architectures using real-world case-studies.

1.5 Research Contributions

Based on the above research objectives and research questions, this thesis yields the following contributions; theoretical contributions and application contributions, as the outcomes.

1.5.1 Theoretical Contributions

1. Founded upon neurophysiological inspiration and the features of big data and digital environments presented in the form of a landscape, proposed a new conceptual framework to address limitations of current AI to; (i) represent the continuously changing natural world in the form of a digital representation informed by multitude of data sources, (ii) enable AI systems with continuous lifelong learning, and (iii) utilize a higher degree of unsupervised learning and self-learning in the AI systems.

2. Advanced the current state-of-the-art in unsupervised self-structuring capability of growing self-organization with inspiration from the neurobiological concept of transience (forgetting) in its learning mechanism in order to reduce over-fitting and the influence of outdated information on the acquired knowledge, while preserving stability and plasticity of the knowledge base.

3. Extended the Growing Self-Organizing Maps (GSOM) (Alahakoon *et al.*, 2000) algorithm to enable the self-structuring neural network to capture the temporal resolution from non-stationary sequential input data streams. The novel extension equips the algorithm to bring forth temporal dependencies within the data stream.

4. Introduced a hierarchical multi-stream self-structuring architecture to provide a holistic digital representation of actions, events and situations obtained from multiple facets and characteristics of the natural environment. This architectural formulation was inspired by the human visual perception system.

5. Developed an unsupervised deep learning based active incremental learning method, Incremental Spatio-Temporal Learner (ISTL), for continuous anomaly detection in surveillance video (Nawaratne *et al.*, 2019c). The proposed method is able to continuously update and distinguish between new anomalies and normality that evolve over time. This technique was developed to address the largely overlooked aspect of video surveillance, which is the evolving nature of anomalous behaviours over time.

6. Developed a novel continuous lifelong machine learning algorithm to continuously acquire and fine-tune knowledge from the environment over sustained periods of time

in order to alleviates catastrophic forgetting in scenarios where sequential information becomes progressively available over time and access to previous information is restricted.

1.5.2 Application Contributions

1. Developed constituents of a comprehensive intelligent smart city platform exercising the capabilities of unsupervised lifelong learning representations coupled with supervised deep learning. The proposed application utilized the theoretical formation of self-structuring neural network to capture resolutions from non-stationary sequential input data streams of surveillance video footage, extended to a generative latent space.

2. Demonstrated the novel algorithms developed in this thesis using a real-world case-study on neuroscience and mental health. The presented case-study analyzes clinical outcomes of a stroke survivor patient cohort during three time-points: 7-days, 3-months and 12-months post-stroke with the aim to identify distinct clinical profiles and recovery trajectories. Clinical outcome was measured across physical, cognitive and mood domains, disability, stroke impact for work and social adjustment. The temporal resolution of clinical data is analyzed through the theoretical development of self-structuring AI to uncover latent patterns of stroke patient and stroke survivor trajectories.

1.6 Thesis Structure

The rest of this thesis is organized as follows. Chapter 2 provides an overview of neurophysiological functionality of biological brain of humans, and a methodology to tie the neurophysiological inspiration and features of the big data and digital environment together in the form of a landscape, leading to the presentation of an overarching conceptual framework *Multi-layered Self-structuring Knowledge Representation Framework* (MSKRF) that will functions as the basis for the research carried out and described in this thesis.

Chapter 3 explores the pertinence of the MSKRF in facilitating the overall objective of continuous lifelong learning of connectionist models. An in-depth study of viable computation models is provided to lay the foundation for the proposed conceptual framework resulting in identifying Self-Structuring Artificial Intelligence (SSAI) (Alahakoon *et al.*, 2020) with an unsupervised learning paradigm to be the most suitable, thereby, highlighting the Growing Self-Organizing Maps (GSOM) as an effective algorithmic base to lay the foundation for the MSKRF.

Chapter 4 focuses on materializing the internal representation mechanism of MSKRF by implementing the identified biological bases in order to preserve stability and plasticity within the knowledge framework. The overall goal of attaining continuous lifelong learning is described and modelled in Chapter 5 through the implementation of dual learning mechanism to process continuously streaming data to develop a memory mechanism that can continuously update the knowledge preserving what it has learnt previously thus addressing the problem of catastrophic forgetting.

Chapter 6 demonstrates the utilization of novel algorithmic developments from Chapter 4 and Chapter 5 in two application areas in two different contexts and data environments: a smart city and a digital-health. Finally, Chapter 7 presents a discussion of contributions and future work, followed by the conclusion of the thesis.

Chapter 2

A Conceptual Framework for AI Agents in Data Intensive Digital Environments

Data is ubiquitous in both natural and digital environments. With the advent of the digital age, large volumes of digital data is being generated from diverse artificial sensors, in forms of numbers, text, audio, images, video, etc. The availability of such rich sources of data and the availability AI based systems to capture, represent, manipulate and generate insights from these data sources have resulted in wider usage and popularity of AI even in routine day to day routine activities. Although starting to be widely used, the ability of many AI based systems to operate in big data ecosystems are still in its infancy. As described in Chapter 1, the inability of continuous lifelong learning is a key limitation for current AI to be successful in these big data environments.

In retrospect, human evolution for over 6 million years reveals a remarkable advancement of humans, where the human brain evolved throughout different conditions from simple lifestyles to complex, under unpredictable environmental conditions and circumstances. This evolution has provided humans with unprecedented capabilities to adapt to changing environmental conditions. Therefore, an in-depth exploration of the constituents of human brain can pave the way for a new direction to develop more adaptable AI systems. This is the direction of new thinking we intend to bring in this chapter, bringing together the neurophysiological inspiration, the features of the big data and digital environment presented in the form of a landscape leading to the proposal of an overarching conceptual framework as the basis for the thesis.

An overview of the constituents of this chapter is briefly presented in Fig. 2.1. First, we explore the advancement of digital ecosystems with respect to how AI systems perceive natural environments and the importance of these digital environments to support humanity when in need. We then discuss the challenges and limitations that exist in conventional AI

systems to function in such volatile digital environments, leading to the need for structural and functional enhancements. Second, we study the findings of the structural, functional and behavioural facets of biological brain, primarily focusing on the visual perception system and memory system to understand the ability of humans to demonstrate complex behaviours and skills whilst having a memory formulation that can continuously learn and adapt. Third, we analyze key limitations in current state-of-the-art AI models focusing on the needs of a new paradigm of AI systems to cater to the needs and harness the opportunities created in modern digital environments. Based on these findings and investigation results, we propose seven key constituents in human neuronal system that have the potential to aid and inspire its artificial counterpart to advance their capabilities in perceiving and representing the natural environment in order to process Big Data, derive insights that can be transformed to actions and recommendations.

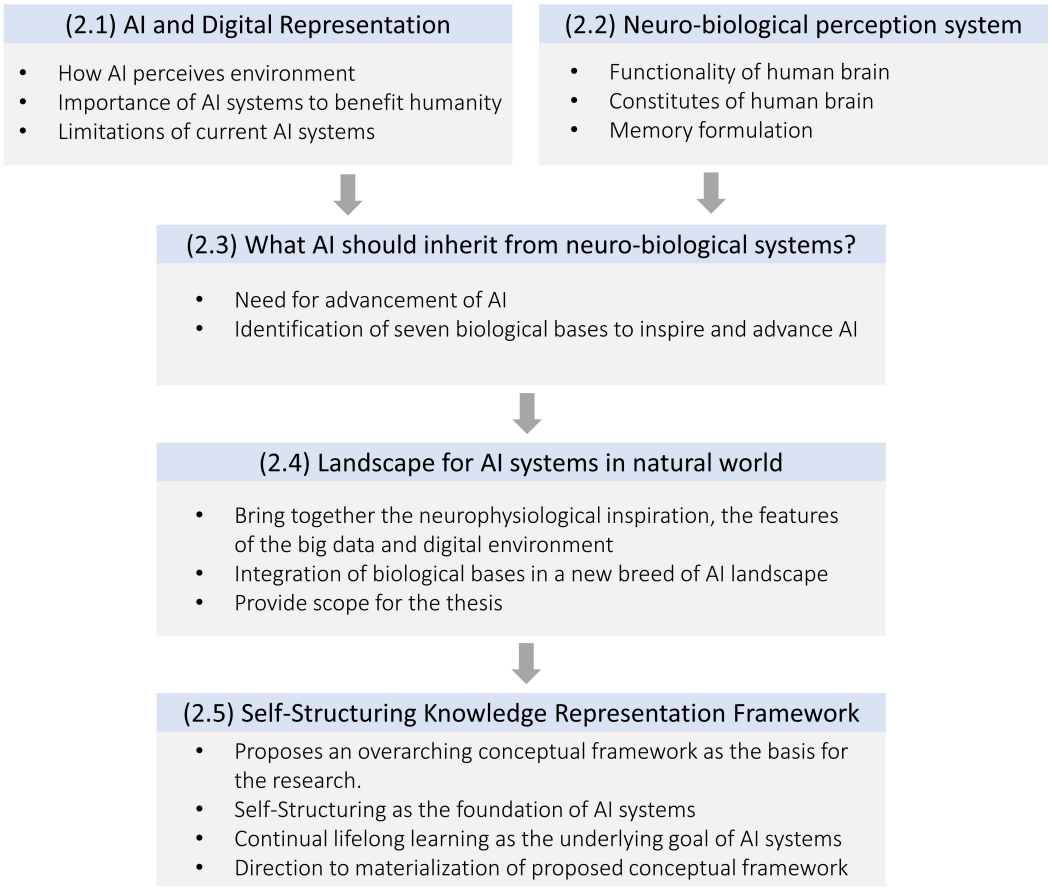


Fig. 2.1 Chapter Overview

In this chapter, we position the need for AI in natural ecosystems with its digital counterparts, proposing a landscape containing the key constituents. With the landscape used as

the foundation for developing the scope for this thesis, we present a conceptual framework for AI systems to continuously acquire information, represent its knowledge in a digital form, and continuously update the acquired knowledge based on the continuous evolution of natural environment. With the conceptual framework, we identify the underlying need for continuous lifelong learning in AI in order to acquire knowledge continuously from data streams, whilst having self-structuring knowledge representation as the foundation of learning. The conceptual framework provides a new direction the design and development of a new breed of AI systems.

2.1 Digital Representation of Natural Environments

This section aims to explore the advancement of digital ecosystems with respect to how AI systems perceive natural environments and the importance of these digital environments for supporting human needs. The constituents discussed in this section relate to module (2.1) *AI and digital representation* of the chapter overview presented in Fig. 2.1.

Greek philosopher Aristotle supposedly first classified the famous five senses: vision, hearing, taste, smell and touch. Over the years more senses have been included such as senses for balance, time, temperature, etc. (Mind, 2012). Humans operationalise these senses through biological organs such as eyes, nose, ears, tongue and skin. Comparable with human perception of the natural world that originates from human sensory inputs, the digital world derives this functionality through edge devices such as cameras, microphones, speed detectors, and a diverse range of smart detector devices to sense the natural environment.

For example, an autonomous vehicle (also known as self-driving car or driver-less car) is capable of sensing its environment and move safely with little to no human intervention. This requires a wide range of sensors to perceive its surrounding such as sonar, radar, lidar, global positioning system (GPS), odometer, inertial measurement units, and a wide array of smart cameras. In the past five years, self-driving cars have been improved and commercially made available (Shreyas *et al.*, 2020; Jha and Patnaik, 2020). Ride-hailing companies such as Lyft and Uber are migrating their human driver-based cars to autonomous vehicles, which are currently being manufactured for the market by major vehicle manufacturers such as Tesla, Porsche and BMW.

Many houses today are designed as smart-homes by automating their functionality and utilities. A smart home would include automated lighting control, climate control, entertainment systems, smart kitchen appliances, smart energy management and smart security and safety devices such as smoke detectors and surveillance cameras. Research by Statista (Solutions, 2019) has concluded that in 2020 the global smart home market is valued

at US$ 71.629 billion but is expected to have an annual growth rate of 20.7% until 2023, estimated at US$ 151.955 billion globally by 2023.

These sensors cut across many areas of modern day living, offering the ability to measure, infer and understand environmental indicators from delicate ecologies and natural resources to urban environments. The proliferation of these devices in a communication-actuation network creates the Internet of Things (IoT), wherein, sensors and actuators blend seamlessly with the environment around us (Gubbi *et al.*, 2013). The new era of IoT is a diverse space which encompasses a large variety of hardware and software creating a paradigm that machines and devices interact with each other in a global network, collecting information from the atmosphere and the environment without human intervention (Nawaratne *et al.*, 2018). Gartner Inc. forecasts that by 2020, the devices connected in IoT will grow up to 20.4 billion units from 8.4 billion which is in use worldwide by 2017 (Gartner, 2017). This expansion will create a worldwide network generating a digital representation of a number of aspects including natural behaviours of individuals and things within the natural world.

Such ecosystems would be able to capture and thus provide a digitalised view of global environmental changes. An ecosystem developed on IoT that are capable of sensing the natural environment, providing a digital representation, can be identified as a smart environment, where the digital representation can be transformed to insights and actions using AI. This will provide seamless possibilities for self-management on transportation systems, power plants, utilities, water supply networks, waste management, safety and security, information systems, schools, libraries, hospitals, and other community services. Decision and support systems can be built making use of this digital platform, sending alerts, and influencing emergency infrastructure (McLaren and Agyeman, 2015; Hart and Martinez, 2015; Cohen and Muñoz, 2016; Fourtané, 2018).

2.1.1 Limitations of Traditional AI in new Digital Environments

With the proliferation of Big Data and IoT, data sensing and capturing technologies have made significant advancements by generating large, multi-modal, multi-source, dense, high frequent and non-stationary datasets (Tole *et al.*, 2013). This has resulted in the natural environment being represented and reflected in the digital world at a clearer resolution. With multiple data capturing technologies (e.g., IoT, CCTV, wearables) as well as human interaction, opinions, moods and emotions capturing have led to creating a more holistic digital representation of events and situations from diverse data sources. As the changes in such situations are captured over time, the new digital world provides a closer representation of the natural world and it is dynamics than ever.

In the past, representation of the natural world in digital form was mostly stationary and in silos; thus, the traditional AI algorithms were designed specifically to address such discrete data representations. This has resulted in an asymmetric learning paradigm where the historic data providing insights to solve current and future problems. This asymmetric learning paradigm is also known as isolated learning because it does not consider any other related information or account for the current knowledge (Hong *et al.*, 2020). The fundamental problem with this isolated learning paradigm is that it does not retain and accumulate knowledge learned continuously to be used in future learning. Without the ability to accumulate knowledge, an AI system typically needs a large number of training examples in order to learn effectively, where labelling of training data is often carried out manually, which is highly labour-intensive and time-consuming and, in many instances, not possible due to complexity and volatility of data and difficulty in identifying labels for each and every task.

As such the digital environments today are geared to generate non-stationary and high-frequent continuous data volumes where asymmetric/isolated learning becomes intractable and be deemed inefficient and unrealistic in most of the real-world scenarios. Streaming data might disappear after a given period of time and/or not allowed to be stored at all due to storage or privacy constraints (Aljundi, 2019). For instance, the surveillance video feeds generated by a Closed-Circuit Television (CCTV) device would accumulate up to 50 GB per day, and accumulation over a month would require 1.5 TB of storage and not allowed to be stored due to sensitivity. In such scenarios, storing these voluminous data to train AI models would deem inefficient and unrealistic. This has necessitated a symmetric learning paradigm where the AI system is not only geared to learn from past data to solve the current task but to harvest the non-stationary and continuous data streams to acquire knowledge for past and future tasks.

In contrast, humans and other animals have acquired capabilities to learn and adapt to diverse environmental conditions through the evolution for over 6 million years. For decades researchers have worked towards perfecting AI systems inspired by biological systems and thus, as proposed by Kiritsis (2011), we argue that the advancements of AI can greatly benefit by studying the essential characteristics of the highly evolved human perception system.

2.2 Human Perception of Natural Environment

The previous section presented a detailed review of AI and digital representation of the natural environment, leading to challenges and limitations that needs to be addressed to successfully function in modern digital ecosystems. In contrast, this section looks into the

biological counterpart of AI. An in-depth exploration of how biological perception system facilitates humans to perceive the natural environment is presented in this section, related to module (2.2) *Neuro-biological perception system* of the chapter overview presented in Fig. 2.1. In this section, we attempt to answer the first research question (RQ1): What aspects of the structural and functional facets of neurophysiological studies can be used as a foundation premise to develop techniques for computational continual lifelong learning in data intensive digital environments?

2.2.1 Biological Brain

Adult human brain consists of approximately 100 billion intricately connected neurons that makes possible memory, vision, learning, thought, consciousness and various functions of the mind (Shatz, 1992; Herculano-Houzel, 2009; Quiroga, 2017). The precision of connectivity and wiring in the brain makes it the most remarkable development of human anatomy. It is fascinating that this remarkable complex structure begins its development during the fetal development in the first few weeks after fertilization, while many of the sensory organs are not even connected to the embryonic processing centers of the brain (Konkel, 2018).

While each neuron is connected with nearly 10,000 other neurons, not all of these connections are active. Some connections are constantly reinforced while some are used in seldom. Quiroga (2017) presented an analogy for these connections in the brain using vehicular traffic in an arterial road network. The connections that are constantly reinforced are like a high-traffic highway that offers a convenient link between two places, while others resemble deserted routes, one that could in principle connect two places but in practice does not. Similar to the unused roads, the neural connections that are seldom used may disappear over time. Building on this analogy, changing the connectivity of the brain is similar to re-routing vehicles on a congested road. These changes will eventually bring about changes in what information these neurons encode. This phenomenon is known as the *neural plasticity*, and a key mechanism used in the brain to generate and store specific memories.

The concept that memories relate to neural connectivity was discovered in 19^{th} century, however, the most important contribution to this hypothesis was offered by Hebb (1949). Hebb postulated that the joint activation of neurons reinforces the connection between them. This phenomenon can be summarised as; "Neurons that fire together, wire together". That is, if two neurons tend to fire at the same time, it is likely that they encode similar information and thus, it makes sense that they are connected. Thereby, the connection between the two neurons are reinforced. In contrast, the wiring between neurons that tend to fire at different times are weakened (Quiroga, 2017).

Mountcastle (1957) first found traced touch and flutter signals through the thalamus and into the cerebral cortex of a cat, where he described for the first time successive neurons responded to the same patch of the skin and had similar response properties. These findings were a breakthrough in modern understanding of the brain's functionality, in which certain neuronal cells in the mammalian brain respond selectively to specific sensory stimuli. The classic experiments of Hubel and Wiesel (1962) extend and confirm the findings of Mountcastle, where they demonstrate how neurons along the visual pathway extract increasingly complex information from the patterns of light cast on the retina to construct an image. Hubel and Wiesel presented a topographical structure in the visual cortex of mammalian brain that represents the visual field, where nearby cells process information from nearby visual fields. In their work, Mountcastle, Hubel and Wiesel revealed the organization of cortical neurons, fundamental properties of objects in the environment and resulting mammalian perception of the natural world around them.

To provide a high-level understanding of the biological brain, key features and components are discussed in this section. An illustration of the human brain indicating the parts that are involved in memory is presented in Fig. 2.2. First, a highly uniform, pink/gray shaded outer surface can be found resembling a smooth cauliflower with a number of ridges and valleys (termed gyri and sulci). This is the neocortex, a thick sheet of neural tissue that envelops the inner parts of the brain. The neocortex is responsible for the intelligence of human, i.e., perception, language, thought, attention, mathematics, and planning (Molnár and Pollen, 2014). While the other brain structures such as brain stem, basal ganglia, amygdala etc. are important to the functioning of the human, the essential aspects of human intelligence occur in the neocortex, the thalamus and the hippocampus (Hawkins and Blakeslee, 2007).

The neocortex, approximately 2000 cm^2 of size (Hofman, 2014), consists of six horizontal layers labelled from the outermost inwards as I to VI. These layers are segregated by cell type and neuronal connections. All these neuron types contain branching structures called axons and dendrites. Dendrites bring information to the cell body and axons take information away from the cell body. Information from one neuron flows to another neuron across a synapse, i.e., when an axon from one neuron touches the dendrite of another neuron forming a connection called a synapse (Shepherd, 1975).

The entire neocortex structure has a consistent shape with limited landmarks, e.g., fissure separating the two cerebral hemispheres and a sulcus that divides the back and front regions. The structure is developed with specific regions for certain functions. For instance, a stroke in a person's right parietal lobe can cause the loss of ability to perceive on the left side of the body. A stroke in the left frontal region, i.e., Broca's area (Anwander *et al.*, 2007), compromises the person's ability to use the rules of grammar, although vocabulary and the

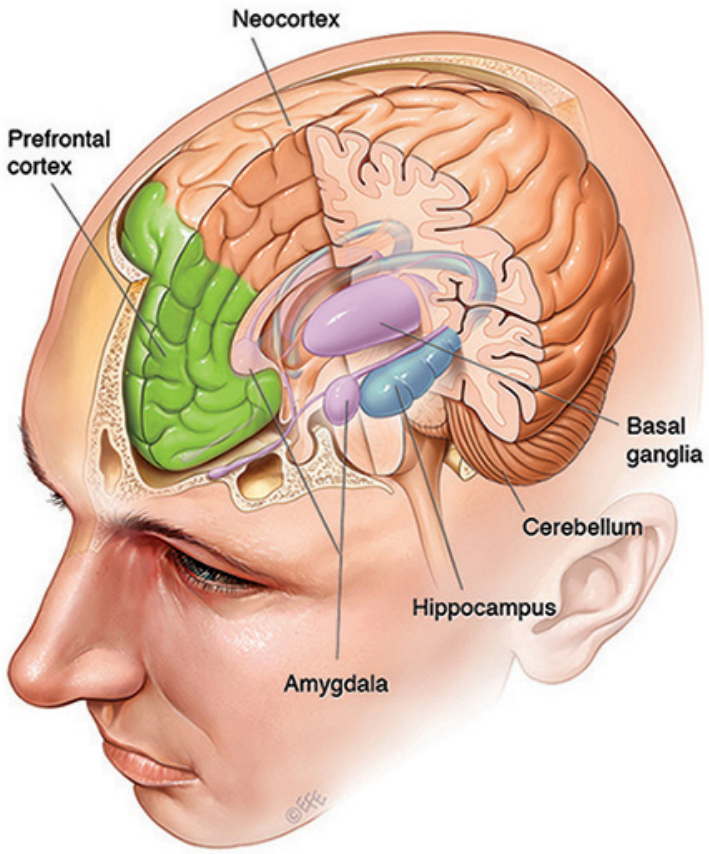


Fig. 2.2 The parts of the brain involved in memory (Illustration by Levent Efe)

ability to understand the meanings of words are unchanged (Hawkins and Blakeslee, 2007). Similarly, a number of functional areas of the brains have been discovered in the past century, yet, much remains to be discovered (Schoenemann, 2006). The left hemisphere is responsible for processing the right visual field and the right hemisphere the left visual field.

The complex neocortex structure is arranged in a branching hierarchy. In general terms, "hierarchy" describes the relation among a set of elements indicating which element lies “lower” or “higher” of another (Bond, 2004). In the context of neuroscience, the concept of hierarchy has been employed to examine the order of connectivity within the neurons in human brain. In this hierarchy, lower areas feed information up to the higher areas by a certain neuronal pattern of connectivity, while higher areas send feedback down to lower areas using different connection patterns (Hawkins *et al.*, 2017). Further to this, there exists lateral connectivity between the neurons interconnecting them in the same hierarchical level.

The brain processes sensory inputs from multiple sensory modalities, i.e., vision (sight), audition (hearing), tactile stimulation (touch), olfaction (smell), and gustation (taste), via the

hierarchical structure of the neocortex. The lowest of the functional regions, the primary sensory areas, are where sensory information first arrives in the cortex. These regions process the information at its rawest, most basic level. For instance, the primary auditory area is called A1, and connected to a hierarchy of auditory regions above it. This has a primary somatosensory area named S1 and a hierarchy of somatosensory regions connected higher the hierarchy (Lewis *et al.*, 2000). The next section provides an in-depth exploration of the hierarchical structure of the neocortex using the visual perception system.

2.2.2 Structural and Functional formulation of Biological Visual Perception System

Visual perception is a cognitive capability in the human mind that enables a human to perceive the ambient environment using light in the visible spectrum reflected by the objects in the environment. Visual perception forms a comprehensive and complex system in the brain involving approximately 60% of the brain. Roughly 20% dedicated for visual-only processing while the remainder is used in perception for combined sensory inputs such as touch, audio, video, attention, etc. (Keller *et al.*, 2012).

The process of visual perception originates when the eye focuses on a light signal either directly from a source or reflection from an object onto the retina, which is situated in the back part of the human eye. The retina acts as the physical optical sensor providing visual stimuli, a gateway from the eye to the brain. This visual sensation captured at the eye is processed by ganglion cells and transmitted to the Lateral Geniculate Nucleus (LGN), the superior colliculus of the midbrain, and the primary visual cortex. The perception of the visual stimuli occurs as a consequence of this retina generated neuronal signal being processed at the central nervous system, i.e., visual cortex. The characteristics such as brightness, sensitivity, colour, contrast, field of view, depth of field, spatial and temporal sensitivity are interrelated to produce the visual perception (Cardullo *et al.*, 2011). A simplified schema of the human visual pathways is illustrated in Fig. 2.3.

The visual cortex, located in the occipital part of the neocortex and approximately 64 cm^2 in size. It resides above the cerebellum and both hemispheres of the brain. The visual cortex process sensory information in a hierarchical manner, where different sub-regions of the visual cortex process information that are hierarchically advanced. The visual cortex composed of two subsections: primary visual cortex and secondary visual cortex.

The information from LGN, which has a spatial arrangement corresponding to a retinotopic organization, first flows to the primary visual cortex that corresponds to the Visual 1 (V1) area. The second section corresponds to visual areas; V2, V3, V4 and V5. The informa-

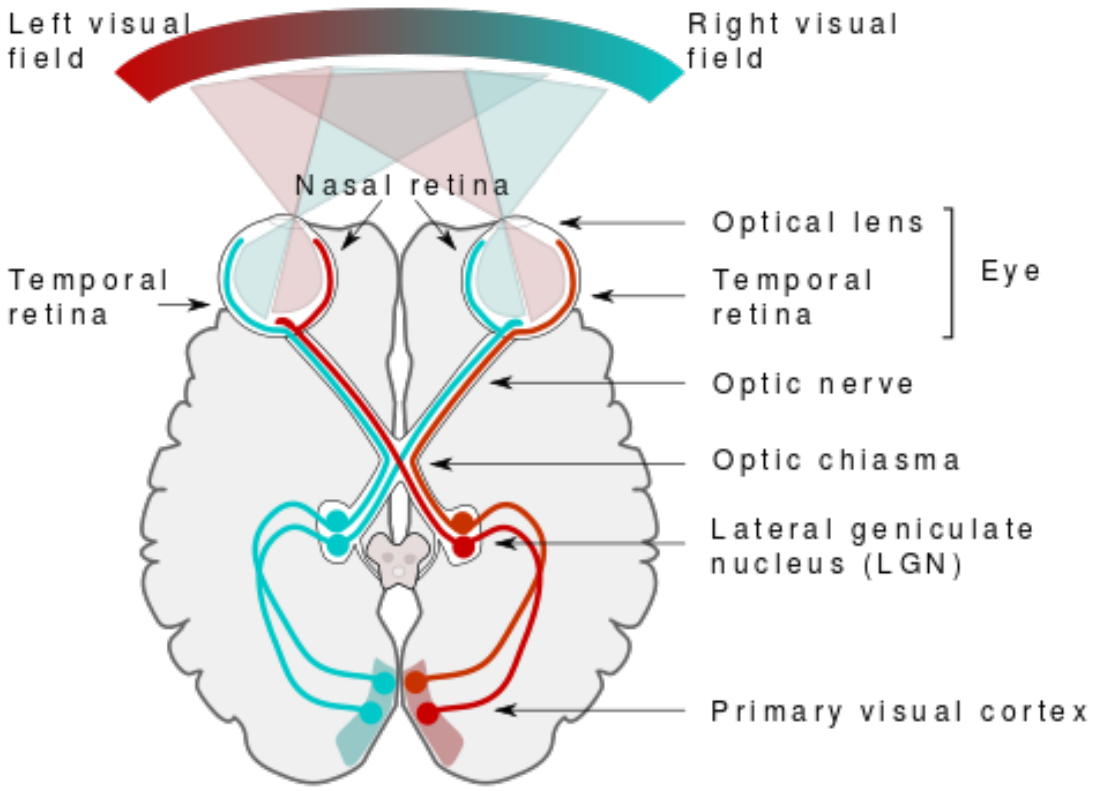


Fig. 2.3 A simplified schema of the human visual sensory. Adopted from Commons (2015).

tion from V1 follows to the deeper regions of the secondary visual cortex hierarchically. The first region, primary visual cortex (or V1) is highly developed and specialized in processing static objects and simple pattern recognition. As proposed by Li (2002), V1 generates a saliency map highlighting the important information of the visual input to guide the shifts of attention. This retinotopic map is a well-defined map of the spatial information in vision, that transform the visual input from retina to V1, in which even the blind spots are mapped into V1 using a phenomenon known as cortical magnification (Barghout-Stein, 1999).

The information encoded by V1 is better described as edge detection rather spatial coding. That is identifying points in the visual input at which the brightness changes sharply or provides discontinuities in the visual field. For instance, a visual input of a wall that is colored black on one side and white on the other side, the maximum contrast occurs at the dividing line. Thereby, this dividing line is encoded, while few neurons code the brightness information of the overall visual input.

The V2 area receives strong direct feed-forward connections from V1, and is the second major area in visual cortex. The neurons in V2 encode the visual inputs received from V1 area based on their orientation, spatial frequency and color. The encoding mechanism of V2 is common as the V1, however, the V2 area has a larger receptive field. Addition to the common encoding, V2 is specialized in modulating orientation and binocular disparity, distinguishing the foreground from the background (Qiu and Von Der Heydt, 2005). The encoded information processed at V2 is forwarded to further processing in two pathways, named the Ventral stream and Dorsal stream. Goodale *et al.* (1992) discovered the existence of these two major cortical pathways; the ventral stream and the dorsal stream, that are known to specialize to detect *what* the visual perception is and locate *where* the visual perception is in the view. This is applicable to both vision and hearing. The structural formulation of the two pathways are illustrated in Fig. 2.4.

The Ventral stream, also known as the "What Pathway", is associated with form recognition and object representation. The Ventral stream provides a description of the elements as well as utilized in judging the significance of these elements. It is also associated with storage of long-term memory. The Dorsal stream, also known as "Where Pathway", is associated with motion, representation of object locations, and guides the control of eyes and arms, especially when visual information is used to guide saccades or reaching. The Dorsal cortical stream located on the parietal lobe of the brain. The Dorsal stream responds to a limited field of view compared to the Ventral stream. The Dorsal stream transforms inward visual stimuli to a head-centered coordinate system. After the visual signals are processed at V2, the cortical organization and information processing is directed to these two streams.

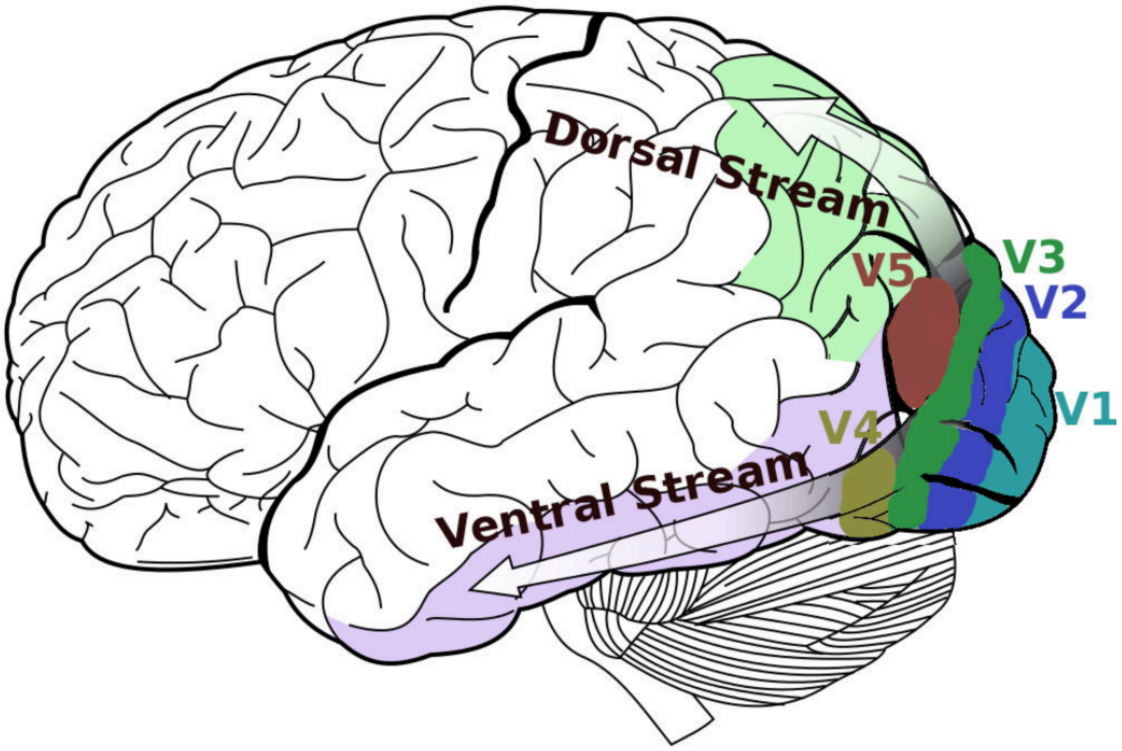


Fig. 2.4 The visual areas and processing pathways of human brain. Adopted from Barros (2017).

The Ventral stream composites of cortical regions V4 and Inferior Temporal Gyrus (IT). The feed-forward connections from V2 and V1 are received by the visual area V4. The neurons in V4, similar to the functionality of V2, encodes orientation, spatial frequency and colour. In addition, V4 encodes visual patterns with small complexity such as geometric shapes (circle, square, rectangle). Schmid *et al.* (2013) discovered that V4 exhibits long-term plasticity and are strongly modulated by attention mechanism. The information processed at V4 is then forwarded to IT region, the region responsible for higher levels of visual processing. IT region is associated with the representation of complex object features, such as global shape, and in addition it is responsible for face perception and numerical recognition (Haxby *et al.*, 2000).

The Dorsal streams composites of cortical regions V3 and V5. The neurons in region V3 are essentially associated with global motion (Braddick *et al.*, 2001). Receiving feed-forward connections from V1 and V2 regions, V3 encodes the visual input in a coherent motion of large patterns, proposing the characteristics of the motion. The feed-forward connections V3 receive from V2 are stronger than the ones receive from V1. The region V5, also known as Middle Temporal Region (MT), is mainly responsible for highest level of motion perception

including the integration of local motion signals into global perceptions and the guidance of eye movements (Born and Bradley, 2005). V5 receives feed-forward neuronal connections from V1, V2 and V3, however, unlike other visual areas, the strongest input to V5 is received from V1 area. This can be intuitively explained as the motion of an object is perceived by the brain from the motion of edges of the object. The neurons in V5 are primarily responsible for encoding speed and direction from the motion of the input visual field.

2.2.3 One Algorithm to Rule Them All

Vernon Mountcastle was the first to discover the uniformity in appearance and structure of the neocortex (Mountcastle, 1978). Drawing from his findings, Mountcastle proposed that a similar canonical circuit consisting of cortical columns underlies everything the neocortex does. Mountcastle indicated that due to the uniformity of all the regions in neocortex, perhaps the same basic operations are being performed by the neurons in each of the regions. For instance, the neurons in region that performs auditory processing operates in the same mechanism as the neurons in visual cortex region (Creutzfeldt, 1977; Hawkins *et al.*, 2019).

Recent studies have discovered that the development of neocortex is extremely plastic, as such the regions can adapt or train itself depending on the inputs flows to it. On this basis, researchers have surgically rewired the brains of new-born ferret such that their eyes send signals to the cortical region where auditory processing generally occurs. As a result, the ferrets developed a functioning visual pathway in the auditory portions of their brains (Nitta *et al.*, 1993). Similarly, human neocortex is extremely plastic, where studies have proven that changes in somatosensory input can remodel human cortical organization (Fraser *et al.*, 2002). For instance, congenitally blind adults use the rearmost portion of their cortex to read braille, which ordinarily becomes dedicated to vision (Hawkins and Blakeslee, 2007). As such, these studies have strengthened the initial findings of Mountcastle (1978).

To this end, all parts of the neocortex operate through a common principle (a common computational algorithm) with the cortical column being the unit of computation. Due to this uniformity, the regions of the neocortex are plastic and thus enables the human brain to adapt to available somatosensory (Hawkins and Blakeslee, 2007).

2.2.4 Complementary Learning Systems for Continuous Knowledge Acquisition

Humans have the ability to continuously acquire and fine-tune knowledge throughout their lifespan. This capability is governed by the capability of human brain that learns and memorise through sensory inputs from the environment. The learning is characterized by

the extraction of the structure of perceived experience with the aim to generalize to novel situations. Thereby memory requires the collection of separated experiences to occur at specific space and time.

A prominent discovery on learning and memory of the human brain was that the neocortex and the hippocampus complement each other for the continuous learning. This was formalised as the Complementary Learning Systems (CLS) theory (McClelland *et al.*, 1995). The neocortex gradually acquires structured knowledge representations while the hippocampus quickly learns the specifics of individual experiences. The CLS theory proposed that learning system is necessarily slow for two main reasons. First, each experience represents a single sensory sample from the environment. Given this, a small learning rate allows a more-accurate estimate of the underlying population statistics by effectively aggregating information over a larger number of samples (McClelland *et al.*, 1995). Second, the optimal weight of each connection in the memory leans on the values of all of the other connections. Before these connections are exposed to experiences, the initial weights of these connections are noisy and weak, leading to a slower initial learning. This slow learning form has been both theoretically and practically proven and particularly important in deep neural network architectures that consists of many layers (LeCun *et al.*, 2015).

Although there are advantages of gradual learning system such as neocortex of human brain, it suffers from two drastic limitations. First, such as system is unable to base its learning from an individual experience. For instance, consider experiencing a life-threatening situation - an encounter with a lion at a watering-hole. After such a situation, it is important to learn to avoid such encounters with lions and probably avoid that particular location of watering hole. Second, is the catastrophic interference (French, 1999) relating to stability-plasticity dilemma (Carpenter and Grossberg, 1987), in which the new information severely disrupts the existing knowledge. That is, even for a single experience, it could update the neuronal weights significantly to accommodate the current experience. This could lead to substantial adaptation of existing memory or identify as false knowledge. For instance, the new experience on encountering a lion would lead to the adaptation or changes to knowledge of other less-threatening animals the person may already be familiar with.

A second, complementary, learning system can address both these limitations, affording the rapid and relatively individuated storage of information about individual items or experiences such as the encounter with the lion (Kumaran *et al.*, 2016). The CLS theory proposed that the hippocampus and related structures in the Medial Temporal Lobe (MTL) support the initial storage of experience-specific information, including the features of the watering hole as well as those of the lion. This proposal has been captured in models of the role of

the hippocampus in recognition memory for specific items and in sensitivity to context and co-occurrence of items within the same event or experience.

The CLS based complementary memory modules and their interactions are depicted in Fig. 2.5.

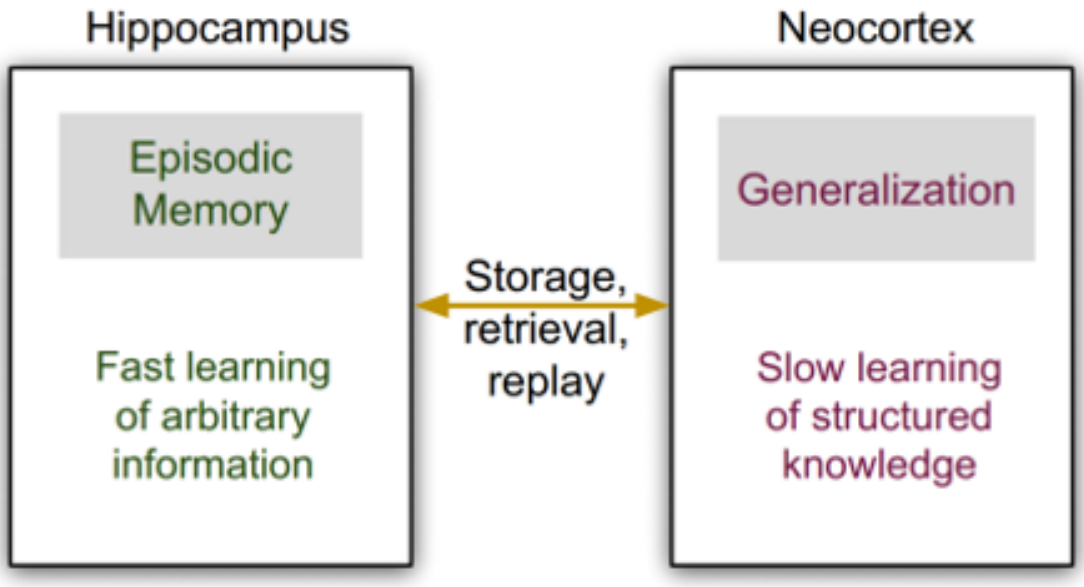


Fig. 2.5 Complementary Learning Systems Theory. Adopted from Parisi *et al.* (2019).

As mentioned above, the neocortex serves as the basis for the gradual acquisition of structured knowledge about the environment, characterized by a slow learning rate and builds overlapping representations of the learned knowledge. Conversely, the hippocampus allows rapid learning of the specifics of individual items and experiences, exhibiting short term adaptation of episodic memory (O'Reilly *et al.*, 2014). The interplay of hippocampus and neocortex is crucial in concurrent learning of regularities (statistics of the environment) and specifics (episodic memories), which is identified as the memory consolidation. It is interesting to note that the hippocampal replay promotes preferential treatment to unusual/significant memories such as high reward, information content, novelty and surprise. The phenomenon of memory replay, i.e. memory consolidation generally occurs during when the living being is at rest, mostly during the Rapid Eye Movement (REM) sleep (Taupin and Gage, 2002). Overall, the CLS theory holds the means for effectively generalizing across experiences while retaining specific memories in a lifelong manner.

This section detailed the biological perspective of human perception of the natural environment. Four aspects of neurophysiological perception system are described; i) constituents of human brain, ii) structural and functional aspects of visual perception, iii) uniform structural formulation of the neocortex, and iv) complementary memory formulation in human brain. First, we detail the findings on neocortex structure of the human brain that holds responsible for this sensory perception of the natural world. Neocortex provides means to process the sensory input from the environment, acquire, fine-tune and update knowledge based on the information processed for the prime objective of continuous lifelong learning of human beings. Second, we explored structural and functional aspects of biological visual perception discussing regions of neocortex responsible different functions and different pathways (Dorsal and Ventral streams) to process aspects of information. Third, based on findings from Mountcastle (1957), we identified that all parts of the neocortex operate through a common principle (a common computational algorithm) with the cortical column being the unit of computation. Fourth, we examined the CLS theory that states effective learning requires two complementary systems: one, located in the neocortex, serves as the basis for the gradual acquisition of structured knowledge about the environment, while the other, centered on the hippocampus, allows rapid learning of the specifics of individual items and experiences. We consider these neurophysiological details as inspiration for advancement and development of new breed of AI systems that can lead to our ultimate goal of continuous lifelong learning.

2.3 What should AI systems inherit from neurophysiological systems for developing an efficient digital representation of natural environments?

This section presents the neurophysiological inspiration and the associated features of the big data and digital environment as a preliminary step towards proposing the novel continuous lifelong learning system. At the highest level, this section relates to the module (2.3) *What AI should inherit from neuro-biological systems?* of the chapter overview presented in Fig. 2.1.

Previously, we discussed the need for a digital representation of natural environment and their current limitations in order to enable AI systems to digitally capture and perceive the environment, generate insights based on the perception and transform the insights into actions/recommendations (Section 2.1). The emergence of artificial sensory devices to perceive the natural environment has only been in existence for less than half a century, where

the interconnectedness emerged very recently, by the very end of the 19^{th} century (Baoyun, 2009; Foote, 2016).

In contrast, we explored the basis of human perception system (Section 2.2), in which human sensory and brain collaboratively provide the capability to perceive the natural environment. We have witnessed a remarkable evolution of humans for over 6 million years, where early humans just began to walk upright and make simple tools. The brain size of human has been increasing even during this period, however comparatively slowly (Appleyard, 2011). The faster expansion of brain capacity of humans started from 2 million years to 800,000 years ago, and from then, the human brain size evolved most rapidly during a time of dramatic climate change. Larger, more complex brains enabled early humans of this time period to interact with each other and with their surroundings in new and different ways. As the environment became more unpredictable, bigger brains helped our ancestors to survive (Brains, 2019). These deep roots confirm the remarkable improvement of the human brain adapting to the environment making them the epitome of intelligence in nature thus far.

As AI becomes increasingly widespread, a number of research work attempted to associate their relationship with biological intelligence. These studies have focused on understanding the workings of natural intelligence on the assumption that AI systems should mimic the mechanism of biological intelligence (Said and Masud, 2013). In this thesis, we identify key constitutes of biological brain that have the potential to inspire and advance a new breed of AI systems. Further, as proposed by Kiritsis (2011), the advancement of AI should be followed through the inheritance of the essential characteristics and inspiration from the highly evolved human perception system. On this premise, we propose a collection of seven biological bases, that have been essential for the survival and continuance of human beings to amalgamate in the development of AI systems in order to achieve continuous lifelong learning. These biological bases are:

1. Invariant representation of memory in neocortex.
2. Persistence and transience of memory.
3. Sequential information storage capability of neocortex.
4. Auto-associative recall of information from neocortex.
5. Hierarchical abstraction in memory storage.
6. Multi-modal information fusion.
7. Complementary Learning Systems Theory.

We discuss these bases in detail in following subsections with respect to their role and functionality in biological systems, what limitations of AI systems can be addressed through each of these bases and in terms of their importance for the advancement of AI.

2.3.1 Invariant representation of memory

Biological brain does not persist with and recall information with complete fidelity, but, only important relationships of the world independent of the details (Hawkins and Blakeslee, 2007). In the context of visual perception, visual cortex builds the representation of visual information that allow objects to occur relatively independent of size, contrast, spatial-frequency, position on the retina, angle of view, lighting, etc. (Rolls, 2008). For instance, when a person encounters a known person, regardless how much of the face of the person has changed or how distance the face is or variations of the face such as moustache, beard, etc., usually it is not difficult to identify the face. By taking a close look at the brain functionality in this context, specifically the activity of the neurons in the primary visual cortex (V1), the pattern of activity is different for each different view of the face. Even with a slight movement or change in lighting condition, the neuronal activation is different. However, the activation of cortical regions in higher levels such as V4 and IT cortical regions (Ventral stream that is responsible for object detection) are relatively stable for such differences. This stability of neuronal activation can be considered as the invariant representation.

In conjunction with the biological notion of invariant representation, a general theory of how humans construct reality has been postulated over two millennia. This theory is based upon the information the humans receive through their eyes, and the distinction between sensation (the physical stimulus impinging on the sensory organ) and perception (the interpretation of that stimulus) (Quiroga, 2017). Ancient Greek philosopher Aristotle postulated that starting with the information received through the senses, the mind generates images that are the basis of thought. These images, are the humans' interpretation of reality, an interpretation that generates concepts from abstractions by eliminating details and extracting the sheer meaning (Sorabji, 1974). Contemplating on the recent neurophysiological findings, the ancient idea has been well proven by the invariant memory representation of the biological brain (Hawkins and Blakeslee, 2007). These invariant representations of sensory inputs are extremely important for the survival and operation of the brain as it enables to learn with a single trial about reward or punishment associated with the object/occurrence, the location, how it was encountered, and then to correctly generalize to other views of the same object/occurrence (Rolls, 2008).

Considering the artificial counterparts, most state-of-the-art (SOTA) computational models suffer from this invariant representation. For instance, for the simplest task of object

detection, SOTA computational models should be trained with large volumes of training samples to encode a correct representation for each object (Alom *et al.*, 2018). Yet, with a slight change of a pixel or orientation or scale would lead to a completely different prediction (Su *et al.*, 2019). However, in an era where mission critical systems such as autonomous vehicles entirely depend on these object recognition systems to detect its surrounding, such errors are prone to be critical. Therefore, it is of the highest importance to provide the invariant representation capability to AI systems, in order to enhance the efficiency and effectiveness of these computational models.

2.3.2 Persistence and Transience of memory

"In the practical use of our intellect, forgetting is as important as remembering." - James (2007)

The predominant focus of neurobiological studies of memory have been on persistence (remembering). Through decades, many researchers identified the ideal memory system as one of perfect persistence, i.e., a memory system that transmits the greatest amount of information, with the highest possible fidelity, across the longest stretches of time (Richards and Frankland, 2017). However, few case-studies contradict the notion of persistence that said to govern the effective memory system. The famous tale of Patient S. is one such example.

Soviet clinical neuropsychologist Dr. A. R. Luria described of Patient S. as; "Patient S., a man with a "vast memory" who could only forget something if he actively willed himself to do so" (Luria and Solotaroff, 1987). In despite Patient S.'s remarkable capability to persist the memory, he was handicapped by this apparent super-human memory. Despite the fact that he was able to remember incidents in exquisite detail, his memory was inflexible to generalize across instances. For instance, given an image of a house, Patient S. was not able to recognize the fact that it is a house unless he has previously seen it. This points to the importance of transience (forgetting) as an essential element of an effective memory system.

As presented by Richards and Frankland (2017), the most intuitive explanation for the need for transience mechanism in the memory system is to help the memory to *make room* for new memories. Nonetheless, considering the sheer number of neurons, i.e., approximately 100 billion neurons (Azevedo *et al.*, 2009), it would seem that there exists ample capacity to store many more memories than a human makes in a lifetime. Thus, it requires further insights to understand the underlying reason for the transience.

In his work, Richards et al. proposed that memory transience is required in a world that is both volatile and noisy. In such volatile environments, transience provides means to adaption, allowing human beings to achieve more flexible behavior. Thereby, forgetting

allows to adapt by preventing overfitting to peculiar occurrences (Richards and Frankland, 2017). According to this perspective, over persistence would lead to inflexible behavior and incorrect predictions. Thus, we can derive the interaction of persistence and transience is mandatory for an effective and efficient memory system to serve its true purpose: i.e., not only the transmission of information through time, rather, optimize decision making (Schacter *et al.*, 2007).

The persistence and transience properties of biological brain, with its essential invariant representation capabilities of neocortical structure, serves two major purposes:

1. Reduce the influence of outdated information on memory-guided decision-making.

2. Prevent overfit to specific past events, thereby promoting generalization.

Overfitting and generalization are two key concerns in any computational model. Generalization is a term used to describe a computational model's ability to respond to new data, i.e., after being trained on a training set, a model can digest new data and make accurate predictions. The success of the computational model entirely depends on the model's ability to generalize. However, if the model has been trained too well on training data, it will not be able to generalize. It will make inaccurate predictions when given new data, making the model incompetent even though it is able to make accurate predictions for the training data. This is called overfitting in the context of machine learning. Many computational models attempted to reduce the overfitting to make the computational models more generalized by using a number of different regularization techniques such as drop-out (Srivastava *et al.*, 2014) in connectionist models, weight decay over time (MacKay and Mac Kay, 2003), sparse coding (Olshausen and Field, 1996) and noise injection (Hinton and Van Camp, 1993). These approaches for regularization resemble forms of partial forgetting.

On this basis, in developing brain inspired computational models, it is essential to identify the parallels between how transience can be used computationally and how it appears to be implemented in the brain. Thereby, making computational models more robust to changes in nature and adapt accordingly in its life-long learning.

2.3.3 Sequential storage and Auto-associative recall of information

Storing memories and recalling are essential characteristics in humans and other mammalians in order to make predictions, recognize patterns and generate behaviour. As these major functions originate from the neocortex, the ability to store and recall information are key attributes of the cortical areas. In this light, neuro-biological researchers have discovered that the neocortex may use a sequential approach in storing of patterns (Rodriguez *et al.*, 2004).

Habitually, when a person tells a story, he/she can only relate one aspect of the story at a time. It is not possible to explain everything in the story at once, but requires to finish one part before moving to the next part. Not only because the language is sequential, but, when the person thinks of the story, it is not possible to recall all the information regarding the story at once. It requires to bring up the thought process sequentially, as a series of thoughts or events. Furthermore, once the person comprehends information of a particular event, the subsequent events will automatically be recalled into memory. Consider the English alphabet as an example. It is difficult to recite the letters of the alphabet backwards (unless practiced), despite the fact that reciting the alphabet forward is simple and straightforward. Same with reciting a song or a poem. As presented by Jeff Hawkins, storing information in the cortical structure sequentially, and recalling in an auto-associative manner are inherent aspects of human neocortical memory system (Hawkins *et al.*, 2009).

Hawkins proposed that sequential storage and auto-associative memory are essential developments in biological neocortical systems that provides a promising way-forward for computational models (Hawkins and Blakeslee, 2007). Early memory models and computational models did not generally store sequences of patters, thus deviating from the inherent biological neocortical structure. Later, Elman (1990) inspired this sequential memory concept in their Simple Recurrent Network (SRN), which is a three-layer network with the addition of a set of context units. SRN maintains a state, allowing it to perform such tasks as sequence-prediction that are beyond the power of a standard multilayer perceptron. Further, development of this idea brought the advancement of Recurrent Neural Network (RNN) that has shown great success in computational tasks such as speech recognition, text generation, time series forecasting etc. Existing work attempted to implement sequential memory in computational models focus on supervised learning paradigm, in which volumes of human annotated training data are required. These techniques, however, are contrary to how human brain is characterized due to this supervised nature. Further, digitization has brought upon big data volumes generating seamlessly, where labelling and human annotation is unrealistic making the current supervised sequential learning techniques unsuitable for the future.

On this basis, in developing computational models to represent the natural environment, the recurrent/sequential information processing behavior is essential since the natural environment is proliferated with sequential data streams.

2.3.4 Hierarchical abstraction in memory storage

Regions of the neocortex are formed in a hierarchy (Felleman and Van, 1991). When sensory inputs enter the neocortex, initial regions (e.g., primary visual cortex for visual input) detect

basic characteristics of the input. The output of the first region passed onto the next region, that combine these basic features into more complex features, and this process is repeated and continued until several levels up in the hierarchy, in which the neurons respond to complete concepts (or objects) (Hawkins *et al.*, 2019). This hierarchical model is characterized by its hierarchical and feedforward organization.

Considering the visual inputs, neurons in V1 region, with small receptive fields, are sensitive to basic visual features. For instance, neurons in V1 respond predominantly to edges and lines. These neurons projects to neurons at next stage of the hierarchy that code further complex patterns, e.g., basic shapes such as circle, square. By the IT region, the neurons respond in a viewpoint-invariant matter to complete objects, such as faces, people, vehicles, etc. This biological phenomenon is effectively illustrated in Fig. 2.6.

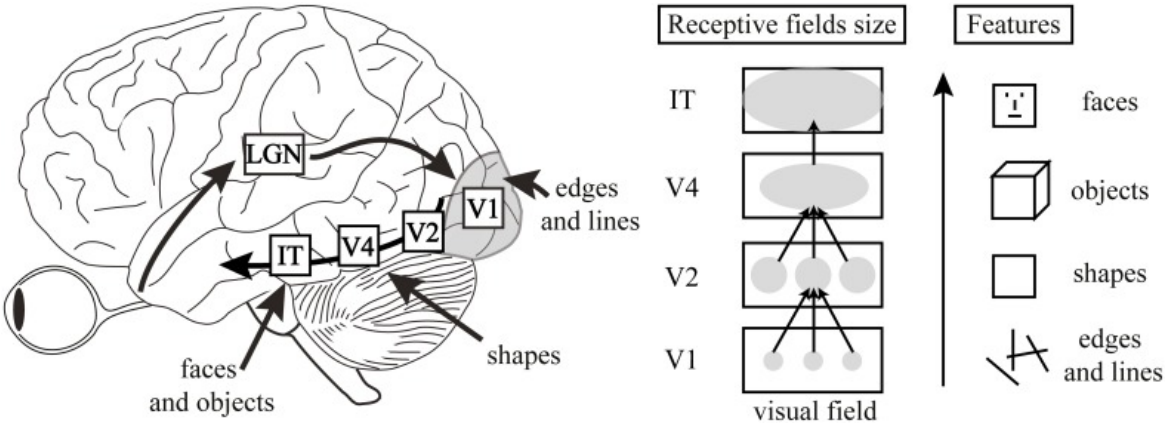


Fig. 2.6 Hierarchical, feed-forward model of the brain for visual perception. Adopted from Herzog and Clarke (2014).

This hierarchical model of the neocortex that comprise of hierarchy of feature extractors is mimicked by recent computational models in supervised deep learning, e.g., Convolutional Neural Networks (CNN), which have shown a great success in many applications such as face recognition, object detection, audio transcription etc. (LeCun *et al.*, 2015). Resembling the hierarchical organization in neocortex, these computational models attempt to encode advanced level of abstraction of its input incrementally in their hierarchical layers. This success justifies the importance of this hierarchical nature of input representation in supervised learning paradigms, thus, provides a promising way forward to adapt hierarchical representation in AI systems that can perform in unsupervised learning paradigm to represent the natural world efficiently.

2.3.5 Multimodal information fusion

The natural environment is highly complex; thus, human nervous system is equipped to sense this complex environment using multiple, different sensors. When an event occurs, more than one sensor detects the event, generating redundant neural signals. This underlies multisensory processing, that is of substantial adaptive value and has been extensively examined in the cerebral cortex of mammals (Stein and Meredith, 1993; Stein *et al.*, 2014). Recent studies on neurophysiology have discovered that the biological brain is already capable of processing multisensory information at the fetal development (Sours *et al.*, 2017), specifically examined in two cortical regions; Intraparietal Sulcus (IPS) and Superior Temporal Sulcus (STS). Both IPS and STS receives convergent inputs from visual, auditory and somatosensory areas. IPS has been implicated to signal the position of the body in space, eye movement, and the geometrical properties of objects such as shape, size, and orientation (Seltzer and Pandya, 1980).

STS is equipped for face and voice perception, processing of complex visual stimuli, and visual-auditory processing (Hein and Knight, 2008). For instance, the two cortical pathways of visual perception, the Dorsal stream and Ventral streams (as discussed in Section 2.2.2), are converged at the STS cortical region to generate a perception of the visual stimuli on *what* the stimuli is and *where* it is located in the view of the beholder. Further, STS is assumed to be implicated in social perception, demonstrating increases activation related to voices versus environmental sounds, stories versus non-sense speech, moving faces versus moving objects, biological motion, and theory of mind (i.e., false belief stories versus false physical stories) (Grossman and Blake, 2001; Beauchamp, 2015). In this light, robust lines of evidence attest that multisensory convergence and processing develops within the adult IPS and STS brain regions.

Analogous to the multisensory convergence in human brain to perceive the natural environment, equivalent mechanism is vital for its artificial counterpart to represent and perceive the natural environment. For instance, Jayaratne *et al.* (2018) discussed the need for bio-inspired multisensory information fusion as a requirement for autonomous robots to form an unambiguous and meaningful representation of their surroundings. Authors presented an empirical evaluation on an audio-visual dataset consisting of utterances to evaluate the quality of multimodal fusion over individual unimodal representations. The results dictated that multimodal representation achieves significant improvements over the unimodal representations. Drawing from these empirical research, we believe that multimodal information fusion is an essential element in the development of computational models to perceive the natural environment.

2.3.6 Complementary Learning Systems Theory

Biological beings are equipped with the capability to continuously acquire (learn), fine-tune (adapt) and transfer (share) knowledge and skills throughout their lifespan. This ability is known as Continuous Lifelong Learning (CLL) and it is mediated by a comprehensive neurocognitive mechanism that together activate with specialized sensorimotor skills and long-term memory consolidation and retrieval (Hamker, 2001; Parisi *et al.*, 2019; Riemer *et al.*, 2019). An effective CLL system should learn over a continuous stream of non-stationary environment, and should attain two key capabilities:

1. Continuously learn and adapt over time (plasticity).

2. Retain the previously learnt knowledge (stability).

Current advent of digitization of the natural environment using AI systems with integrated computation models are frequently exposed to continuous streams of non-stationary environments, resulting in the necessity of CLL. A major limitation of existing AI systems is the catastrophic forgetting (also known as catastrophic interference), which leads to an abrupt performance decrease or, in the worst case past knowledge being completely overwritten by the new (McCloskey and Cohen, 1989).

In biological brain, the catastrophic forgetting is primarily addressed by the complementary learning systems (as discussed in Section 2.2.4) in which the interplay of dual memory systems, i.e., hippocampus and neocortex, enable concurrent learning of regularities and specifics of the environment. This provides humans the ability to progressively learn over a sustained time span by accommodating novel knowledge while retaining previously learned experiences.

However, it is important to understand the delicate difference between catastrophic forgetting versus transience. The transience of the neocortex endorses the knowledge adaption by selective forgetting, while catastrophic forgetting refers to complete wipe out of the memory in order to accommodate new knowledge. Therefore, the computational models should be equipped with transience property in order to provide an invariant abstraction and adaption of the new knowledge, while addressing the inherent catastrophic forgetting issues.

In essence, this section presented scholarly work leading to the discovery of relationships between the biological intelligence and artificial intelligence. We justified the necessity of biological mimicry in AI systems that can aid in augmentation and expedition of the digital ecosystem to represent the natural environment. On this basis, we identified seven biological bases, that have been essential for the survival and continuance of biological beings. Further,

we discussed these bases in detail with respect to their role and responsibility in the human perception system, what limitations of AI agents can be addressed through each of these bases, and in terms of their importance for the advancement of AI agents. We believe these biological bases have the potential to advance computational models in representing the natural environment in order to satisfactorily perform in digital environments by processing Big Data, deriving insights that ultimately transform into actions and recommendations.

2.4 A Landscape for Digital Representation of Natural Environments

The overall focus of this chapter is to provide a direction of a new thinking in developing AI systems. As such, previous section identified and documented seven key constitutes of biological brain that have the potential to inspire and advance a new breed of AI systems. This section intends to provide a conceptual framework that ties together the neurophysiological inspiration, the features of the big data and digital environment presented in the form of a landscape and propose an overarching conceptual framework as the basis for the research carried out and described in this thesis. This section is at the stage of proposing a landscape that relate pre-identified biological bases to big data environments and computational needs. Overall, this section relates to the module (2.4) *Landscape for AI systems in natural world*, of the chapter overview presented in Fig. 2.1.

This thesis acknowledges the remarkable evolution of biological beings over millions of years that has provided them with ultimate capability to represent the natural environment in order for survival by developing predictions and actions as necessary. In contrast, AI systems, that came into existence only by the latter part of 19^{th} century, are merely capable of representing the natural environment in a fractionalised form, focused on isolated tasks in hand. With the emergence of interconnected artificial sensory devices, AI systems encompass the mere potential to augment and expedite its representation of the natural environment in order to effectively advance their capabilities. However, this requires utmost inspiration and adaption of the biological bases encoded in humans for development of advanced and futuristic AI agents.

This section positions the need for AI in natural ecosystems with its digital counterparts, proposing a landscape for the development of futuristic AI agents underpinning biological bases of human neurophysiological system. This landscape model is developed in order to provide the scope for this thesis and to develop of an architectural overview for AI agents in natural environments. Drawing on basis of the proposed landscape, we later propose a

conceptual framework for AI systems to continuously acquire information, represent its knowledge in digital form, and continuously update the acquired knowledge based on the continuous evolution of natural environment.

The broader view and context in which the thesis is developed is illustrated in the proposed landscape in Fig. 2.7. At the highest level, the foundational basis for this thesis development resides within the natural environment, perceiving the environment through the artificial somatosensory, and providing feedback and actions to the environment through computation models. A representation based separation between the sensory input (i.e., data) and computation models is proposed through the work in this thesis. The separation is modulated through a hierarchy of representation layers: latent representation and cognitive representation.

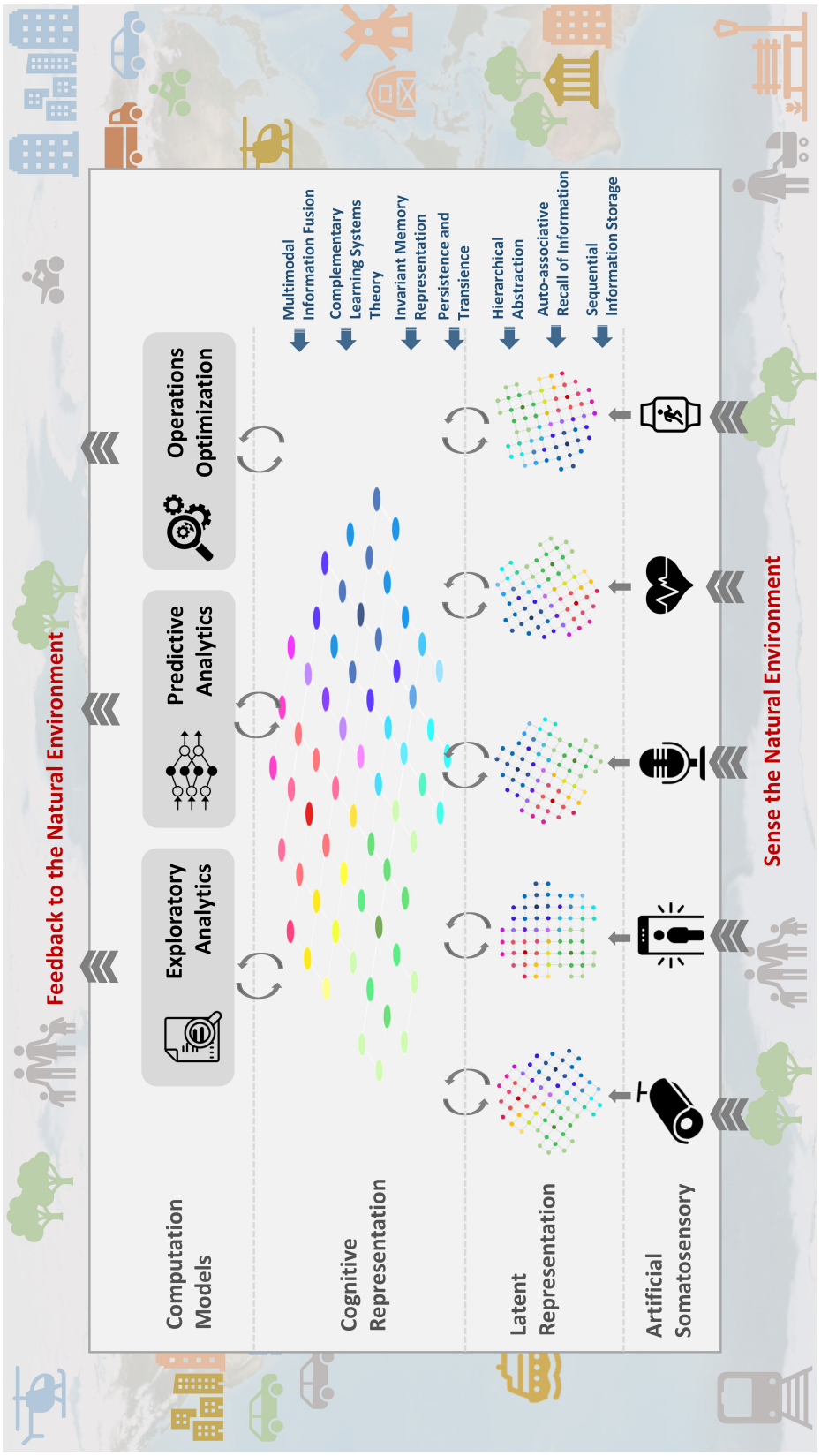


Fig. 2.7 Landscape Model for AI agents in the natural environment.

The proposed landscape model is designed to receive inputs from the natural environment through artificial somatosensory, resembling biological beings that sense the natural environment through their biological somatosensory system inclusive of eyes, nose, ears, tongue and skin. The somatosensory system of artificial counterparts consists of edge devices such as surveillance cameras, microphones, speed detectors, and a diverse range of smart devices. Based on the application domain and context, combinations of these edge devices are to be selected.

In general, the input from artificial somatosensory systems are directed to the computational models through a pre-processing pipeline to analyse and generate insights/action recommendation that should feedback to the environment (Zhang *et al.*, 2019). The pre-processing is manually carried out by a human-in-the-loop. The human, i.e., domain expert or data analyst, will attempt to clean the raw input data received from somatosensory systems, engineer features according to the task at hand and even annotate data with labels before presenting to computational models. Such a mechanism was adequate for traditional environments with small scale structured stationary datasets. However, with the digitization and connected IoT, we are on the verge of moving to an environment with the availability of Big Data and a heavy demand for real-time outcomes. The IoT will generate massive volumes of continuous data streams using number of sensors that provide diverse facets of environment. The complexities arise from these advancements make traditional data pipelines insufficient to perform in such digitized environments, thus, deems a new genre of AI that provide means to advance contemporary needs.

In traditional AI systems, a number of biologically inherited functionality are modulated in either sensory data acquisition modules or computation models (presented in Section 2.3). For instance, sequential information processing, persistence, transience and handling catastrophic forgetting is primarily handled in computational models in current machine learning paradigms. In contrast, biological counterparts often develop a comprehensive invariant representation of the environment through its somatosensory system by the synthesis of inputs from multiple sensory modalities. These invariant representation lies along hierarchically connected representation layers (e.g., Ventral stream, Dorsal stream) and information pass along these representation layers are then utilized for insight generation and decision making. This characterization of biological perception system provides means to advance AI systems. Therefore, inspired by the biological cognitive system, we propose the landscape architecture to decouple data and computation models by proposing intermediate hierarchical layers of representation. In addition, we propose the multitude of essential bases inspired by biological perception system to be modulated by the proposed hierarchical representation layers: latent representation and cognitive representation.

The top most layer of the landscape model consists of computational models developed for specific tasks such as exploratory analytics, predictive analytics, anomaly detection, classification, optimization, simulation, etc. Ultimately, we intend the representation mechanisms we develop to be able to transform and augment the data acquired from artificial somatosensory system to enable aforementioned computational algorithms to be carried out efficiently, ideally without a human-in-the-loop.

2.4.1 Latent Representation

The aim of the Latent Representation (LR) is to provide an unambiguous structure to represent data from a single modality, i.e., single input sensor. Irrespective of the type of data, LR should be able to understand the distribution of the data streamed through the sensor and provide a topographic representation of it. Further, LR is expected to evolve over time along with the changes that occur within the data distribution.

In conventional vocabulary, "latent" refers to "hidden". In machine learning literature, the data points that can be observed in input data space can be mapped into a latent space, in which the similar input data points are positioned in close proximity (Tang *et al.*, 2019). Recent work in representation learning proposed that the success of machine learning generally depends on data representation because of the different representation techniques provide means to combine and modulate diverse explanatory factors of variation behind the data. Developments in unsupervised representation learning have utilized deep learning, casing advances in probabilistic models, autoencoders and manifold learning (Bengio *et al.*, 2013).

This thesis proposes to represent the input space using an invariant representation and provide the ability of sequential information storage for non-stationary streams of input data through the latent space representation. The latent representation is expected to modulate persistence and transience in its representation and to embody efficient mechanisms to suppress catastrophic interference. The objective of modulating these biological bases in LR is to adapt the representation along with continuous change in data distribution over time. That is, when the natural environment evolves overtime with introducing new types of behaviour and new types of data, LR should be able to update and adapt its representation capability and its structure accordingly.

In practice, the sensory input from the artificial somatosensory system will generate separate LRs to learn an effective representation of each of the sensory modality. Followed by the LR, input stimuli are directed to the cognitive representation layer in order to synthesize information from multiple sources.

2.4.2 Cognitive Representation

Human's capability of information fusion, in which ambiguous sensory cues to construct an unambiguous representation of the surrounding is paramount to their survival and successful interaction with natural environments (Stein and Meredith, 1993). This provides a number of advantageous to humans including noise reduction, disambiguation of ambiguities, complementary information leading to a better representation of surroundings, and ultimately provide a holistic representation of the surrounding (natural environment) within the human brain (Jayaratne *et al.*, 2018).

On this basis, we aim to design the Cognitive Representation layer (CR) to resemble the information fusion mechanism that occur in biological perception systems. CR will provide means to fuse and synthesize input stimuli received from multiple modalities (i.e., multiple sensors), developing a holistic representation of the natural environment (or the context in consideration). In addition to information fusion, CR is responsible for continuously acquire, fine-tune and modify its representation adapting to the evolving nature of the environment. Thereby, we expect to embed a complementary memory mechanism that resemble neocortex and hippocampus structures of biological brain in order to continuously acquire knowledge with minimal affect from catastrophic interference. This provides means of continuous lifelong learning based adaptation to its representation as discussed in section 2.3.6.

2.4.3 Computation Models and Feedback

The computation models layer can include diverse computation techniques such as exploratory analytics, predictive modelling, optimization, simulation and machine learning techniques. In contrast to typical machine learning pipelines that are densely depend on the input data and feature engineering, we intend to use the CR and LR to decouple and augment input stimuli prior to be processed by computation models. In addition to make learning algorithms less dependent on feature engineering, the proposed hierarchical representation layers are expected to provide means to learn, identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data. The outcomes of the computation models are ultimately used for generation of insights and decision making, which in turn provide feedback to the natural environment.

2.5 Multi-layered Self-structuring Knowledge Representation Framework

Relating the neurophysiological inspiration, the features of big data and digital environment, Section 2.4 presented the landscape of AI systems to position in natural world providing a high-level overview and scope for the thesis. Drawing upon the landscape, this section proposes an overarching conceptual framework, in which the conceptualisation as well as the design and development of new AI system can be materialized. This section relates to the module (2.5) *Self-Structuring Knowledge Representation Framework*, of the chapter overview presented in Fig. 2.1.

Traditional AI has been built for pre-identified and well-defined problems/domains. However, new digital ecosystems are driven by dynamic and volatile environments with a variety of data sources that generate continuous streams of data. What is being proposed as a new AI framework, should be able to support such environments while being able to continuously learn from new experiences. Thereby, the conceptual framework proposed in this section constitutes of two key components, LR and CR, which cater to the 'unknown' latent and cognitive derivations which will occur during actual situations/events. Both LR and CR should also be able to adapt its representation along with continuous changes in data distribution over time, to resemble evolving external environment. If these two representation layers develop with a pre-defined structure, the knowledge representation will saturate very early, and will be unable to acquire new information while retaining previously learning knowledge intact. This leads to an abrupt performance decrease or, in the worst case, past knowledge being completely overwritten by the new. Therefore, it is not possible or very difficult to predefine the appropriate representation structure for these layers. Taking account of the need for automatically structure the representation, the solution brought by this thesis is Self-Structuring Artificial Intelligence (SSAI) with an unsupervised learning paradigm. SSAI is not only of pivotal importance for the proper development of neural structures and internal representation, but also help bootstrap the emergence of cognitive abilities encoded into representation (Lungarella and Sporns, 2005).

Drawing on the landscape architecture for AI systems, we propose the *Multi-layered Self-structuring Knowledge Representation Framework (MSKRF)*, that consists of four layers: (i) Sensory Inputs, (ii) Latent Representation, (iii) Cognitive Representation, and (iv) Continual Knowledge Acquisition. The key elements of the proposed MSKRF conceptual framework are illustrated in Fig. 2.8, demonstrating the contribution of SSAI and the representation layers providing a foundation for materialization of the framework. We propose the seven biological bases identified in this thesis to be enforced in the development of latent representation and

cognitive representation as delineated in left side pane of conceptual framework diagram. The interplay between latent and cognitive representation layers enable continuous knowledge adaption due to the incremental changes in input stimuli. The essential characteristics from biological perception system should adapted in both representation layers.

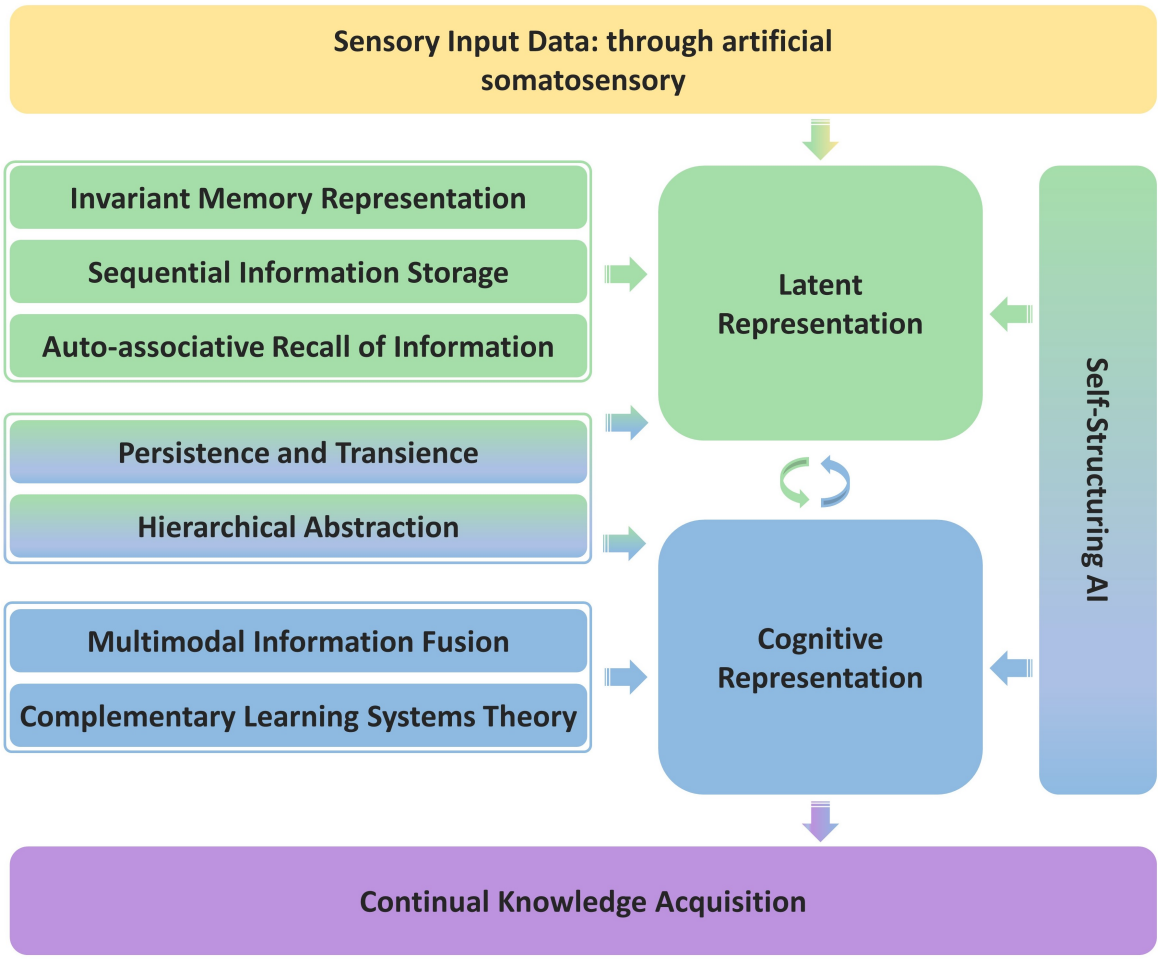


Fig. 2.8 Conceptual Framework for Self-Structuring Knowledge Representation.

The MSKRF originates from obtaining sensory information through artificial somatosensory such as cameras, microphones, smart detectors, etc. Sensory input data module is followed by two representation layers: latent representation (LR) and cognitive representation (CR). The materialization of these representation modules is underpinned by self-structuring capability. Therefore, when the representation modules are materialized as algorithms, we propose to use SSAI as the foundation to facilitate data representation. Additionally, the representation modules should be designed to resemble the biological cortical structure with the capabilities to accommodate the seven biological bases in each layer as presented in Section 2.4.1 and Section 2.4.2.

The MSKRF proposes a bidirectional connectivity between LR and CR. The purpose of LR is to provide an unambiguous structure to represent data from a single modality, while CR fuses and synthesizes the input data captured from multiple modalities. The forward pass of information from LR to CR facilitates to develop of a holistic representation of the natural environment. The backward pass of information from CR to LR facilitates self-structuring of the representation of individual modality influenced by the holistic representation obtained from the fusion.

We formulate the ultimate goal of the MSKRF as to continuously acquire knowledge, update and adapt over time from natural environment, which we denote as continuous Knowledge Acquisition. Traditionally, isolated datasets where people looked at and pre-processes to identify types of data we want to use and then feed to machine learning and data analytics algorithms in order to derive outcomes for decision making. However, with the advancement of data generation, Big Data, IoT, and computational models, more proactive measure needs to be maintained in knowledge acquisition. Thus, we attempt to reinforce lifelong learning concepts using SSAI in order to achieve a more proactive AI paradigm.

The forthcoming chapters are aligned to design, develop, validate and demonstrate the modules discussed in the MSKRF framework. Chapter 3 investigates the background for selection of a candidate algorithm and learning technique for representation modules, in which the theoretical foundation of SSAI is discussed in detail. This development of SSAI core algorithm can be relate to SSAI module of the MSKRF framework. The investigation has resulted in the selection of an SSAI candidate algorithm that has inherent capability to represent knowledge in an invariant representation. This is followed by a practical exploration to demonstrate capabilities of the selected SSAI technique in representing the natural environment for understanding the environment through video data.

Chapter 4 develops a novel algorithm, underpinned by core SSAI principles to accommodate the biological bases discussed in section 2.3 focusing on Latent Representation module of the MSKRF framework. The novel algorithm is developed to overcome current limitations in traditional self-structuring AI: 1) inability to accommodate the temporal nature of the input data, and 2) overfitting and the influence of outdated information on the acquired knowledge. The algorithm development will incorporate five biological bases we identified (out of seven) in Section 2.3, which are:

1. Sequential information storage capability of neocortex.
2. Auto-associative recall of information from neocortex.
3. Hierarchical abstraction in memory storage.
4. Persistence and transience of memory.

5. Multimodal information fusion.

Amalgamation of these bases in the proposed SSAI algorithm leads to the development of a novel algorithm, followed by a series of experiments conducted to demonstrate the validity and usability of the algorithm.

Chapter 5 aims to develop continuous lifelong learning algorithm by extending the novel SSAI algorithm incorporating additional capabilities aligned to Cognitive Representation module and Continuous Knowledge Acquisition module in MSKRF framework. We first attempt to develop a deep learning based cognitive representation model to achieve continuous lifelong learning. A deep autoencoder architecture is utilized for anomaly detection from surveillance video using an active learning approach, which is presented as an experiment. The experiment leads to a discussion of limitations in using deep learning, where the need for unsupervised self-structuring SSAI is highlighted. Developing upon the limitations of the proposed deep learning model, we introduce a novel development of unsupervised CR model using a brain inspired memory formulation and consolidation mechanism, that was proposed as the seventh biological base, i.e., Complementary learning systems theory. The algorithmic development is demonstrated using a series of experiments with benchmark datasets. Chapter 6 demonstrates comprehensive developments of the proposed MSKRF conceptual framework in two application areas providing two different contexts and data environments: a smart city environment and an environment composed of health-care related applications.

In order to experiment and demonstrate the integral capabilities of the MSKRF framework, we utilize video data due to its challenging nature, i.e., the computational complexity and cost of video data processing due to spatial and temporal dimensional structure combined with non-local temporal variations across video frames (Nawaratne *et al.*, 2019c). In addition, we provide a number of case-studies that utilize video data, IoT sensor data and clinical data to demonstrate the wide applicability of the implementation of this conceptual framework.

2.6 Summary and Research Questions Revisited

This chapter proposed a conceptual framework that tied together the neurophysiological inspiration, the features of big data and digital environment, presented in the form of a landscape and proposed an overarching conceptual framework as the basis for the research carried out and described in this thesis. Section 2.1 explored the advancement of digital ecosystems with respect to how AI systems perceive natural environments and identified key limitations in current state-of-the-art AI systems, suggesting the need for a different way of thinking about materializing AI systems. Section 2.2 provided an in-depth review on

structural, functional and behavioural facets of biological brain focusing on visual perception system and memory system, leading to an understanding of the ability of humans to demonstrate complex behaviours, skills whilst having a memory formulation that can continuously learn and adapt.

Section 2.3 evaluated the relationship between biological intelligence and artificial intelligence, proposing seven key constituents in human neuronal system that has the potential to aid its artificial counterpart to advance its capabilities in perceiving and representing natural environment in order to process Big Data, derive insights that can be transformed to actions and recommendations. Drawing on the identified biological bases, we positioned the development of AI systems in natural ecosystems with its digital counterparts, proposing a landscape model. The landscape model leads to conceptualization of a knowledge representation framework for AI systems to continuously acquire information, represent its knowledge in a digital form, and continuously update the acquired knowledge based on continuous evolution of natural environment. We set the ulterior objective of the proposed conceptual framework as continuous knowledge acquisition, whilst having self-structuring knowledge representation as the foundation of learning.

This chapter provided the scope for development of this thesis, whilst addressing the first research question **(RQ1): How can computational continual lifelong learning enable the natural world to be represented digitally, making use of continuous streams of data from a variety of digital sensors? What aspects of the structural and functional facets of neurophysiological studies can be used as a foundation premise to develop techniques for computational continual lifelong learning in data intensive digital environments?** As stated in section 1.4, RQ1 is decomposed into four sub-questions. The sub-questions were addressed in this chapter as follows:

RQ 1.1) How has the new digital world, made up of Big Data and IoT, transformed AI systems in perceiving natural environments, acquiring and updating knowledge for past and future tasks? This question is addressed by the Section 2.1 that explored the advancement of digital ecosystems with respect to how AI systems perceive natural environments and identified key limitations in current state-of-the-art AI systems, suggesting the need for new thinking about conceptualizing AI systems.

RQ 1.2) What are the core structural components and functional mechanisms in the human neurophysiological system that support the continual lifelong learning in humans? This question is addressed in section 2.2 through section 2.3 by reviewing the key functions and structural formulations of the biological brain. In light of the comprehensive

introduction to the biological brain, the thesis presents seven biological bases that have been essential for the survival and continuance of humans.

RQ 1.3) How can these neurophysiological facets of humans be used to inspire artificial representation of natural world in data intensive digital environments? The knowledge is merely a representation of the natural environment in the human brain. Section 2.2.1 provides the mechanism on how the knowledge is stored in the human memory, based on a set of cognitive theories from the literature.

RQ 1.4) How can such neurophysiological inspiration be used to combine the features of big data and digital environment to form an overarching conceptual framework? We base the identified biological bases and a set of cognitive theories to develop a conceptual framework (MSKRF) to model the landscape for AI agents in representing the natural world. The landscape is proposed in section 2.4 while the conceptual framework is presented in section 2.5.

The next chapter relates to the selection of a candidate algorithm and learning technique for the latent representations, in which the theoretical foundation of SSAI is discussed in detail.

Chapter 3

Self-Structuring AI to Facilitate Representation Learning

Chapter 2 brought together the neurophysiological inspiration, the features of big data and digital environment in the form of a landscape and proposed an overarching conceptual framework, *Multi-layered Self-structuring Knowledge Representation Framework* (MSKRF), as the research base for this thesis. MSKRF has the potential capability to form an AI based knowledge representation that capture continuously evolving environmental stimulus and adapt its knowledge representation accordingly, satisfying the overall objective of continuous lifelong learning.

The intermediate knowledge representation mechanism of the MSKRF framework lies in the two hierarchically connected layers: latent (LR) and cognitive (CR) representations that provide the foundation for learning from input by representing sensometry input stimuli from artificial somatosensory in a digital form. This chapter aims to explore possibilities in existing computational paradigms to select the most suitable computational model to facilitate the two representation modules.

New digital ecosystems are driven by dynamic and volatile environments with a variety of data sources that generate continuous streams of data. What is proposed as the knowledge representation framework has to support such environments while also being able to continuously learn from new experiences. The two key components of knowledge representation, LR and CR, will cater to the 'unknown' latent and cognitive derivations which will occur during actual situations/events. Hence, LR and CR should also be able to adapt and automatically structure its representation along with continual changes in data distribution over time to resemble the evolving external environment. If these two representation layers are developed with a pre-defined structure, they will be unable to acquire new information while retaining past learning. This will lead to abrupt performance decrease or, in the worst case, past knowl-

edge being completely overwritten by the new. Chapter 2 proposed Self-Structuring Artificial Intelligence (SSAI) with an unsupervised learning paradigm as a solution to overcome these challenges. SSAI is not only of pivotal importance for development of neural structures and internal representation, but also will help bootstrap the emergence of cognitive abilities encoded into representations (Lungarella and Sporns, 2005).

Chapter 3, 4 and 5 in combination design, develop and evaluate the MSKRF conceptualized in chapter 2. Chapter 3 investigates viable algorithmic paradigms and computational mechanisms to provide an architectural base for the MSKRF framework, founded upon SSAI. Chapter 4 develops an algorithm for LR by incorporating the biological bases discussed in section 2.3 with SSAI as the foundation. Chapter 5 proposes the CR module of MSKRF by formulating these algorithms as complementary learning systems to achieve the overall objective of lifelong learning.

A key contribution of this thesis is to bring to fruition the proposed MSKRF framework from initial conceptualization to a practical and usable AI technology for real applications. The two hierarchically connected LR and CR are major components of this framework, thus, it is imperative that we adapt the most viable learning paradigm to lay the foundation for these representation mechanisms. Thereby, this chapter aims to investigate both biological and natural phenomenon that governs representation in natural environments in order to provide an architectural basis for MSKRF. The subdivision of the chapter to achieve this aim is presented in Fig. 3.1.

First, we provide a theoretical foundation by investigating the feasibility of modeling the indeterministic natural environment using machine learning paradigms to facilitate representation learning in Section 3.1. The investigation narrows down to unsupervised self-organization as a viable prospect. We examine both natural and biological prospects that use self-organization in Section 3.2. Here we identify experience-driven self-organization is a key constituent in biological brains that enable humans to continuously acquire knowledge during their lifetime.

In Section 3.3, we delve into the foundation of SSAI from biological perspective to understand how natural self-organization is facilitated. This investigation resulted in evidence to support the argument, self-structuring is a key facilitator to enable self-organization. We then look into computational formulations of SSAI exploring literature on existing SSAI techniques, their limitations and prospects in Section 3.4 and select a viable candidate SSAI base algorithm for the proposed MSKRF and demonstrated via Smart City application focusing on intelligent video surveillance in Section 3.5.

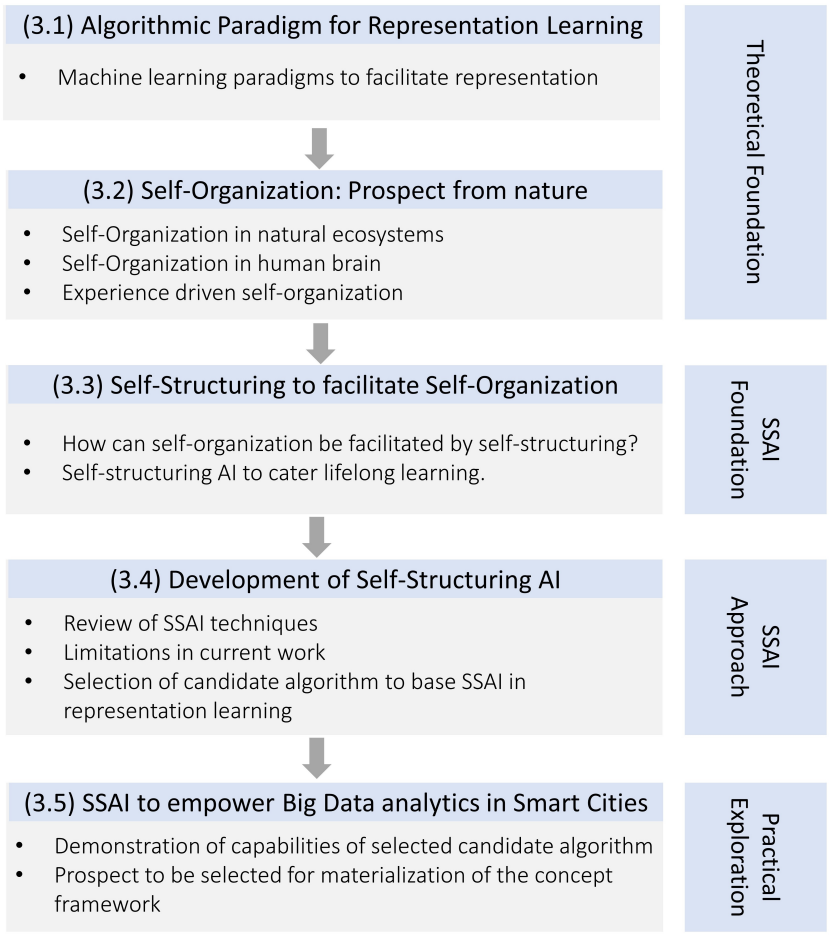


Fig. 3.1 Chapter Overview

3.1 Prospect for Representation Learning

Based on the theoretical foundation of modelling indeterministic natural environments using computational models, this section investigates machine learning paradigms that facilitates representation learning, relating to the module (3.1) *Algorithmic Paradigm for Representation Learning* as presented in chapter overview in Fig. 3.1. Given the importance of representation for effectiveness of AI systems, it is vital to select an appropriate algorithm for the development of representation mechanism from its ideation. The widely used prospects in algorithm developments are: i) statistical modelling, ii) supervised learning, iii) unsupervised learning, and iv) reinforcement learning. The former relates to the development of mathematically-formalized mechanism to approximate the data and optionally to make inferences from this approximation. The supervised and unsupervised machine learning paradigms relate to representations of learning based on the experiences derived from sensometry inputs.

Reinforcement Learning is a machine learning paradigm that relates to taking actions in an environment in order to maximize the notion of cumulative reward.

Inference generally refers to the development of a formalized understanding or test a hypothesis about how a system behaves based on data, whereas prediction aims at forecasting unobserved outcomes or future behavior. Statistical methods have a long-standing focus on such inference problems, as they tend to create and fit a data-specific probability model for inference upon the system. Such statistical model allows to compute a quantitative measure of confidence that a discovered relationship describes as a 'true' effect that is unlikely to result from noise. If enough data are available, explicitly verified assumptions can be made about the data/system (e.g., equal variance) and refine the specified model if needed. In general, the statistical methods have been designed and used with smaller datasets with a few attributes particularly for inference problems (Bzdok *et al.*, 2018). In contrast, machine learning paradigms concentrate on prediction by using general-purpose learning algorithms to find patterns in rich and unwieldy data (Bzdok, 2017; Bzdok *et al.*, 2017).

The nature is unpredictable if not indeterminate. Cziko (1989) argues the complexity of nature based on multiple facets including individual differences in entities, chaos theory, the evolutionary nature of natural entities, the role of consciousness, free will in human behavior and the implications of quantum mechanics relating to the feasibility of modeling the natural environment (Boccaletti *et al.*, 2000). The new digital world has enabled to capture this indeterminate nature using Big Data. Thereby, it is not practical to develop statistical models to represent the sensometry stimuli from the natural environment, and as such developing on the foundations of machine learning paradigms can be thought suitable as they merely derive the representation from previous experience, similar to how humans perceive nature and learn from it.

The machine learning based representation mainly constitutes of three learning paradigms: supervised learning, unsupervised learning and reinforcement learning. Supervised learning algorithms infer a function from a set of training examples. The aim is to approximate a mapping function, such that given a new input, the mapping function to be able to infer (or predict) the output. In contrast, unsupervised learning aims to find hidden structures in unllabeled data. The main difference between supervised and unsupervised machine learning paradigms are the availability of labelled data or ground truth. The supervised learning requires a prior knowledge of what the output values for the training samples should be, while unsupervised learning does not.

Reinforcement learning (RL), in contrast, is the training of machine learning models to make a sequence of decisions (Neftci and Averbeck, 2019). The AI agent learns to achieve a goal in an uncertain, potentially complex environment by employing a trial and error

strategy to come up with a solution to the problem. In the context of representation learning, existing literature widely utilize supervised or unsupervised representations to accommodate RL, instead of using RL as a mechanism to develop a representation (Stooke *et al.*, 2020; Madjiheurem and Toni, 2019; Zhu, 2020).

Humans usually self-learn via exposure and observation to develop a representation of the natural world. For instance, a human child does not require a large number of images of dogs for him to recognize a new picture of a dog. As a toddler, after a few examples of dogs, they learn to differentiate in great detail. Additionally, considering practical scenarios, finding labelled data is unrealistic and generating such labelled data is time-consuming and expensive. Further, AI systems of the future is expected to incorporate higher degrees of unsupervised learning in order to generate value from unllabeled data (Nawaratne *et al.*, 2017). This is the ability we intend to build into the proposed representation architecture thus, we select unsupervised learning paradigm in the creation of representation layers in MSKRF.

In unsupervised representation learning, self-organization can be identified a natural phenomenon, which has been computationally recreated to achieve unsupervised learning that resemble both biological brain and natural phenomenon. Self-organization can be understood as a process where a form of overall order arises from local interactions between parts of an initially disordered system. Self-organization process is spontaneous when sufficient energy is available, without any need for control by any external agents. The self-organization provides the means in chaos theory in terms of islands of predictability in a sea of chaotic unpredictability (Schweitzer, 1997; Khadartsev and Eskov, 2014). On this basis, this thesis intends to build the representation development on the foundation of self-organization.

3.2 Self-organization: Prospect from Nature

Self-Organization is the formation of order and patterns in a system by internal processes without being constrained or forced by external stimulus (Green *et al.*, 2008). The notion of self-organization has been more robustly articulated since its inclusion within the complexity framework, however, it has been a pervasive idea in scientific thought with subject of inquiry in different fields of knowledge: physical sciences, chemical sciences, biology, computational methods, environmental science, economics and cognitive systems (Skår, 2003; Anzola *et al.*, 2017; Capra, 1996). The current section presents an overview of Self-Organization from a natural prospect, related to the module (3.2) *Self-organization: Prospect from Nature* as presented in chapter overview in Fig. 3.1.

3.2.1 Self-organization in Ecosystems

Self-organization theory holds that the distinctive characteristics of living systems require explanation in terms of life's emergent properties at different levels of organizing principles expressed at different levels of biological organization; from cells through whole organisms to communities and ecosystems. For instance, consider a plant distribution on a mountainside. The ecosystem is constrained by the cold, limiting the altitude at which plant species can grow. The cold acts as an external constraint. Simultaneously, competition for resources and land sites leads to self-organization within the plants community by truncating the range of altitudes where plant species can grow as illustrated in Fig. 3.2.

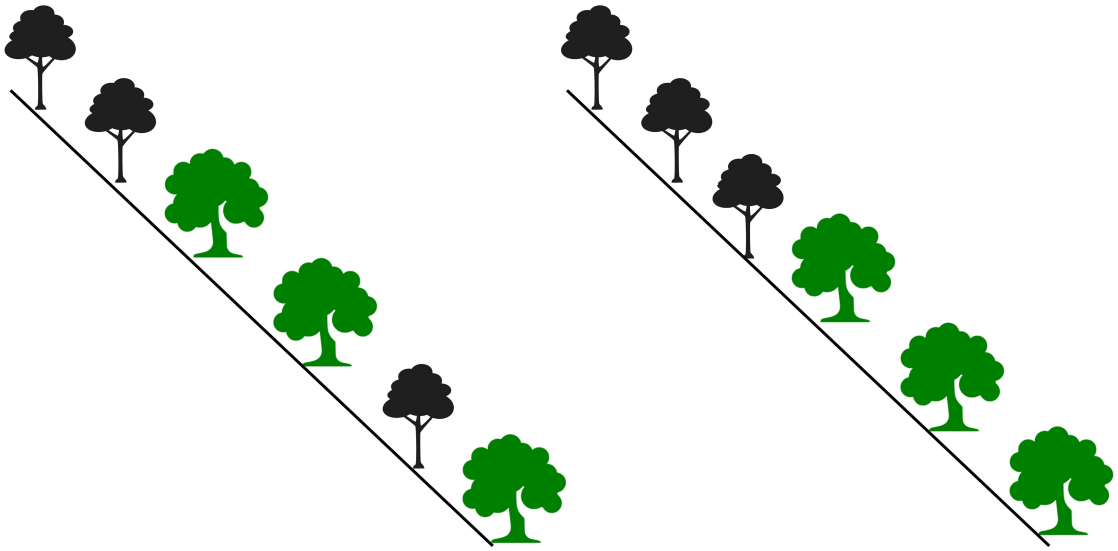


Fig. 3.2 Effect of competition on plant distributions on a gradient.

In nature, the phenomenon of self-organization can be observed universally, e.g., flocking behaviour of birds and fish, an ant colony. Similarly, the self-organization behaviour can be observed in many aspects in business studies, social sciences, medicine and technology, e.g., liquid crystallization, spontaneous folding of proteins, organization of market economy (Krugman and Krugman, 1996; Camazine *et al.*, 2003; Dimitrov, 2003). Self-organization can further be noted in terms of interaction between components of a system and associated with the ideology of complexity and emergence, as captured by the popular saying "the whole is greater than the sum of its parts.", features of a system emerge out of interaction within the components due to the process of self-organization.

3.2.2 Self-organization in Human Brain

Apart from the self-organization that occur in natural ecosystem, human brain can be sourced as a prominent domain in which the self-organization theory is encompassed in its foundation. As discussed in section 2.2.1, the human brain can be identified as one of the most complex systems known to humankind, that contains over 100 Billion strongly connected neuronal cells (Bassett and Gazzaniga, 2011). For instance, a single neuron can have more than 10,000 connections to other neurons. The human brain serves in processing cognitive functions, and clinical evidence and numerous results from animal experimentation indicate that these cognitive functions have to be learned (Singer, 1986). This arouses the question: *What steers the numerous neurons that encompass human brain so that it can serve cognitive functions and produce macroscopic phenomena such as coherent steering of muscles in locomotion, grasping, vision, in particular to pattern recognition and decision making?*

The brain structures that serve these cognitive functions require sensory experience for their maturation. Singer (1986) proposed that self-organization processes are implemented in mammalian brain structure in order to optimize genetically determined blueprints of neuronal connectivity. Thus, neuronal activity becomes an important shaping factor in the development of structural and functional architecture of mammalian brain. To this extent, neuronal activity is modulated by sensory signals and environmental factors that influence the development of brain neuronal networks.

Further, David Kahn proposed a theory that suggests self-organization modulates human brain to cope in the newly changing environments through dreaming. A different facet of the *self* is created as a result of a self-organizing process in the brain, through dreaming. Self-organization in biological systems often occur as means to cope with an environmental change that existing systems cannot cope. In dreaming, self-organization serves the function of organizing disparate memories into a dream since the dreamer herself is not able to control how individual memories become weaved into a dream. The self-organized dreams provide a wider repertoire of experiences. This expanded repertoire of experience results in an expansion of the self beyond that obtainable when awake (Kahn, 2013).

On the grounds that self-organization plays a vital role in both ecological and biological settings, we propose that self-organization should be utilized as a foundation in the development of the representation mechanism in this thesis. Thereby, a further investigation on diverse aspects and essential characteristics of self-organization in biological brain is assessed in following sections to inspire the development of the representation mechanism proposed in MSKRF.

3.2.3 Experience-Driven Self-Organization

Based on a number of clinical experiments using animals, Singer (1986) inferred that experience-driven self-organization is a key component of human cognitive ability, and should not be considered as a passive imprinting process but rather as an active dialogue between the brain and its environment. Experience-driven development plays a crucial role in human brain with topographic maps being a common feature of the neocortex structure of human brain for processing sensory inputs (Nelson, 2000).

Shatz (1992) demonstrated that the development of topological maps in biological brain happens through self-organization of the input connections from the thalamus and are shaped by visual experiences. He proposed that the brain wires itself as the fetus develops, similar to how a computer is manufactured: the chips and components are assembled and connected according to a pre-set circuit diagram. According to this analogy, a flip of a biological switch at some point in prenatal life turns on the computer. This notion resembles the cortical structure of the fetal brain is recorded as a biological blueprint, presumably DNA, and the organs will commence its function only once the wiring is completed. Thereby, it has been hypothesized that while intrinsic factors such as DNA or genes drive initial development of biological brain, extrinsic factors such as somatosensory experience based self-organization provides mean to development of the brain to achieve higher complexity and performance throughout lifetime (Sur and Leamey, 2001; Hirsch and Spinelli, 1970; Parisi, 2017).

While self-organization forms the foundation of knowledge acquisition in human brain, there exists a number of phenomena that enable smooth and seamless integration of knowledge in the memory. Stability and Plasticity in biological brain is an important aspect that needs to be considered and addressed with regards to self-organization. The following subsections provide introduction to the stability-plasticity phenomenon.

3.2.4 The Stability-Plasticity Dilemma

Humans have the ability to adapt to the environment by effectively acquiring novel information, refining knowledge on the basis of new experiences and transfer the consolidated knowledge across multiple domains. It is acknowledged that human do tend to gradually forget previously learnt information throughout their lifespan, however, rarely does the learning of novel information catastrophically disrupt consolidated knowledge (McCloskey and Cohen, 1989; French, 1999). This ability of learning without catastrophically forgetting in human brain is mediated by a rich set of neurophysiological principles that regulate the stability-plasticity balance of respective brain areas that are responsible for the development of human cognitive system (Lewkowicz, 2014).

The stability-plasticity dilemma is a widely recognised phenomenon in artificial neural networks that have been inspired by biological neural systems. The basic notion is that learning requires plasticity for the integration of new knowledge, and also stability in order to prevent the disruption of previous knowledge (Mermillod *et al.*, 2013). High plasticity will result in previously encoded data being constantly erased, whereas high stability will impede new information being integrated with existing knowledge.

Neurosynaptic plasticity provides means for the human brain yielding physical changes in its neural structure to allow humans to learn, remember and adapt to dynamic environments. The neural structure in biological brain is particularly plastic during critical periods of early development in which neural network acquire their imminent structure driven by sensorimotor experiences (Shatz, 1992; Parisi *et al.*, 2019). The plasticity becomes less prominent as the biological system stabilizes through stages of development. Thus, moving along the development stages, the neural structure preserves only a limited degree of plasticity for its adaptation and reorganisation at a reduced level (Ikezu and Gendelman, 2016; Quadrato *et al.*, 2014). Studies on development of human cortex have demonstrated a consistent tendency to decrease levels of plasticity with increasing age of humans (Uylings, 2006).

The stability-plasticity dilemma regards the extent to which biological and artificial systems must be prone to integrate and adapt new knowledge, while the extent in which the adaptation process should be compensated by internal mechanisms that stabilize neural activity in order to prevent catastrophic interference with consolidated knowledge (Uylings, 2006).

3.3 Self-Structuring to Facilitate Self-Organization

This section provides a theoretical foundation on how the self-organization can be facilitated through self-structuring in computational models in order to achieve continuous lifelong learning by looking through biological organizations and existing computational models. We relate this section to the module (3.3) *Self-Structuring to Facilitate Self-Organization* as presented in chapter overview in Fig. 3.1.

Effective sensorimotor representation is not only crucial for the proper development of biological neocortical structures, in fact pivotal to bootstrap cognitive abilities. As such, the presence of an informational structure in sensory data is vital for this early development of biological neural structure. Lungarella and Sporns (2005) argues that statistical structure of information in sensorimotor experience may be partly based on an effectively coordinated motor activity, and such self-generated motor activity may have a powerful influence in shaping the informational structure ensuring high quality sensory information. Based on the

premise by Lungarella and Sporns (2005), embodiment plays an important role as necessary to complement neural information processing, in which the biological neuronal system attends to and processes streams of sensory stimulation, ultimately generating sequences of motor actions that guide the production and selection of sensory information. Thus, it is apparent that information structuring by motor activity and information processing by neural systems are parallel functions that are linked through sensorimotor loops.

This phenomenon of biological self-structuring to supplement neural learning has been worked on multiple research directions. In the midst of 19^{th} century, Craik (1952) developed the idea that primary function of higher cognitive areas is to develop a symbolic working model based on the associative structure of objects and events in the environment. His view expresses the environmental regularities as cognitive maps. Statistical regularities of the environment are argued to be important for learning, memory, intelligence, inductive inference, and in fact, for any area of cognitive related studies where the information processing mechanism of the brain promotes survival by exploiting them (Barlow, 2001). The ideology behind these inferences is that sensory systems are not only optimized for processing and effectively coding environmental stimuli, further they are well adapted to the statistical structure of the natural environment (Lungarella and Sporns, 2005).

The structural representation of the environment in biological cognitive system is thoroughly explored in attempts to understand the representation of images by mammalian visual perception system considering statistics of images from the natural environment such as trees, rocks, bushes etc. (Field, 1987). Various coding schemes are compared in relation to how they represent information in such natural images, and results support Barlow's theory that the goal of natural vision is to represent information in the natural environment with minimal to non-redundant form (Barlow, 1961, 1979, 1983).

This phenomenon in biological brain has been further explored in the development of AI systems. For instance, Betsch *et al.* (2004) used second-order statistics and wavelets to analyze videos of natural scenes recorded by a camera attached to a cat's head, where the cat explored several outdoor environments and videos of natural stimuli. The videos were recorded from the animals' perspective. The results demonstrated an enhanced occurrence of horizontal orientations compared to average orientations, different mechanisms of fixation point selection as compared to humans, and faster changes in natural stimuli of contour positions compared to orientations. Application of information theoretical measures have been attempted to quantify data streams obtained from robotic systems, and these studies have demonstrated a clear distinction of sensorimotor strategies that can lead to variation of statistical structures in visual input streams (Spoms and Pegors, 2003; Tarapore *et al.*,

2006; Sporns and Pegors, 2004). This has led to the recognition of wide variations in neural structure shown in primary visual cortex (V1) (Webber, 1991).

On this basis, the sensory experience of artificial embodied systems should hold capabilities to effectively coordinate sensorimotor activities through representation, thereby, in need of algorithms and computational models developed in representing environment that can self-structure based on the sensory input. Drawing on the previous discussion (Section 3.2), we note the need for self-organization in artificial memory structures for effective representation of the environment. This thesis argues that the computational bases for representation should also be equipped for self-organization. Through the evolution of millions of years, the biological brain has been gradually adapted to achieve self-organization through self-structuring capability by having a flexible cortical structure. However, most computational bases are restricted for narrow and focused applications making them restrictive and rigid in their structure. Thereby, in order to achieve self-organization in computational bases, it is imperative to develop algorithms that have more fluidity and flexibility that are better suited to replicate the forces of nature.

On this account, we propose the creation of representation modules (latent and cognitive) in the conceptual framework (MSKRF) should contain self-structuring as its foundation. Subsequently, self-structuring foundation would lead the AI system to self-organize based on the environmental inputs, providing an effective representation of the environment.

3.4 Computational Models of Self-Organization

Based on the previous discussion, we identified the importance and need of using unsupervised machine learning paradigm with self-organization for the development of a representation mechanism. As such, an effective computational model to design and develop latent and cognitive representations of MSKRF should be able to self-structure based on the sensory inputs (input data). This section focuses on investigating SSAI computational models that enable continual lifelong learning in machine learning literature, leading to the selection of a viable candidate for MSKRF. We relate this section to the module (3.4) *Development of SSAI* as presented in chapter overview in Fig. 3.1.

A number of computational models have been developed that can demonstrate self-structuring through statistical learning with a nonlinear approximation of the distribution of the input data. Early models that implement biologically inspired self-organization have focused on simplified statistical and mathematical models that govern the formation of topographic maps (Parisi, 2017). Later, a range of artificial self-organizing neural networks have been proposed to resemble the dynamics of biological cortical system such as neural

plasticity and Hebbian learning (Hebb, 1949). Most these neural networks founded upon the Hebbian learning as the basis of unsupervised learning, inspired by the neural weight adjustment mechanism in biological cortical system. The Hebbian theory is often summarized as "Cells that fire together wire together", emphasizing that any two cells or systems of cells that are repeatedly activate at the same time will tend to become *associated*, thereby, that activity in one facilitates activity in the other (Lowel and Singer, 1992).

This section aims to provide an overview of artificial neural network based computational models that have contributed to better understand the underpinning neural mechanism for the development of cortical organization for AI systems. The core ideology of self-organizing neural networks is to make different regions of the neural network to respond similarly to certain input samples starting from an unorganized (or random) state. In a typical self-organization based neural network, training phase composites of developing an organized topographic map through competition and correlative learning process such that a set of neurons represent prototype vectors encoding a submanifold in the input space, thereby learning a significant topological relation of the input space without supervision, i.e., without any labelled input samples (Kohonen, 1990; Parisi, 2017).

3.4.1 Self-Organizing Feature Maps

Teuvo Kohenen's Self-Organizing Feature Maps (abbreviated as SOFM or SOM) is a human cerebral cortex inspired neural network that produce a nonlinear high-dimensional input space into a reduced dimensional discretized representation, while preserving the topological relationship in the input space (Kohonen, 1982, 1990). On the basis of Hebbian Learning (Hebb, 1949), competition and correlative learning (Webber, 1991), as input signals are presented, neurons compete amongst for ownership of the input and the winner strengthens its relationships with this input.

Prior research suggests that the paradigm of competitive learning might constitute a viable mechanism based on the fact that the response properties of the cells of the visual cortex could develop to form coding units suitable for representing the visual stimuli encountered in natural life (Webber, 1991). Thereby, the competitive learning can be based as suitable in representing a self-structuring foundation in computation models, which in turn makes SOM a viable candidate for the development of representation modules in the proposed conceptual framework (MSKRF).

SOM consists of a lattice of competitive neurons connected to adjacent neurons by a neighborhood relation. Initially the neural network is mapped as a lattice of n neurons each having a weight vector, $W_k, (k = 1, 2, ...n)$, representing the input space. The coordinate systems in the SOM represents the output space. The number of neurons (n) should be

predefined. In the training phase, each input vector of a given dataset D, $x_i, (i = 1, 2, ...|D|)$, is presented to the neural network to calculate the Best Matching Unit (BMU) based on the distance between the input and the weight vectors of the neurons. Usually the Euclidean distance (Danielsson, 1980) is selected and the distance is denoted as the quantization error E (Behrooz and Behzad, 1993), which is given by equation 3.1.

$$E_i = ||W_k - x_i|| \tag{3.1}$$

Thereby, the neuron with the lowest quantization error is selected as the BMU (the winning node).

$$BMU(x_i) = argmin_j(E_i) \tag{3.2}$$

Subsequently, the weights of the winner node and its neighbours are updated as equation 3.3, where $w_k(t+1)$ is the updated value of the weight vector of the k^{th} neuron while $w_k(t)$ is the previous weight value of the same neuron. The η_t and $h_b(t)$ represents the learning rate and Gaussian neighborhood function respectively. The learning rate η_t decays over time between [0, 1].

$$w_k(t+1) = w_k(t) + \eta_t \cdot h_b(t) \cdot [x_i - w_k(t)] \tag{3.3}$$

Typically, the input data are presented to the SOM for a predefined number of times, which is denoted as training iterations.

SOM algorithm has a fixed topology, where the dimensions of the reduced topographic representation needs to be defined in advance. SOM is typically employed for exploratory analytical tasks, in which a little or no information about the data is presented upfront. It is often known only at the completion of the exploratory analysis that a different sized SOM would have been more appropriate for the application. Therefore, multiple simulations have to be run for different sized networks to pick the optimum network (Alahakoon *et al.*, 2000). Thereby, having to predefine the topological structure would constrain the mapping accuracy. A solution to this problem would be to determine the structure as well as the size of the network during the training of the network.

3.4.2 Growing Self-Organizing Computational Models

The benefits of self-structuring neural networks have intrigued the interest of researchers in the recent past, where many have proposed new neural architectures both in supervised and unsupervised paradigms. Having SOM as the underlying concept and the foundation, more

complex and dynamic algorithms have been presented, where the main advantage lies on their ability to grow structure to represent the input data space better in contrast to their fixed structure counterparts.

Fritzke proposed the Growing Cell Structures (GCS) algorithm that is based on SOM, but replacing the basic two-dimensional lattice structure by a network of nodes whose connectivity defines a system of triangles (Fritzke, 1991, 1994). Starting with a triangle of cells at random positions in R^n, the triangle is distributed over the area of non-zero probability in the space. Heuristics are used for insertion and deletion of nodes and connections to the network (Fritzke, 1994). GCS results in a graph network in the structure $G = (V, E)$, where V is the set of nodes and E is the set of edges connecting the nodes. The GCS works well with relatively low-dimensional data however, mapping cannot be guaranteed to be planer for high-dimensional data, limiting the power of visualization for exploratory analysis and simulation.

Neural Gas (NG) algorithm, proposed by Martinetz *et al.* (1991), starts with no connections and a fixed number of units floating in the input vector space. NG uses a similar competitive learning as used by SOM, however, network connections are generated between winner units and its closest competitor. NG has a fixed number of units that needed to be predefined, thus, there exists similar drawbacks as SOM. As an extension to NG, Fritzke (1995) proposed an incremental extension of SOM and NG, named Growing Neural Gas (GNG). Unlike SOM and NG, this model has no parameters which change over time and is able to continue learning, adding units and connections, until a performance criterion has been met.

The GNG network starts with a set of two neurons in the input space. At each training iteration, network connections are generated (if not exists already) between winner units and its closest competitor. The weights of the network nodes are updated based on the quantization error between winner node and input vector. If the number of given inputs is a multiplication of a predefined parameter λ, a new neuron is created halfway between those two neurons that have maximum accumulated error (Fritzke, 1995). The GNG algorithm consists of an age based node removal mechanism to remove rarely used connections and neurons without connections. The growth of GNG network stops when a predefined criterion is met, i.e., some performance measure, network size, or a given number of training epochs (Parisi, 2017).

The GNG generates new neurons at a fixed growth rate, i.e., new nodes are added only when the number of iterations is an integer multiple of some predefined constant, λ. Solution to this limitation is extending the self-structuring in a way such that the learning algorithm can add nodes whenever the network in its current state does not sufficiently match the input. In this way the network grows very quickly when new data is presented, but stops growing

once the network has matched the data. On this basis, Marsland *et al.* (2002) proposed Grow When Required (GWR) algorithm, that generate new neurons whenever the network in its current state does not sufficiently match the input.

As a criterion for neural growth, the GWR training algorithm considers the amount of network activation at time t computed as a function of the distance between the current input $x(t)$ and its BMU, where the weight of the BMU is w_b. The activation $a(t)$ as define by:

$$a(t) = exp(-||x(t) - w_b||) \tag{3.4}$$

In addition, Marsland extended the learning of the GWR algorithm using biological inspired habituation. Habituation is a form of learning in which a response to a stimulus decreases after prolonged presentations of that stimulus (Groves and Thompson, 1970). Thereby, the GWR algorithm accounts the number of times that a neuron has fired so that recently created neurons are properly trained before creating new ones (Marsland *et al.*, 2002). GWR is implemented with a firing counter $\theta \in [0,1]$ to express how frequently a neuron has fired. θ is updated based on the efficacy of a habituating synapse reduces over time, and the value of θ is given by,

$$\theta(t) = \theta_0 - \frac{S(t)}{\alpha} \cdot (1 - exp(\frac{-\alpha_t}{\tau})) \tag{3.5}$$

where θ_0 is the resting value, $S(t)$ is the stimulus strength, and α and τ are constants that control the behavior of the curve.

The use of an activation threshold and firing counters to modulate growth of the network lead to generation of a larger number of neurons at early stages of the training and then tune weights of existing neurons through subsequent training epochs (Parisi, 2017). Nonetheless, the network graph structure of GWR depends on locality of input data. Therefore, the network can develop different dimensionality for different regions of the network, which can result in visualization difficulties and inability to focus on the representation space in different granular levels, i.e., zoom out of the network to visualize global patters and zoom in to visualize granular patters in the representation space, that are useful for data mining by identifying clusters in the data.

3.4.3 Growing Self-Organizing Maps

Drawing on the base of cortical self-structuring, Alahakoon *et al.* (2000) proposed an improved growing variant of self-organization, Growing Self-Organizing Maps (GSOM), introducing a map topography that self-structure by adapting its size and shape. A natural

topology of the data space is ensured due to the unconstrained learning and node growth in the GSOM.

The GSOM network structure starts with a minimal number of neurons (typically, four neurons) and grows on boundary based on heuristics and input representation. GSOM consists of two phases; first, the *growing phase* where the neural network grows new neurons and adjusts neuronal weights to sufficiently represent the input space, and second, the *smoothing phase* in which the neuron weights are fine tuned.

In the growing phase, once the input vectors, x_i, are presented to the neural network over a number of cycles (t), best matching neurons (BMU) are calculated using a distance measure, e.g., Euclidean distance. The BMU selection can be computed similar to that of SOM as explained in equations (3.1) and (3.2). Subsequently, the weights of the BMU and its neighbouring nodes are updated using equation 3.3.

The learning rate η_t is a decreasing function of time between $[0,1]$ as shown in equation 3.6, where R is the learning rate update constant and ϕ_t is the size of the neuronal map at time t.

$$\eta_t = \alpha \times \left(1 - \frac{R}{|\phi_t|}\right) \times \eta_{t-1} \tag{3.6}$$

α is an exponentially decaying function, which we define as below, where η_0 is the initial learning rate and T is the total number of growing iterations.

$$\alpha = e^{\frac{-t \times \ln(\eta_0)}{T}} \tag{3.7}$$

Over the number of training iterations, the BMUs accumulate quantization error based on the distance between the input and its weight vector. A neuron is said to under represent the input space once the accumulated quantization error of the neuron is greater than the *Growth Threshold* (GT), as defined below;

$$GT = -D \times \ln(SF) \tag{3.8}$$

The GT is determined by the number of dimensions D in the input space, and novel introduced parameter named the *Spread Factor* (SF), where $SF \in [0,1]$. The SF can be utilized to control the spread of the network structure independent of the dimensionality of the dataset.

In the case of the BMU is under-representing the input space, i.e., accumulated quantization error is greater that GT, new nodes are inserted into the neural network to sufficiently represent the input space. If the BMU is on the boundary, the map is grown from boundary by adding new neurons to the map, as illustrated in Fig. 3.3 (a), (b) and (c). Otherwise, the error

Δe_i is spread among neighbouring neurons using equation 3.9, where *factor of distribution* (FD) is the parameter to regulate error distribution. The newly added neuron is initialized to match the weights of the existing neighbouring neurons.

$$\Delta e_i^n = \Delta e_i \times (1 + FD) \tag{3.9}$$

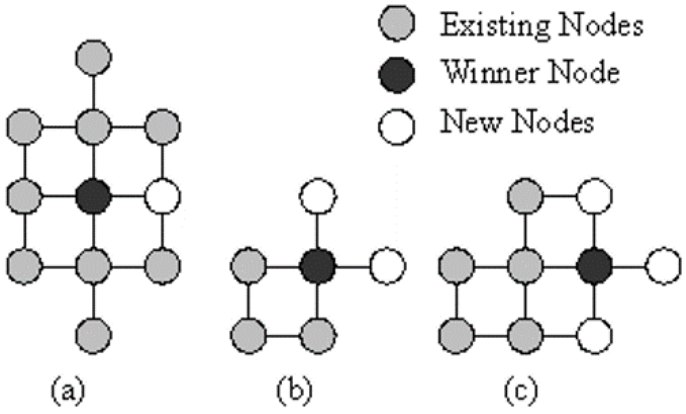


Fig. 3.3 Growth of new nodes in GSOM algorithm. Adopted from Alahakoon *et al.* (2000).

In the smoothing phase, similar to the growing phase, the inputs are presented to the GSOM and weights are adjusted based on the inputs. However, no new neuron will be inserted in this phase as the purpose of the smoothing phase is to smooth out any existing quantization error.

The spread factor of GSOM algorithm enables the GSOM to branch out to represent the dataset at hierarchical levels of granularity, controlling the spread of the neuronal network such that finer clustering of the dataset can be achieved hierarchically as required by the data analyst and on the regions of interest. In a typical exploratory analysis experiment, a low spread factor can be given at the beginning of analysis and then gradually increasing the spread factor value for further observations of selected regions in data.

Due to this flexible shape of the self-structuring network, GSOM can represent a set of data with a fewer number of nodes (at an equal amount of spread) compared to the SOM. This becomes a significant advantage when training a network with a very large data set, as when the number of nodes get reduced, both processing time and the need of computation resources get reduced. In addition, GSOM produces a two-dimensional structure, which results in convenient visualization compared to other growing self-structuring variants such as GNG and GWR.

Thereby, we intend to select GSOM self-structuring algorithm as the basis and foundation of the representation modules (latent and cognitive) in the conceptual framework

(MSKRF) presented in section 2.5. Due to aforementioned advantages of GSOM over other variants, using GSOM as the foundation for representation would lead the artificial system to self-organize based on the environment inputs, providing an effective environmental representation.

Invariant Memory Representation in GSOM

The knowledge representation in GSOM resembles the biological brain which does not persist and recall information with complete fidelity, but, only important relationships of the world independent of the details (Hawkins and Blakeslee, 2007). The objective of invariant memory representation in biological brain is to allow objects, situations and events to occur relatively independent of size, contrast, spatial-frequency, position on the retina, angle of view, lighting, etc. (Rolls, 2008). For instance, when a person encounters a known person, regardless how much of the face of the person has changed or regardless of how distance the face is or variations of the face such as moustache, beard, etc., it should not be difficult to identify the face. These invariant representation of sensory inputs are extremely important for the survival and operation of the brain as it enable the brain to learn in a single trial about reward/punishment associations of the object/occurrence, the location, how it was encountered, and then to correctly generalize to other views of the same object (Rolls, 2008).

The GSOM represents its knowledge in a generalized form irrespective of subtle variations. The competitive learning of GSOM with respect to the distance function will make sure to distinguish across information that have higher variation while generalize to capture inputs with less-variations as similar. The less-variant (or 'similar') information/inputs will be represented using prototypes in the GSOM neural maps. For instance, consider a GSOM representation of animals of the Zoo dataset (Dua and Graff, 2017). Given a lower spread factor, the representation will result in clustering similar animals into single prototypes. For instance, less variant mammals, birds, and fish are grouped into single prototypes as shown in Fig. 3.4 (A). The required degree of invariance can be calibrated by changing the spread factor parameter of the GSOM. This is shown in Fig. 3.4 (B), in which the increased SF has distributed the previously clustered prototypes further apart while keeping highly similar inputs in clustered prototypes.

Thereby, we claim the invariant memory representation is an inherent capability of GSOM. Selecting GSOM as the candidate SSAI algorithm will make MSKRF profoundly adhere to the first biological base proposed in this these, section 2.2.2, which is the invariant representation of memory in neocortex.

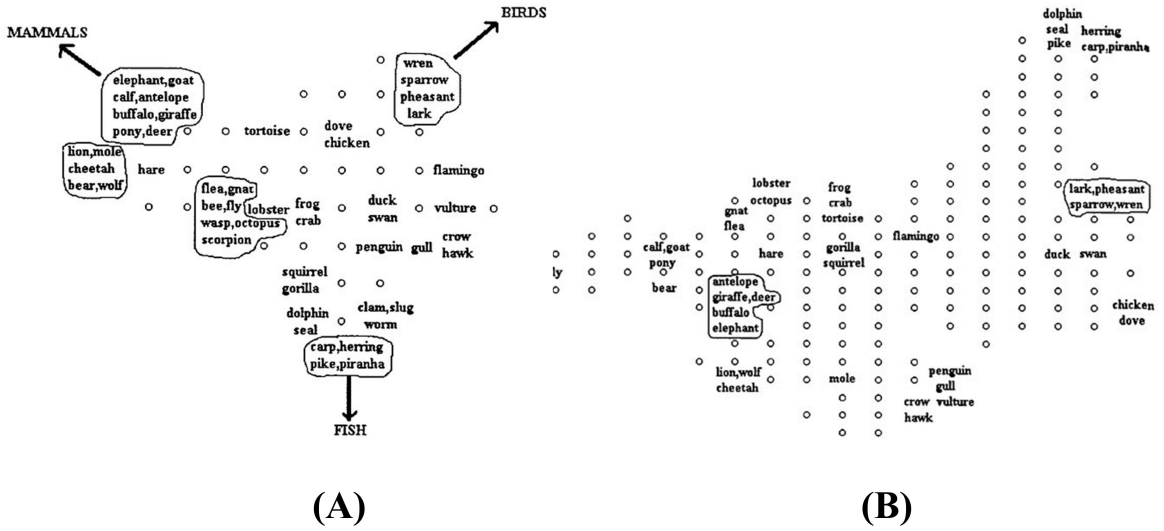


Fig. 3.4 Invariant memory representation of GSOM. A) SF=0.1 and B) SF=0.85. Adapted from Alahakoon *et al.* (2000).

3.5 Practical Exploration of Self-Organization using SSAI

The theoretical foundation discussed in the previous sections lead to the selection of GSOM algorithm as the basis of representation modules. This section aims to conduct a practical exploration of self-organization and representation using selected SSAI algorithms to demonstrate how the Latent representation of MSKRF can be materialized using SSAI. We relate this section to the module (3.5) *SSAI to empower Big Data analytics in Smart Cities* as presented in chapter overview in Fig. 3.1.

We investigate the capabilities of GSOM with respect to the base SSAI techniques in a smart city context for this demonstration. In the context of smart cities, a number of sensory modalities are available such as vehicular traffic data, pollution data, weather data, parking data, surveillance data, etc. Among these data types, video surveillance data can be considered as one of the most complex type of data in terms of computational overhead (Nawaratne *et al.*, 2019c). With this practical exploration, we demonstrate the capability of GSOM to represent input video data in exploratory manner for further processing.

We position this practical exploration in MSKRF as presented in Fig. 3.5, where the uncovered regions show the components of current demonstration. Surveillance cameras can be considered as the sensors in the smart city context generating video data feeds of the surveillance environment. We use GSOM and SOM computation algorithms as the SSAI platform that is already equipped to represent memory/information in an invariant form, in order to validate the importance of using GSOM over SOM. Using the SSAI algorithm, we

then generate a latent representation of the input data space derived from surveillance video. The covered regions are not explored in this chapter as we will extend and develop advanced computing techniques to uncover those regions in forthcoming chapters.

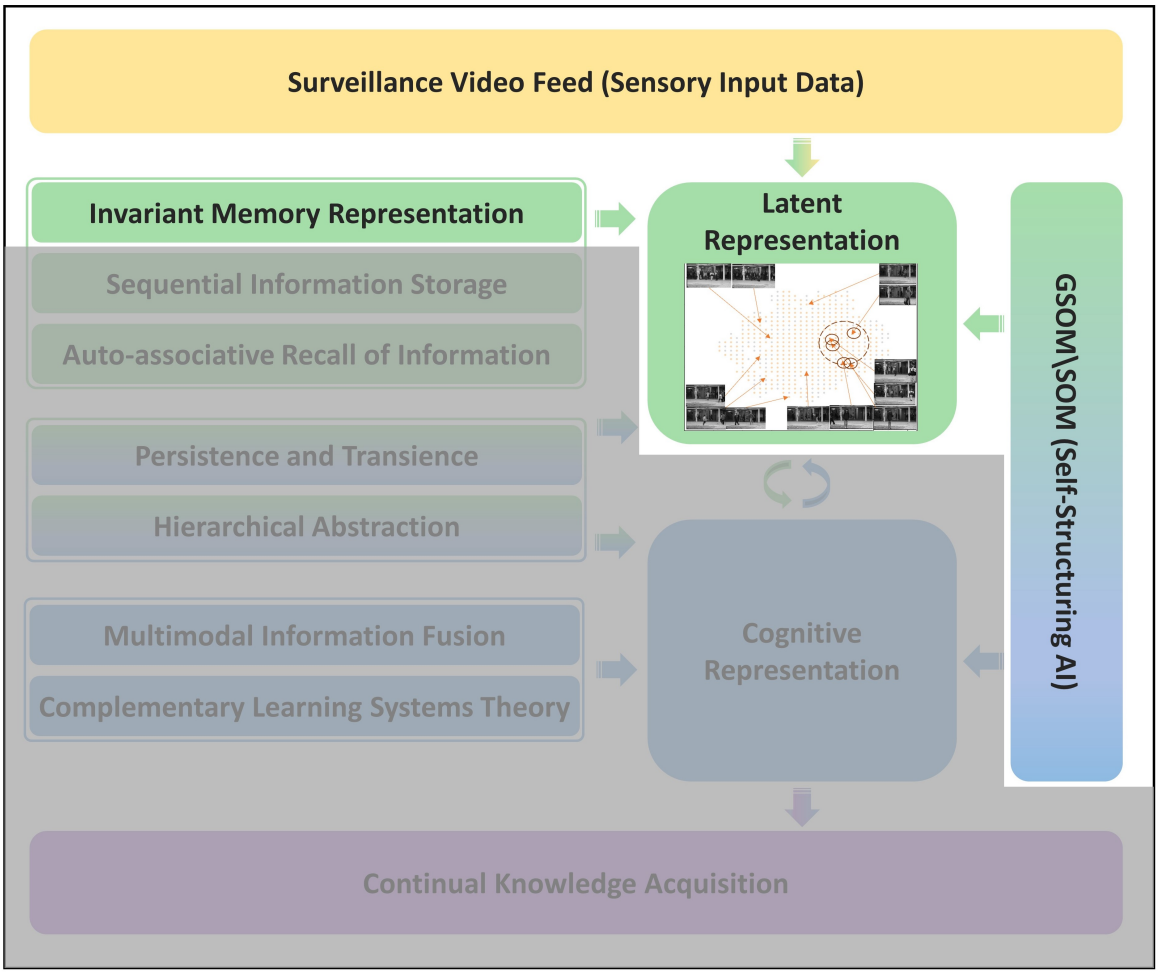


Fig. 3.5 Positioning the demonstration of SSAI in MSKRF.

Prior to the demonstration of the concept framework, the next section will provide an overview of the importance of selecting Smart City environment as the domain/context to conduct the SSAI practical exploration.

3.5.1 Smart City Context

Urban migration is an increasing trend in the 21st century and it is estimated that more than 68% of the worlds' population will live in urban environments by 2050. Such migration will strain the abilities of cities to cope and this situation has created an urgent need for finding smarter ways to manage the challenges such as congestion, traffic and transport, increased

crime rates, social disorder, higher need and distribution of utilities and resources, etc. with smart cities being proposed as the solution (Hashem *et al.*, 2016). Using technological advancement as the base, smart cities are expected not only to cater to the needs of a huge increase in population, but also provide improved living environments, utilize resources more efficiently and responsibly as well as be environmentally sustainable. Basic technology facilitated functions within smart cities will include the collection of data using sensors, CCTV cameras, smart energy meters as well as social media engines that capture real-time human activity which are then relayed through communication systems like fibre optics, broadband networks, internet and Bluetooth.

AI has the potential to make sense of the large and diverse big data and use insights derived to optimize operational costs and resources, and enable sound citizen engagement in smart city environments. As discussed by Guelzim and Obaidat (2016), public safety and security could be enhanced by AI through sophisticated surveillance technologies, accident pattern monitoring, linking crime databases and combating gang violence. AI can also help with crowd management, estimation of size, predicting behaviour, tracking objects and enabling rapid response to incidents.

Although the utilization of AI, machine learning and advanced analytics have provided much value for smart cities, there is one key constraint that limits the realization of the advantages from such technological advances. That is the utilization of AI and advanced analytics is currently carried out in silos and as isolated applications due to the lack of information integration and sharing mechanisms. Since each incident, occurrence, behaviour and situation will be different in terms time taken, number of people and objects involved, spatial features, background and types of data representations, smart city environments are non-deterministic and deciding the appropriate architecture or structure of machine learning models becomes a difficult of even impossible task. Therefore, the two main requirements of AI and machine learning for dynamic and volatile environments (non-deterministic) such as smart cities are:

1. The ability to self-learn and adapt without pre-labelled past data,
2. The ability to self-structure the architecture or network structure to represent a particular situation.

As a solution, the proposed MSKRF framework underpinned by SSAI can be used as the base technology for representation modelling with the ability self-structure and self-learn (unsupervised learning) to match individual situations. The self-structuring capability of the model is harnessed to reduce the need for human involvement in utilizing machine learning at the front-end data capture and initial analysis stage where real-time detection of anomalies,

pattern and trend detection and prediction could be achieved. The captured data could then be passed on for further specialised processing and advanced analytics. As such, we demonstrate the capability of SSAI in MSKRF through the lens of a smart city environment in the next sub-section.

3.5.2 Experiment Objectives

With the aim to overcome the challenges of representing smart city environment in digital space, this experiment attempts to represent the data intensive environment in a digital world using the self-structuring algorithm, GSOM. The objectives of the experiment are three-fold;

1. First, we evaluate representation of raw smart city data volumes that generate at the artificial sensors (i.e., surveillance cameras). We derive SOM based local feature maps to demonstrate limitations existing in structuring the input data space into a latent space representation. The objective of this experiment is to explore the validity of using a self-structuring AI core algorithm in order to represent the input space with context based structural adaptation. This experiment validates the benefit of structural adaptation in the AI technology.

2. Second, we develop local feature maps using GSOM algorithm. We extend the representation with multi-granular structural adaptation and show how the granularity calibration can be implemented based on requirements of the context. This capability enables the AI to self-structure representations at different level of abstraction which could be calibrated to link objects of interest or events from different source data or timelines.

3. Third, we compare the optimal representation for selected datasets with respect to their context and the granularity, demonstrating the capability of structural adaptation to optimally represent the input data for further processing. The feature in self-structuring AI enables automatically representing the object or event optimally without human involvement.

3.5.3 Dataset and Feature Extraction

The experiment utilizes two benchmark video action recognition datasets from smart-city contexts which involved sparse to dense crowd behaviour. First, the CUHK Avenue dataset Lu *et al.* (2013a), a video recording of street activities at the City University, Hong Kong that were acquired using a stationary video camera with a resolution of 640×360 pixels. CUHK

Avenue dataset contains total of 37 video samples ranging from human behaviour such as people walking on walking paths and groups of people meeting on the walking path to unusual behaviour such as people throwing objects, loitering, walking towards the camera, walking on the grass and abandoned objects. Few example frames from the Avenue dataset are illustrated in Fig. 3.6.

Fig. 3.6 Examples of CUHK Avenue dataset

Second, the UCSD pedestrian dataset (Mahadevan *et al.*, 2010a), that was recorded focusing towards a pedestrian walkway. This dataset captures different crowd scenes, ranging from sparse to dense crowd flow in its video samples. Here we utilize the training data set which contains 34 video samples with a resolution of 238×158 pixels. For the experimental purposes, we manually labelled the videos based on their crowd density into 4 categories (low, mid, high, very high). Few example frames from the Avenue dataset are illustrated in Fig. 3.7.

As the video samples have different dimensionality, we pre-process the inputs by resizing the extracted frames to 30×30 pixels, and normalizing pixel values by scaling between 0 and 1. Then we extract the histogram of optical flow (HOF) (Wang and Snoussi, 2012) and histogram of gradients (HOG) (Dalal and Triggs, 2005a) to be utilized as feature vector for online structural adaptation.

3.5.4 Representation of Smart City Video on SOM

In the first step of the experimental procedure, we develop a SOM based representations for UCSD pedestrian dataset (Fig. 3.8) and Avenue datasets (Fig. 3.9). For this feature map development, we used 20×20 feature map with a learning rate of 0.01. In the training phase, the SOM feature maps were trained for 100 iterations.

Fig. 3.7 Examples of UCSD pedestrian dataset

The latent representation of the pedestrian flow of UCSD dataset can be visualized in the SOM map shown in Fig. 3.8. The colour code elicits the density of the pedestrian flow. It can be identified here that the dense pedestrian movements are clustered from the centre to middle of the SOM map. However, low-dense and mid-dense videos are scattered throughout without any fine-grained cluster region. Analysis of the SOM visualization shows the self-structuring of the neural network has become restricted due to the rigid structure (pre-defined 20 x 20 grid). Thereby, initial learning of dense pedestrian movements has been firmly clustered while rest of the scenarios were constrained to grow as required by the feature representation.

As illustrated in Fig. 3.9, the latest representation of the avenue dataset is plotted with detailed frames each SOM node has been clustered. A fine grained analysis resulted in identifying 4 unique regions, which have different characteristics of images. Fig. 3.9 (A) is composed of frames that are idle with minimal crowd present and negligible motion activity. Fig. 3.9 (B) contains frames in which large gathering of people stays idle and slight movements at distance. Fig. 3.9 (C) contains rapid movements of people across the frame (left-right), whereas Fig. 3.9 (D) contains movements of crowd to and from the camera position.

Both SOMs that were developed upon the UCSD and Avenue datasets are constrained by the initial node size constraint. Thus, structural adaptation of the node map is restricted, making the natural spread limited to a pre-defined threshold. Due to this limitation, we propose a growing variant of SOM to be utilized for the structural adaptation of the video input data. From a smart city perspective, the diversity and distribution of input data in a single stream (e.g., video data) could be unpredictable, thereby, forming a pre-defined

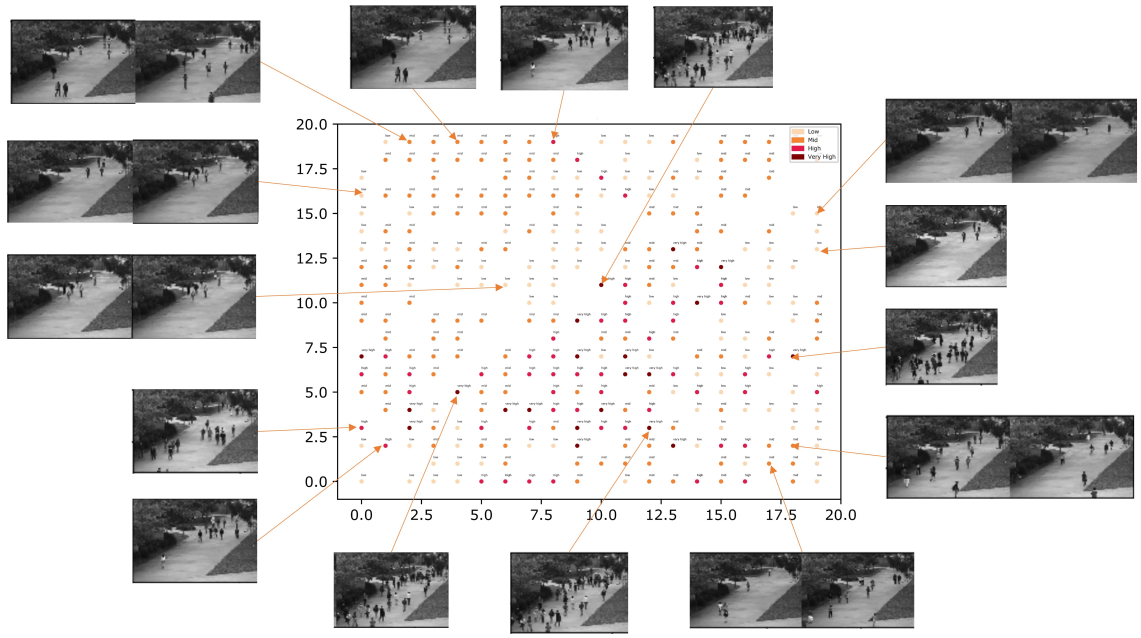


Fig. 3.8 SOM based local representation for UCSD Pedestrian Dataset

structure is unrealistic. This further emphasize the need for unconstrained self-organizing algorithm such as GSOM.

3.5.5 Representation of Smart City Video on structural adapting network

The second experiment aims to demonstrate capabilities of GSOM to overcome the aforementioned limitation of constrained self-organization as shown previously. In addition, we explore multi-granular structural adaptation capability lies with GSOM and how the granularity calibration can be implemented based on the requirements of the context. In this step, we develop growing feature maps for two selected datasets using the GSOM algorithm. Due to the unconstrained nature, we do not need to specify the map size constraint. We let the GSOM algorithm to structurally adapt based the dynamic nature of the input feature space on the video data.

The growing feature map for multiple granularities for UCSD dataset is depicted in Fig. 3.10. The structural adaptation has been conducted for three levels of abstractions, i.e., spread factor of 0.3, 0.5 and 0.8. For demonstration purposes, we have selected 8 distinct video frames from different crowd density characteristics. (plotted identically in Fig. 3.10) The colour code elicits the density of the crowd flow. Based on the structural adaptation, it

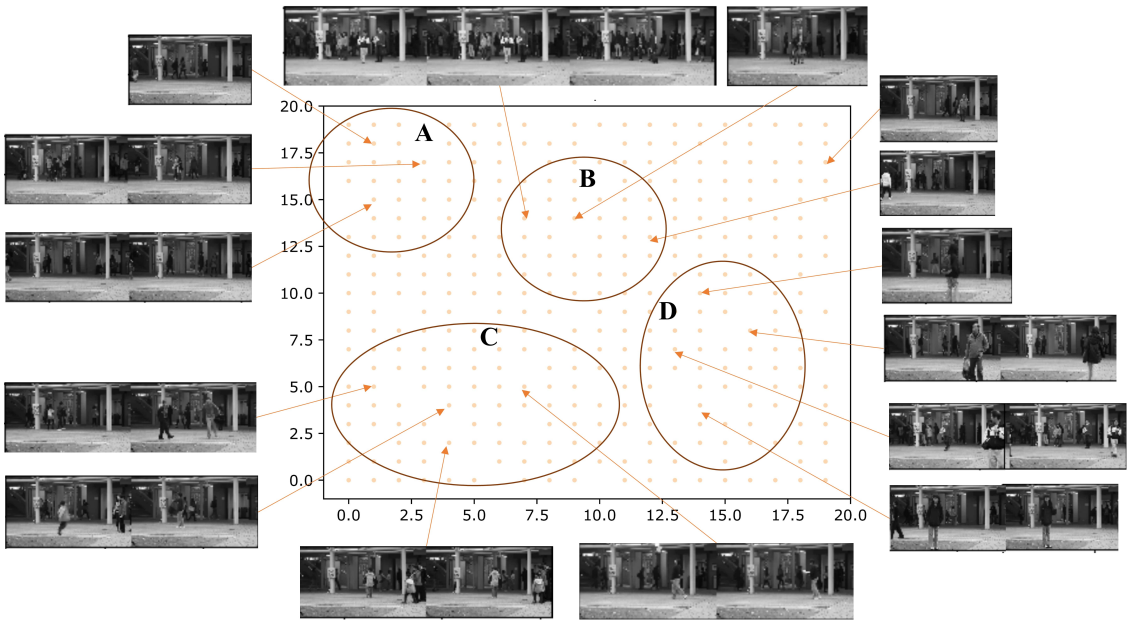


Fig. 3.9 SOM based local representation for Avenue Dataset

can be seen that the network growth is wider when the spread factor increases. At the same time, the data points that were closer in Fig. 3.10 (a) have parted when spread increases, i.e., Fig. 3.10 (b). This will enable the GSOM map to represent input data space in detailed with different calibrations of spread factor.

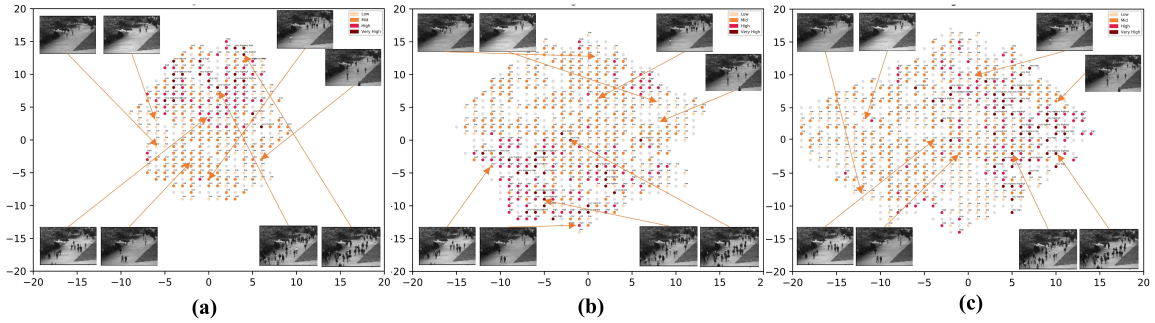


Fig. 3.10 GSOM based local feature maps with multiple granularities for UCSD Pedestrian Dataset (Spread Factor – (a) 0.3, (b) 0.5, (c) 0.8, Learning Rate – 0.01, Learning Iterations - 100)

Similarly, we derive GSOM representation maps (Fig. 3.11) for Avenue datasets with three different abstraction levels. Same observations we identified in UCSD dataset applies here, such that, the calibration of spread factor change the granularity of representation of the input data.

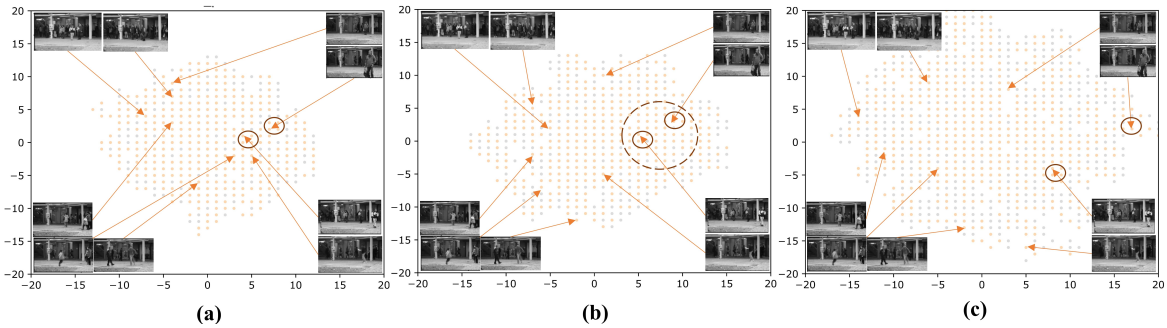


Fig. 3.11 GSOM based local feature maps with multiple granularities for Avenue Dataset (Spread Factor – (a) 0.3, (b) 0.5, (c) 0.8, Learning Rate – 0.01, Learning Iterations - 100)

It is apparent from the multiple feature maps with multiple abstraction levels that the structural adaptation is possible with GSOM algorithm. Here, the growth/ structural adaptation of the neuron network is not restricted by a pre-defined map size constraint, but allow the input data space to decide how it should spread, restricted only using a friction parameter (i.e., spread factor). In contrast to SOM, GSOM has the capability to adapt its structure with respect to input data space dynamically and thereby optimal to utilize in the proposed framework. In the context of smart city, multi-granular exploration capability is imminent as the distribution and diversity of input data is not known prior. For instance, consider a scenario that requires facial recognition from a surveillance camera. Depending on the targeted person's position from the camera (i.e., distance from the camera lens), it might be required to change the granularity in which the face should be identified. We will further explore the positioning capability of the propose AI framework in following experiment.

3.5.6 Structural Adaptation with Context Awareness

The objective of the third experiment is to evaluate the optimal structural adaptation of the feature maps for pre-defined context requirements. We evaluate the optimal structural adaptation of the feature maps for pre-defined context requirements. Here we define the context requirement for the UCSD data source as to detect the flow of pedestrians, and context requirement for the Avenue data source as to detect forward facing frames that are optimal to run through face recognition due to pixilation constraints (i.e., direct facing people with a large pixel coverage).

Fine grain analysis of the UCSD feature maps show that the minimally spread feature map (Fig. 3.12 (a)) is optimal to represent the crowd density, based on the color-coded nodes and manually tagged sample video frames.

We analyzed the feature maps of Avenue dataset and manually tagged video frames as highlighted in the Fig. 3.12 (b). This region analysis shows that feature map of SF=0.3 is tight, thus both forward facing frames as well across walking frames are closely represented, making it difficult to distinct between them. In contrast, the feature map of SF=0.8 shows a sparse representation. However, the feature map with SF=0.5 provided a fine cluster of forward-facing frames. Therefore, it is evident that for the current context requirement of identifying forward facing frames, the feature map of SF=0.5 is optimal. Further, it is important to note when the context requirement differs, the optimal representation could become different from what is current optimal.

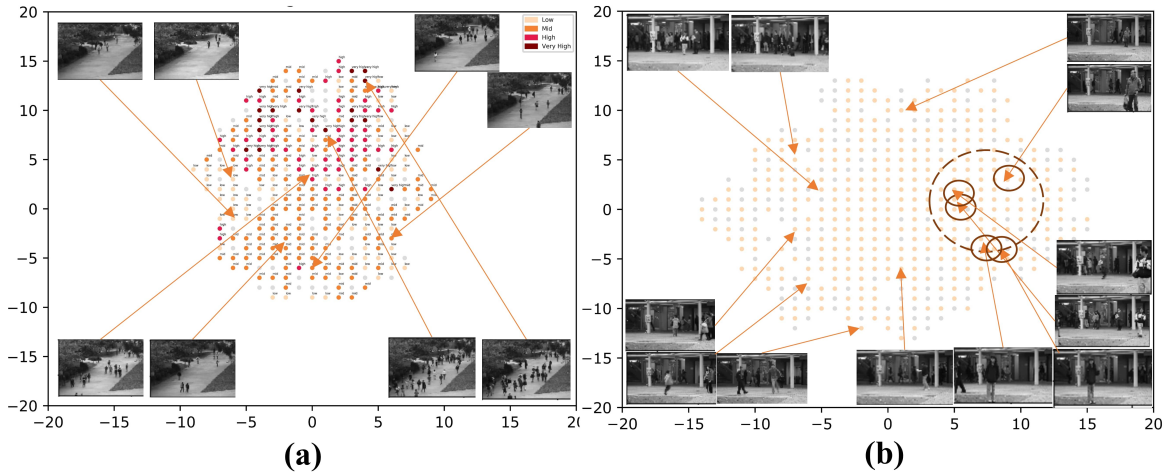


Fig. 3.12 Optimal local feature maps derived based on the context requirements. SF-0.3 for UCSD feature maps, SF-0.5 for Avenue feature map

In the context of smart cities, diverse sensory devices have specific requirements. For instances, there are CCTV cameras that are designed to capture anomalies from the surveillance context, detect faces in entrances, automatically detect vehicle license plates and/or speed of vehicles and for general surveillance. These specific cameras would have different objectives, thus, each would need to calibrate for specific context requirement. For example, anomaly detection cameras should be calibrated to capture high granular movements to capture subtle changes in motion, facial detection cameras should be calibrated to obtain optimal representation of face regions and vehicle detection cameras should be calibrated to zoom into vehicle license plate and extract that specific region. Thereby, having the capability to derive optimal structural adaptation for each local feature map for pre-defined context requirements is a prominent capability provided by the proposed self-building AI framework to smart city context applications.

3.5.7 Analysis of computational overhead

The experiments were carried out on a multicore CPU at 2.4 GHz with 16GB memory and GPU of NVIDIA GeForce GTX 950M. The average processing time for activity representation (HOG and HOF features) and self-organization was respectively 12 milliseconds and 144.7 milliseconds per frame. Thus, the overall computational is 156.7 milliseconds per frame, i.e., the algorithm is able to process 7 frames per second (FPS). From the runtime analysis, it is evident that the maximum computation occurs at self-organization phase, thus, by speeding up this process using parallelized implementation, it would be possible to further enhance runtime efficiency.

3.5.8 Discussion

The emerging information revolution makes it necessary to manage vast amounts of unstructured data rapidly. As the world is increasingly populated by IoT sensors that can sense their surroundings and communicate with each other, a digital environment has been created with vast volumes of volatile and diverse data. Traditional AI and machine learning techniques designed for deterministic situations are not suitable for such environments. With a large number of parameters required by each device in this digital environment, it is desirable that the AI is able to be adaptive and self-structure, rather than be structurally and parameter-wise pre-defined.

This experiment explores the benefits of using self-structuring AI and machine learning with unsupervised learning for empowering big data analytics for smart city environment, highlighting a case-study of video surveillance, and action recognition. By using the GSOM, a novel capability suite of self-structuring AI is introduced to overcome the limitations of traditional AI and enables data processing in dynamic smart city environments.

We derive both SOM based and GSOM based representation to validate the importance of GSOM self-structuring capability over SOM in order to represent the input space with context based structural adaptation. Resembling the invariant representation of memory in neocortex, GSOM provide an effective invariant representation of the sensory input obtained from the edge devices (i.e., CCTV). On this basis, we justify the importance and applicability of GSOM algorithm to provide the base of representations in the proposed Knowledge Representation Framework (MSKRF).

Second, we experiment the representations generated using GSOM with multi-granular structural adaptation and show how the granularity calibration can be implemented based on requirements of the context. This capability enables the AI to self-structure representations at different level of abstraction which could be calibrated to link objects of interest or events

from different source data or time lines. This prompts a resemblance to the hierarchical abstraction in biological brain as presented in section 2.3. Thereby, we can further justify the implacability of GSOM as the representation mechanism of the proposed framework.

3.6 Summary and Research Questions Revisited

The previous chapter (Chapter 2) brought together the neurophysiological inspiration, the features of the big data and digital environments and presented in the form of a landscape and proposed a conceptual framework, *Multi-layered Self-structuring Knowledge Representation Framework* (MSKRF), as the basis for the research carried out and described in this thesis. The proposed MSKRF is set to form an AI based knowledge representation to capture continuously evolving environmental stimulus and adapt its knowledge accordingly, abiding the overall objective of continuous lifelong learning. The intermediate knowledge representation mechanism of the MSKRF framework lies in the two hierarchically connected layers: latent (LR) and cognitive (CR) representations, which provides the foundation for learning from input by representing sensometry input stimuli from artificial somatosensory in a digital form.

This chapter intended to explore options available in existing computational paradigms to select the most suitable computational model to facilitate the two representation modules. For effective knowledge representation, LR and CR should be able to adapt and automatically structure its representation along with continual changes in data distribution over time, to resemble evolving external environment. Thus, it is pertinent that we adapt the most viable learning paradigm to lay the foundation for these representation mechanisms. Thereby, we introduced Self-Structuring Artificial Intelligence (SSAI) with an unsupervised learning paradigm as the solution for this.

On this basis, this chapter (Chapter 3) investigated both biological and natural phenomenon that governs representation as whole in natural environments in order to select and validate the most viable SSAI learning paradigm, as illustrated in 3.1. Initially, we explored the feasibility of modelling the indeterministic natural environment using machine learning paradigms in order to facilitate representation learning, which narrows down to unsupervised self-organization as a viable prospect for representation learning. We examine both natural and biological prospects that use self-organization for multitude of representation mechanism. We identified experience-driven self-organization is a key constituent in biological brains that enable humans to continuously acquire knowledge in their lifetime.

Followed by a detailed theoretical foundation, we provided an in-depth review on existing computational models that have the capability to self-organize focusing on their prospects and

limitations. The comparative analysis lead to the selection of a viable candidate algorithm to base SSAI for the proposed MSKRF, which is the Growing Self-Organizing Map (GSOM) algorithm. The chapter concludes with a practical exploration to demonstrate the SSAI capabilities of the GSOM algorithm, carried out in a Smart City platform forcing on intelligent video surveillance.

The conceptual formulation of SSAI and the smart city based practical exploration was presented in the journal article entitled *self-building artificial intelligence and machine learning to empower big data analytics in smart cities* (Alahakoon *et al.*, 2020). An extended study to utilize SSAI in IoT data interoperability environments was presented in the journal article entitled *Self-evolving intelligent algorithms for facilitating data interoperability in IoT environments* (Nawaratne *et al.*, 2018).

Overall, this chapter partly addresses the second research question, RQ2; **What are the computational and machine learning constituents of continuous lifelong learning for materializing the proposed conceptual framework?** As stated in section 1.4, RQ2 is decomposed into four sub-questions, in which the first and second sub-question have been addressed in this chapter:

RQ 2.1) What are the computational and machine learning foundations for representation learning in the digital world that have been proven through both neurophysiological and ecological studies? This sub-question is addressed in section 3.1 through section 3.3 by reviewing key functions and natural formulations of self-organization and the need for self-structuring foundation in order to represent the natural world. Drawing on the inspiration from nature and its natural phenomenon, we brought forward computational models that are capable to self-structure.

RQ 2.2) What are the structural and algorithmic limitations in current AI for achieving continuous lifelong learning and what fundamental architectural changes will address these limitations? This sub-question is addressed in Section 3.4 with a comprehensive exploration of existing self-structure-capable computational models, to identify structural and algorithmic limitations in current AI that limit the achievement of continuous lifelong learning. We determined that the pre-defined structure of existing self-structure based computation methods limit their representation capability, thus allowing a self-growth without a constrained structure will enable an effective representation. Thereby, we narrowed down the literature scan to Growing Self-Organization Maps (GSOM) algorithm which captures the essence of an effective self-organization based representation mechanism. Further we demonstrated the validity and robustness of GSOM algorithm through a practical exploration

under Smart City context in Section 3.5.

Despite the strengths and capabilities of GSOM to represent natural environments, Big Data challenges poses a volatility in data streams. Thereby, digital representations of ecosystems are necessary to accommodate temporal structure of data streams to continuously adapt its knowledge representation. The next chapter describes attempts to address these novel demands in computation to represent temporal structure to attain increased plasticity without compromising the stability of computational models.

Chapter 4

Recurrent and Transience Self-Organization with Bio-inspired Stability and Plasticity

The neurophysiological inspiration for the research and the characteristics of the big data and digital environment are brought together to form a landscape, proposing an overarching conceptual framework, *Multi-layered Self-structuring Knowledge Representation Framework* (MSKRF) in Chapter 2. MSKRF has the potential to form an AI based knowledge representation to capture continuously evolving environmental stimulus and adapt its knowledge representation accordingly, supporting the overall objective of continuous lifelong learning. The intermediate knowledge representation mechanism of the proposed MSKRF lies in the two hierarchically connected layers: latent (LR) and cognitive (CR) representations, which provides the foundation for learning from input by representing sensometry input stimuli from artificial somatosensory in a digital form. Chapter 3 explored the pertinence of these representation layers in facilitating the overall objective of continuous lifelong learning of connectionist models. An in-depth investigation of viable computation models to lay the foundation for these representation mechanisms resulted in identifying Self-Structuring Artificial Intelligence (SSAI) with an unsupervised learning paradigm to be the most suitable, thereby, identified Growing Self-Organizing Maps (GSOM) as an effective algorithmic base to lay the foundation for representation layers in MSKRF.

While AI has become increasingly widespread, a number of research work have attempted to claim their relationship with biological intelligence, in which higher chances of a given technology succeeding when working on the assumption that AI systems should mimic the mechanism of biological intelligence (Said and Masud, 2013). Kiritsis (2011) stated that the advancement of AI should be followed through the inheritance of the essential

characteristics and inspiration from highly evolved human perception system. Through a thorough analysis literature on the workings of the biological brain (Section 2.3), this thesis identified a collection of seven biological bases, which have been essential for the survival and continuance of humans, that can provide the inspiration for enhancing the capabilities of AI systems in order to achieve continual lifelong learning.

GSOM, the chosen candidate SSAI algorithm to develop the foundation of MSKRF representation layers, is capable of representing its knowledge in an invariant form, satisfying the first of the biological bases. This chapter focuses on implementing five of the remaining six biological bases for MSKRF representation layers, which are:

2. Persistence and transience of memory,
3. Sequential information storage capability of neocortex,
4. Auto-associative recall of information from neocortex,
5. Hierarchical abstraction in memory storage, and
6. Multi-modal information fusion.

The subdivision of the chapter describing each section is presented in Fig. 4.1. Section 4.1 studies the persistence, i.e., learning to remember, which is the predominant focus of cognitive inspired learning systems in order to understand the limitations brought by mere focus on persistence. We look into mechanisms to preserve stability and plasticity of the representation layers in MSKRF that result in integration of transience as a key solution. This leads to the implementation of a strategic forgetting mechanism for GSOM followed by an experimental evaluation of topology preservation for the new GSOM with transience algorithm. This section will focus on integrating the biological base *Persistence and transience of memory* in the representation layers of the proposed MSKRF.

Section 4.2 explores non-stationary and sequential nature of environmental stimuli, in which sequences of data are given by series of temporally and spatially connected observations. We identify that sequential data plays a vital role in the biological perception system since all sensorimotor data are given as sequences of time-varying stimuli, which makes it pertinent to be incorporated in the AI counterparts. As a solution we implement a recurrent information processing mechanism for the GSOM algorithm in order to capture sequential information from input stimuli. This leads to integration of two biological bases; *Sequential information storage capability of neocortex* and *Auto-associative recall of information from neocortex*.

The natural environment is highly complex; thus, human nervous system is equipped to sense this complex environment with multiple modalities. When an event occurs, more

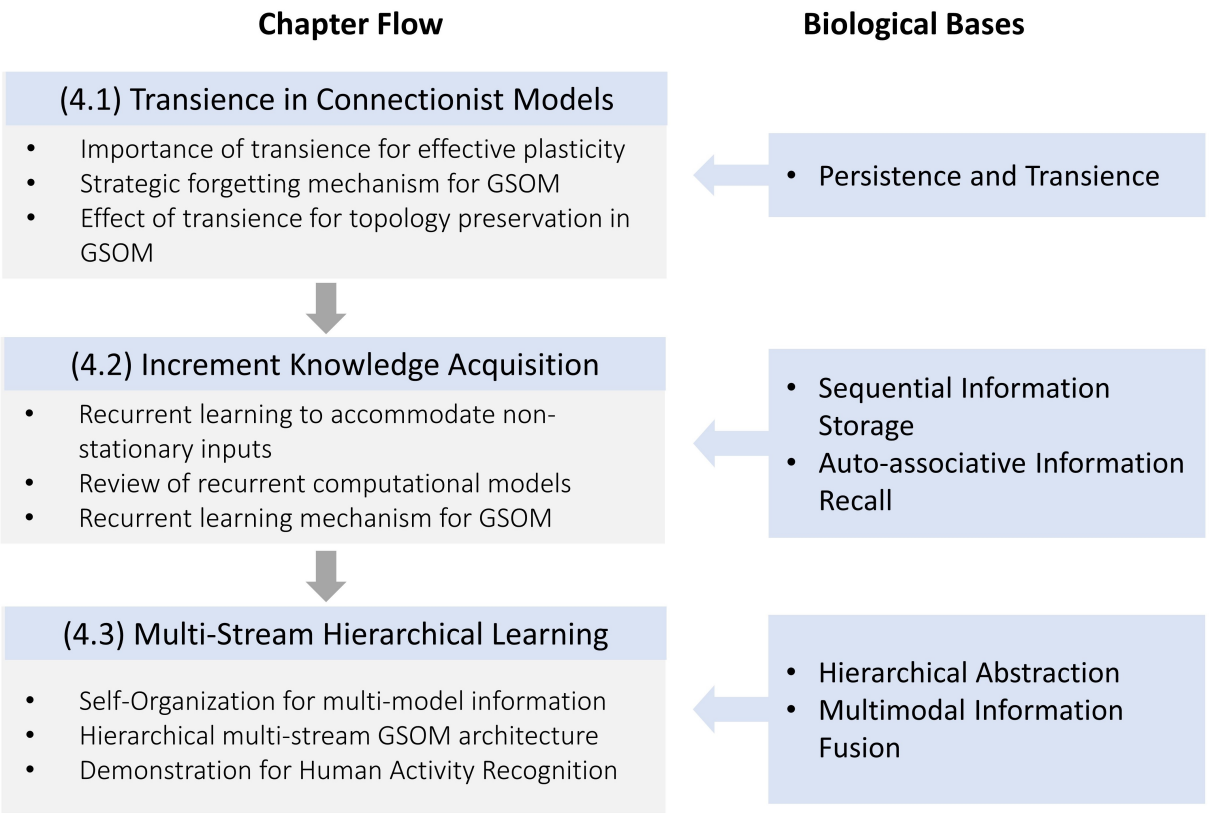


Fig. 4.1 Chapter Overview

than one sensor detects the events, generating redundant neural signals underpinned by the multisensory processing of biological brain. Section 4.3 lays foundation to process multiple information streams by implementing a multi-stream architecture composed of the extended GSOM algorithm. These extensions lead to incorporation of two biological bases; *Hierarchical abstraction in memory storage* and *Multi-modal information fusion*. The proposed multi-stream self-organization architecture is experimented using dense activity recognition video data sets in Section 4.3.1 in order to confirm its validity and usability in real-world settings.

4.1 Transience in Connectionist Models

In this section, we look into mechanisms to preserve stability and plasticity of the representation layers in MSKRF relating to the module (4.1) *Transience in Connectionist Models* in the chapter overview presented in Fig. 4.1. The focus of this section is directed to integrate the biological base: *Persistence and transience of memory* in the proposed MSKRF.

The predominant focus of cognitive inspired artificial learning systems is persistence (learning to remember), where two key limitations exist. First is the influence of outdated information on memory-guided decision-making, and second is overfitting of acquired knowledge to specific events. Recent studies have confirmed the existence of a parallel neurobiological mechanism of transience (forgetting) in human neurophysiology, which states, the interaction between persistence and transience advances intelligent decision-making in dynamic, noisy environments (Richards and Frankland, 2017). Prior research suggests mnemonic transience can be used as means to eliminate overfitting in noisy environments. Overfitting phenomenon occurs when learning models overly fit to a finite dataset leading to inaccurate predictions for unseen and new data (Hawkins, 2004). As such, overfitted models encode inaccurate and false patterns that are overly specific to the noise in training data, but do not generalize to new situations.

Early conceptualization of memory was as a faithful reproduction of past experiences. However, recent discoveries suggest otherwise, such that memories are simplified representation that only capture the essence of past experience, not necessarily in detail (Richards and Frankland, 2017). Thereby, prevention of overfitting (to past experience) would lead simple memories to be more successful in preserving previously learnt knowledge by storing the gist of the learnt experiences. This in turn will lead to better generalization for future events (Kumaran *et al.*, 2016).

A number of research work have attempted to prevent overfitting in connectionist networks. For instance, if too many parameters are used to model the data, the data can be described in a straightforward manner. However, this would lead to overfitting to the training data used, making it unfit to generalize and predict new data points (MacKay and Mac Kay, 2003). Most intuitive solution for this is to use more simpler methods with a smaller number of parameters. Usually this prevention involves restricting the complexity of learning models by limiting the number of parameters (e.g., number of connections) used to model the data. A well-known heuristic for scientists, *Occam's Razor* states this more intuitively, "Entities should not be multiplied unnecessarily." (Blumer *et al.*, 1990). This can be translated as: the simpler the explanation, the broader its applications (Richards and Frankland, 2017).

Most state-of-the-art machine learning models achieve this simplification through *regularization*. As stated by MacKay and Mac Kay (2003), the regularization is the process of constraining machine learning models to promote generalization. Common regularization techniques include drop-out (Srivastava *et al.*, 2014) in connectionist models, weight decay over time (MacKay and Mac Kay, 2003), sparse coding (Olshausen and Field, 1996) and noise injection (Hinton and Van Camp, 1993). These approaches for regularization resemble forms of partial forgetting.

Long Short Term Memory Networks (LSTM) uses a specific learning mechanism to decide which experience to remember and which are to suppress using a forgetting mechanism (Hochreiter and Schmidhuber, 1997). For example, consider a scenario where a machine learning algorithm is expected to predict what will happen next in a movie. Take an example of a particular scene where a woman holds a knife. The machine learning algorithm is expected predict whether the woman is a chef or a murderer. In such scenario, LSTM attempt to address this by using three key modules: Forget, Save and Focus. LSTM holds a separate mechanism to *Forget*; when new inputs are introduced, the algorithm learns to either remember that information or forget it. For example, if a scene of the movie ends, the model should forget the current scene location, and reset any information related to that. In the *Save* mechanism, the model needs to learn whether any information about a particular frame is worth saving (e.g., if the woman walks by a billboard, will it be useful for next predictions if the content in billboard be saved?). *Focus* module provides mechanism to remember the context of the woman (e.g., If the woman holds children, it needs to assign the concept of mother to woman. In a later scene, even if the woman does not hold her children, it should be remembered that she is a mother.). LSTM helps to determine which parts to focus on at any given time while keeping everything safely stored for later use (Fratto, 2018).

Google's DeepMind proposed Elastic Weight Consolidation (EWC) as a mechanism to maintain expertise on tasks which machine learning models have not experienced for a long time (Kirkpatrick *et al.*, 2017). This was introduced based on the neuro-biological process of *synaptic consolidation*. During synaptic consolidation, the brain assesses the task and analyze the neurons that are used to perform the task in order to identify which neurons were more critical. These critical neurons are less likely to be overwritten in subsequent tasks. Similarly, in connectionist models, EWC suggests to selectively slow down learning on the weights that are more important to tasks that have not experienced in long time.

Naftali Tishby proposed The Bottleneck Theory (TBT) that states, "A network rids noisy input data of extraneous details as if by squeezing the information through a bottleneck, retaining only the features most relevant to general concepts" (Tishby, 2017). As the theory states, neural networks go through a fitting phase and compression phase during learning. The network labels its training data during the fitting phase, while during the compressing stage, the network analyze the neurons to identify only the strongest features that are used to perform the tasks. These strongest features will be more relevant to generalize the model. The TBT suggests the compression stage perform as a strategic forgetting step in order to generalize the model, similar to EWC.

Based on the above findings, we can conclude that *strategic partial forgetting* is widely used method to construct memories in connectionist models (Hochreiter and Schmidhuber,

1997; Tishby, 2017; Kirkpatrick *et al.*, 2017). On this premise, we draw parallels between neurobiological and computational mechanisms underlying transience, bringing forth the idea that the interaction between persistence and transience allows for intelligent decision-making in dynamic, noisy environments, as proposed in number of studies reviewed before. In other words, to store only the abstraction of an experience while forgetting statistically insignificant details (Sekeres *et al.*, 2016). Thereby, this thesis intends to amalgamate transience in the GSOM algorithm by integrating a strategic partial forgetting function. This enables to integrate the biological basis *persistence and transience* in the MSKRF as introduced in Section 2.3.2.

4.1.1 Growing Self-Organization with Transience

Drawing on the basis of persistence and transience that allows intelligent decision-making in dynamic and noisy environments, we extend the GSOM algorithm with the integration of transience property to actuate plasticity during knowledge acquisition. We denote the new extension algorithm as Transience-GSOM (TGSOM).

We define the transience threshold M as the maximum lifespan of a neuron without being selected as the winner (BMU) at the competitive growing phase of TGSOM (a detailed discussion of the competitive learning process is presented in Section 3.4.1). For each neuron u_i, we parameterize the age since being selected as a BMU as λ_i. In the growing phase, as each input signal is presented to the neural network, λ_i is updated as follows.

$$\lambda_i = \begin{cases} 0, & \text{if } u_i \text{ is a new neuron} \\ 0, & \text{if } u_i \text{ is selected winner (BMU)} \\ \lambda_i + 1, & \text{otherwise} \end{cases} \tag{4.1}$$

Given ϕ_t is the set of all neurons at iteration t, we update ϕ_t at each iteration as defined in equation 4.2 and equation 4.3, where ϕ'_{t+1} is the set of inactive neurons (i.e., outdated knowledge) at iteration t.

$$\phi'_{t+1} = \{u_i | u_i \in \phi_t \wedge \lambda_i > M\} \tag{4.2}$$

$$\phi_{t+1} = \phi_t - \phi'_{t+1} \tag{4.3}$$

The pseudo code for TGSOM algorithm is presented in algorithm 1, and the python implementation is available at https://github.com/razmik/htgsom.

Algorithm 1: Growing Self-Organization with Transience (TGSOM)

Data: X
Parameters : SF(Spread Factor), M(Age Threshold), T_L(Growing Iterations), T_S(Smoothing Iterations), η_0(Initial Learning Rate)
Result: Neural Map (ϕ_t)

Start with a set of 4 neurons with randomly initiated weight vectors w_i, where $i \in [1,4]$ and $\lambda_i = 0$;
Calculate GT for the given input space using equation 3.8 ;
`/* For all the learning iterations */`;
for $j \in [1, T_L]$ **do**
  for $x_t \in X$ **do**
    Update the learning rate η_t using equation 3.6 ;
    Select the BMU for x_t using equation 3.2 ;
    Set the transience value of BMU as zero, $\lambda_i = 0$;
    Update the weight of the BMU and its neighbourhood using equation 3.3 ;
    Calculate accumulated error for BMU, TE_i ;
    if $TE_i \geq GT$ **then**
      if *BMU is a boundary neuron* **then**
        Generate new neurons based on Section 3.4.1;
      else
        Distribute the error to its neighbours using equation 3.9 ;
      end
    else
      No node growth ;
    end
    Update the transience property λ_i of neurons using equation 4.1 ;
    Update ϕ_t using equation 4.3 ;
  end
end
`/* For all the smoothing iterations */`;
for $k \in [1, T_S]$ **do**
  for $x_t \in X$ **do**
    Update the learning rate η_t using equation 3.6 ;
    Select the BMU for x_t using equation 3.2 ;
    Update the weight of the BMU and its neighbourhood using equation 3.3 ;
  end
end

4.1.2 Preservation of Stability and Plasticity

This sub-section intends to provide a justification on the preservation of stability and plasticity in the newly introduced TGSOM algorithm. In humans and animals, a basic structure of the surrounding environment is initially encoded at the fetal development and we spend the lifetime augmenting, updating and correcting those initial impressions and developing new understanding related to these initial impressions (Shatz, 1992; De Silva and Alahakoon, 2010). To facilitate this experience driven continuous learning, the memory should be able to alter its functional properties and neuronal structure, which is termed "plasticity". William James originally discovered this phenomenon in the year 1890 when he introduced, "Plasticity means the possession of a structure weak enough to yield an influence but strong enough not to yield all at once" (James, 2007). Konorski (1948) discussed the applicability of this phenomenon in artificial counterparts of intelligence emphasizing the difference between immediate "reaction" of nerve cells to incoming changes and the "permanent transformation" of a system of neurons. This led to the conclusion that AI systems should be both stable and plastic in order to facilitate continuous learning and retention of learned memories. A detailed explanation on stability-plasticity was provided in the Section 3.2.4.

The stability - plasticity dilemma is a profound problem relating to the catastrophic inference phenomenon in connectionist models. That is the disruption of old knowledge by new learning in neural network based computational models with distributed representations (De Silva and Alahakoon, 2010). In order to address this phenomenon, the proposed transience function facilitates to encapsulate plasticity in TGSOM without loss of stability as follows: The transience threshold M enforces the removal of neurons which have not been selected as BMU over a period of time thus determines the maximum lifespan of each neuron with respect to their capability in persisting the input space. The recurring process of self-organization and the constituent neighborhood weight adaptation functionality would have disseminated characteristics of the removed neuron to the neighboring neurons, thereby maintaining stability. Therefore, self-organization with transience will discard outdated information and overfitting knowledge in its knowledge acquisition, without the loss of stability (Nawaratne *et al.*, 2019a).

4.1.3 Topology Preservation in TGSOM

The topology preservation of a self-organizing setup is an important property that is exploited in many self-organization algorithms to substantiate its validity. Topology of a self-organizing architecture is defined as the continuity of the map with respect to the input data space. A self-organized map can be considered to preserve its topology if neighboring data points in

the input space are mapped to nearby neurons in the output (self-organized) space (Khalilia and Popescu, 2014). A successfully topography preserved self-organized map can represent the high dimensional input data space in a lower dimensional output map that preserves the topology of the input data. This topological representation yields better and optimal visualization and reveals granular information about the structure and the clusters presented in high dimensional input space. Various qualitative and quantitative approaches are known for measuring the degree of topology preservation. Typically, these measures are based on using locations of synaptic weight vectors. Alahakoon *et al.* (2000) confirms the topology preservation characteristics of vanilla GSOM algorithm. Thus, it is important to uphold the topography persistence with the novel introduction of the transience property, as the removal of synapses (neurons) of the self-structuring architecture might have affected its topology. Thereby, in this section, we quantify the topography of TGSOM using two widely recognized topology measure techniques: Topographic Error and Zrehen Measure.

Topographic Error

Topographic error measure the continuity of the mapping in the neural network. Given a sample vector $x \in M$, nearest weight vector of x in input space is w_i and second nearest is w_j. If the topology is preserved, some of the points in M between x and w_j are mapped to w_i, while the rest are mapped to w_j. If the corresponding neurons n_i and n_j are adjacent, the mapping is locally continuous. If they are not adjacent, the mapping is said to have a local discontinuity or a local topographic error (Kiviluoto, 1996). The topographic error ε_t for the whole map is obtained by summation of the all sample vectors. The topographic error is formally defined as below, where the better topography preservation is elicited by a lower error value.

$$u(x_k) = \begin{cases} 1, & \text{best and second-best units non-adjacent} \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

$$\varepsilon_t = \frac{1}{N} \cdot \sum_{k=1}^{N} u(x_k) \tag{4.5}$$

Zrehen measure

Zrehen (1993) proposed a measure that quantifies local consistency in topographic maps, which measures the separation in feature space of neurons that are neighbours in map representation space. Zrehen measure use the Delaunay triangulation (Chew, 1989) to determine which neurons are neighbours. A pair of neighbouring neurons r and s are locally

Table 4.1 Fundamental Clustering Problems Suite

Dataset Name	Problem	Size	Dimensions	Classes
Hepta	Clearly defined clusters, different variances	212	3	7
Lsun	Different variances and inter cluster distances	400	2	3
Tetra	Almost touching clusters	400	3	4
Chain Link	Linear not separable	1000	3	2
Atom	Different variances and linear not separable	800	3	2
Engy Time	Gaussian mixture	4096	2	2
Two Diamonds	Cluster borders defined by density	800	2	2
Wingnut	Density vs. distance	1070	2	2
Golf Ball	No clusters at all	4002	3	2

organized if every point in the straight line joining their weights vectors w_r and w_s is closest to them than to any other pointer in the map. Equivalently, there should not be any w_t in the sphere with radius $(w_r - W_s)/2$ and center $(w_r + w_s)/2$. Any weight vector w_t that violates the following condition is called an *intruder*.

$$||w_r - w_t||^2 + ||w_t - w_s||^2 \leq ||w_r - w_s||^2 \tag{4.6}$$

The Zrehen measure (ZM) is the sum of all intruders over all pairs of neighbouring neurons, where the better topography preservation is elicited by lower Zrehen measure. In order for better comparison, we present a normalized Zrehen measure proposed by Yarrow *et al.* (2014).

4.1.4 Topographic Evaluation of TGSOM

We validate the topology preservation of TGSOM using the *Fundamental Clustering Problems Suite* (FCPS) dataset, which offers a variety of clustering problems any algorithm would have to face in order to handle real world data (Ultsch, 2005). All the datasets are synthetically generated to be simple and visualized in two or three dimensions. Each dataset represents a certain problem that is solved by known clustering algorithms with varying success. The dataset is described in Table 4.1 and illustrated in Fig. 4.2.

Topographic evaluation of the TGSOM was conducted using 9 individual datasets from the FCPS dataset suite. We evaluated a range of transience values for age threshold ($M \in [1,4,10,20,40,80,120]$ iterations) for 3 levels of granularity ($SF \in [0.3,0.6,0.8]$). TGSOM was trained for 100 learning iterations (T_L) and 60 smoothing iterations (T_S). Factor of distribution (FD) was selected as 0.1 and R as 3.8 based on empirical findings from Alahakoon *et al.* (2000).

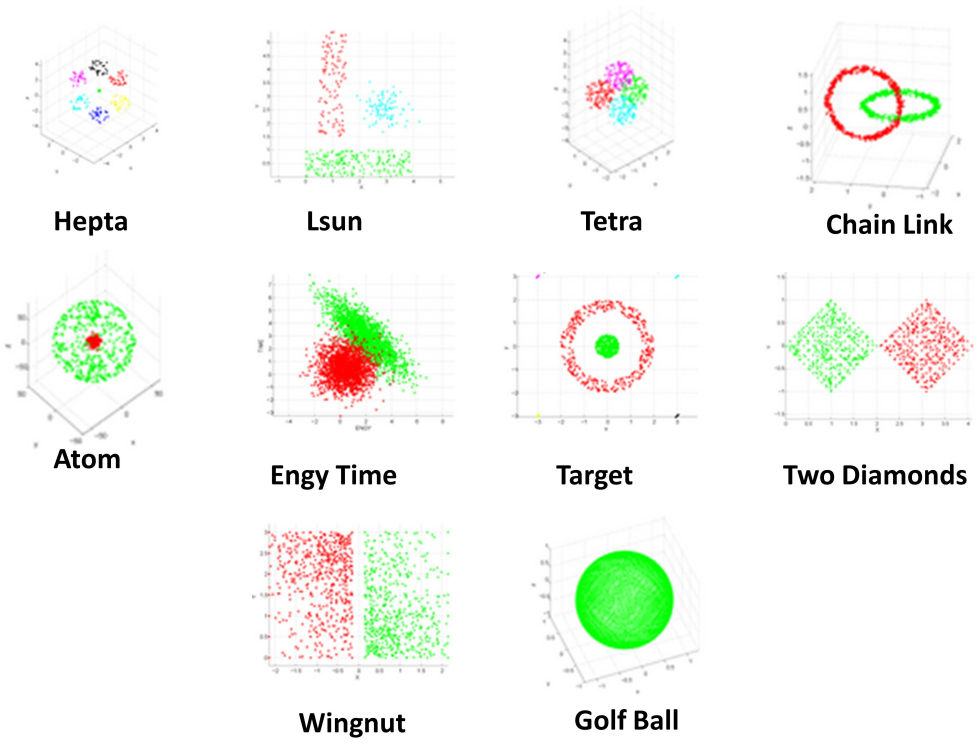


Fig. 4.2 Data distribution of FCPS

Analysis Topology Preservation

The *Topographic error* and *Zrehen measure* for combinations of TGSOM configurations are presented in Fig. 4.3. The Hepta detaset consists of clearly defined clusters with different variances. The topographic error for Hepta datasets at lowest granularity ($SF = 0.3$) is consistent throughout increasing transience thresholds (M), however, the Zrehen measure increases gradually and then reduces with M attains a lower value. For higher granulates $SF \in [0.6, 0.3]$, both the measures are consistent up to maximum M, and lowest when there is no transience. It is to be noted that when M is higher, the number of neurons increase, in turn require more computation resources.

Lsun dataset has different variances and inter cluster distances, and demonstrates a Gaussian distribution with better topography preservation at both lower and higher M. EngyTime (Gaussian mixture dataset) demonstrates consistent topology at different variations of transience thresholds.

Tetra (consists of almost touching clusters), Atom (different variances and linear not separable) and Golf Ball (no clusters at all) tends to have lower topology preservation at lower M and increase the topology with higher persistence. In contrast, Chain Link (contains

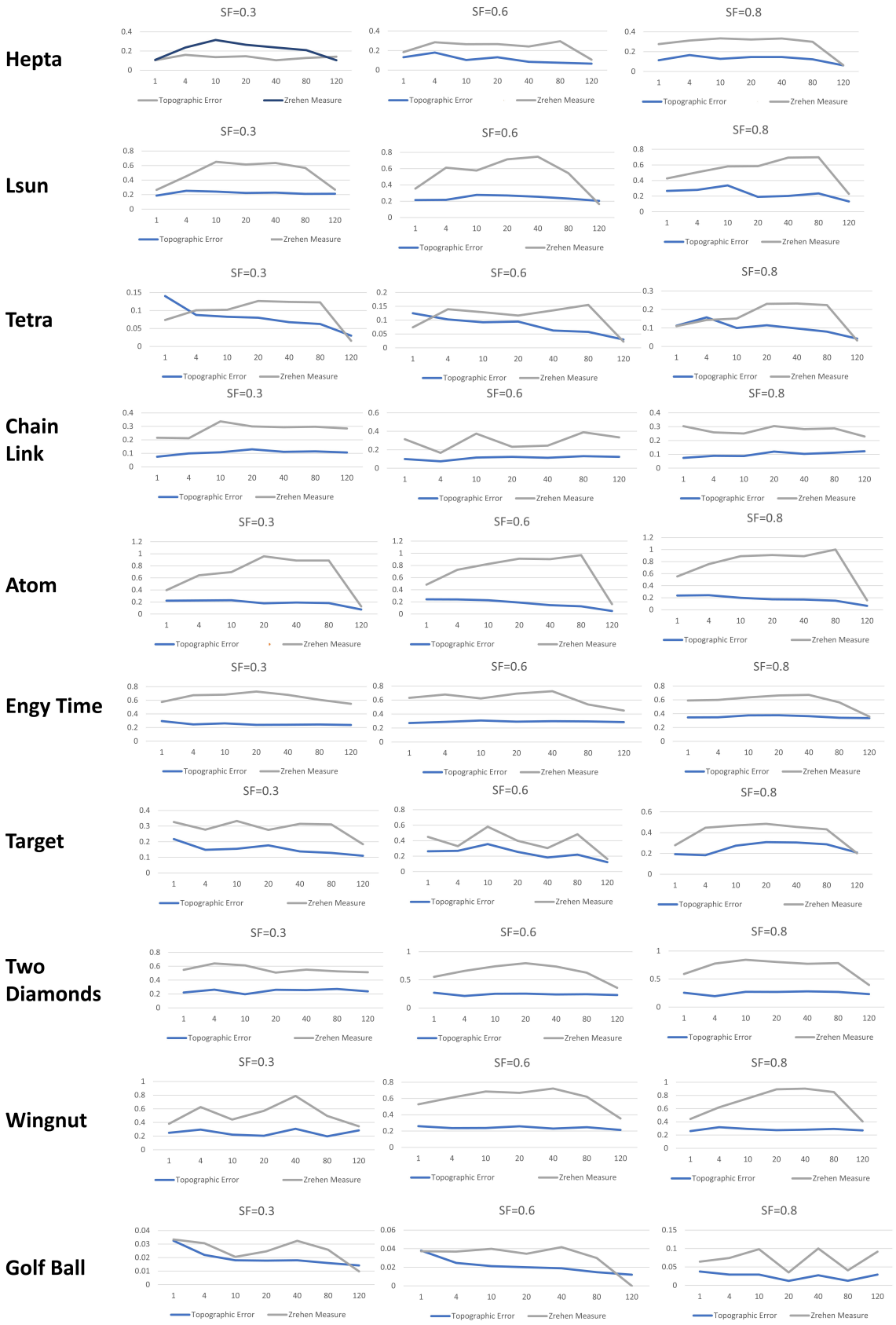


Fig. 4.3 Topography Evaluation for FCPS dataset suite

linear not separable data) shows promising topology preservation with higher transience, i.e., lower M values, with lower persistence.

Drawing on the topology evaluation for different types of data distributions, it is notable that a delicate balance between transience and persistence plays an eminent role for topology distribution. Distributions such as Chain Link provides better topology preservation with higher transience, Lsun, Target, Two Diamonds preserves topology better at either lower or higher persistence, while distributions such as EngyTime's topology is consistence disregarding the level of transience.

Thereby, we can conclude the importance of a delicate balance between persistence and transience in representation learning. As the natural environment (data space) changes and adapts, the artificial counterpart should also be adapted to accommodate these changes. Therefore, the balance of transience and persistence is critical in achieving the adaptation of the artificial counterpart.

Analysis of Cluster Accuracy

In addition to topology preservation, the evaluation demonstrates the inherent capability of TGSOM to be used as a clustering technique. Fig. 4.4 illustrates cluster representations for both linearly separable and non-separable diverse data distributions. For Hepta dataset, Atom dataset and Chain Link dataset, TGSOM representation has generated a separating line with hidden (removed) neurons, due to the use of transience in the TGSOM. This demonstrates an additional advantage of TGSOM over vanilla GSOM algorithm in representation learning.

Furthermore, we provide visualisations of all neuronal map representations of the evaluation detailing the TGSOM latent representations for all the parameter configurations for all the datasets of FCPS suite in Appendix A.1.

4.2 Incremental Knowledge Acquisition

The natural environment is composed of non-stationary sequential data, in which the sequence of data is given by series of temporally or spatially connected observations. Sequential data plays a vital role in biological perception system since all sensorimotor data are generated as sequences of time-varying stimuli. This section aims to investigate capabilities of existing computational models to capture sequential information from non-stationary data streams, thereby, leading to the development of a new unsupervised recurrent information processing mechanism to capture non-stationary sequential data in the proposed SSAI computational model (GSOM) that is used to materialize MSKRF as introduced in module (4.2) *Incremental Knowledge Acquisition* in the chapter overview presented in Fig. 4.1. This leads to integration

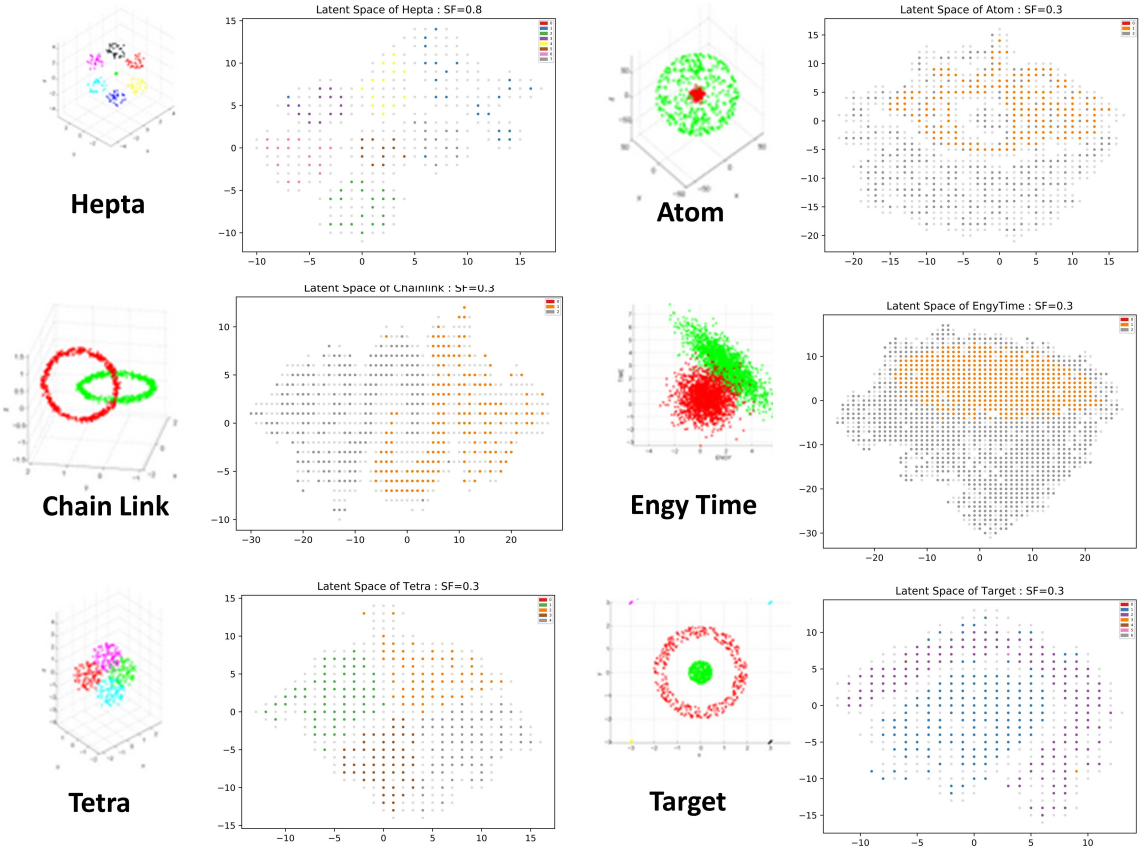


Fig. 4.4 Cluster Evaluation for FCPS dataset suite

of two biological bases; *Sequential information storage capability of neocortex* and *Auto-associative recall of information from neocortex*.

Voegtlin (2002) distinguishes two main approaches in representing the sequential nature of data as *explicit* and *implicit* representations. In explicit representation, time is considered as a dimension of space and delays are used to collect different measures of an input over a time window (Lang *et al.*, 1990). In implicit representation, time is represented indirectly. For instance, Recurrent Neural Networks with *leaky integration units* provide indirect temporal representation (Voegtlin, 2002), where the impact over the knowledge decays over time for a particular time instance. In general, explicit representations of time are limited and oversensitive to temporal deformations of the signal, while implicit representations are powerful and more robust to deformations. One drawback with recurrent input representation is that long-term information tends to decay exponentially over time. However, unless gradient methods are used, these limitations with implicit representations can be easily overcome.

4.2.1 Computational Models for Incremental Knowledge Acquisition

The Simple Recurrent Network (SRN) proposed by Elman (1990) is a well-known example for implicit representation of time. SRN consists of a modified variant of perceptron with a single hidden layer that uses a time-delayed replicate of its hidden layer activities as additional input. The learning happens through back-propagation over both input and hidden layers. Since SRN learns from its own past activities, the representation in its hidden layer is self-referent (Voegtlin, 2002).

Self-reference architectures for self-organization have also been introduced, where recurrent connections are added to original SOM architectures. Temporal Kohonen Map (TKM) is the first to introduce recurrent connections with self-organization (Chappell and Taylor, 1993). TKM consists of recurrent neurons in terms of leaky integrators. The computation of the distance of a neuron w_i from the input sequence $(x_1,...,x_t)$ at time t with similarity measure d_W is:

$$d_i(t) = \alpha \cdot d_w(w_i, x_i) + (1-\alpha) \cdot d_i(t-1) \tag{4.7}$$

where $\alpha \in (0,1)$ controls the rate of signal decay, denoting the quality of the representation of the current input and exponentially weighted past. However, there is no explicit back-reference to previous map activity with TKM, which means that the context is implicitly represented by the weights.

Voegtlin (2002) proposed RecSOM, to overcome the context limitation, introducing a less restricted recurrence that preserves information available within the activation at the last timestamp. In RecSOM, c_i is the context descriptors of each neuron and R_{t-1} is the context vector of the previous time step. N is the number of neurons in the map. $\beta \in (0,1)$ controls the rate of signal decay for the context representation. The activation for RecSOM is shown as below:

$$d_i(t) = \alpha \cdot d_w(w_i, x_i) + \beta \cdot d_i||c_i - R_{t-1}|| \tag{4.8}$$

$$R_{t-1} = (\exp(-d_1(t-1))), ..., \exp(-d_N(t-1))) \tag{4.9}$$

In the context of compact reference representation, MergeSOM combines a compact back-reference with a weighted contribution of the current input and the past (Strickert and Hammer, 2005). Each neuron is equipped with a weight vector w_i and a temporal context c_i. The latter represents the activation of the entire map in the previous time-step. The recursive

activation function of a sequence is given by the linear combination of equation 4.10 and 4.11,

$$d_i(t) = \alpha \cdot d_w(w_i, x_i) + (1-\alpha) \cdot d_w||c_i - C_t|| \tag{4.10}$$

$$C_i = \beta \cdot w_{I(t-1)} + (1-\beta) \cdot C_{I(t-1)} \tag{4.11}$$

where $\alpha, \beta \in (0,1)$ are fixed parameters, C_i is the global context vector and $I(t-1)$ denotes the index of the winner neuron (BMU) at $t-1$. Such context learning mechanism can be applied to lattices with arbitrary topology as well as to incremental approaches that vary the number of neurons over time.

The recurrent sequential information processing has been further adapted in growing variants of self-organizing neural networks. For instance, Growing Neural Gas (GNG) model equipped with context learning by Andreakis *et al.* (2009) (MergeGNG) that use activation functions defined in equation 4.10 and 4.11 to compute winner neurons and create new neurons with a temporal context. Parisi *et al.* (2016) proposed context learning for Growing-When-Required by introducing *Recursive GWR* algorithm, using the same activation function above.

4.2.2 Recurrent-TGSOM for Incremental Knowledge Acquisition

We introduce Recurrent TGSOM (RTGSOM) network with a context descriptor to capture the temporal resolution of the previous time-step. We modify the newly introduced TGSOM (Section 4.1.1) algorithm by modifying its activation function and learning rules to account for temporal information processing.

In this recurrent modification, we update the mechanism to select winner neurons and creates new neurons with a temporal context. In the RTGSOM, the BMU is calculated using:

$$d_i(t) = \alpha \cdot ||x(t) - w_i||^2 + (1-\alpha) \cdot ||c_i - C_t|| \tag{4.12}$$

$$C_i = \beta \cdot w_{BMU(t-1)} + (1-\beta) \cdot C_{BMU(t-1)} \tag{4.13}$$

where, $\alpha, \beta \in (0,1)$ are constants defined to modulate the influence of the current input and the previous input. Following the modification, the complete RTGSOM algorithm is presented in Algorithm 2

Algorithm 2: Recurrent Growing Self-Organization with Transience (RTGSOM)

Data: X

Parameters : SF(Spread Factor), M(Age Threshold), T_L(Growing Iterations), T_S(Smoothing Iterations), η_0(Initial Learning Rate)

Result: Neural Map (ϕ_t)

Start with a set of 4 neurons with randomly initiated weight vectors w_i, where $i \in [1,4]$ and $\lambda_i = 0$;
Initialize an empty global context $C_1 = 0$
Calculate GT for the given input space using equation 3.8 ;
`/* For all the learning iterations */`;
for $j \in [1, T_L]$ **do**
    for $x_t \in X$ **do**
        Update the learning rate η_t using equation 3.6 ;
        Select the BMU for x_t using equation 4.12 and 4.13 ;
        Set the transience value of BMU as zero, $\lambda_i = 0$;
        Update the weight of the BMU and its neighbourhood using equation 3.3 ;
        Calculate accumulated error for BMU, TE_i ;
        if $TE_i \geq GT$ **then**
            if *BMU is a boundary neuron* **then**
                Generate new neurons based on Section 3.4.1;
            else
                Distribute the error to its neighbors using equation 3.9 ;
            end
        else
            No node growth ;
        end
        Update the transience property λ_i of neurons using equation 4.1 ;
        Update ϕ_t using equation 4.3 ;
    end
end
`/* For all the smoothing iterations */`;
for $k \in [1, T_S]$ **do**
    for $x_t \in X$ **do**
        Update the learning rate η_t using equation 3.6 ;
        Select the BMU for x_t using equation 3.2 ;
        Update the weight of the BMU and its neighbourhood using equation 3.3 ;
    end
end

4.3 Multi-Stream Hierarchical Self-Organizing Architecture

This section intends to capacitate MSKRF in terms of processing multiple information streams by implementing a multi-stream architecture. This multi-stream architecture is designed to incorporate two biological bases; *Hierarchical abstraction in memory storage* and *Multi-modal information fusion* that are crucial to understand information obtained by multiple modalities and multiple streams. The discussion and proposal of new architecture in the section relates to the module (4.3) *Multi-Stream Hierarchical Learning* as presented in chapter overview in Fig. 4.1.

The natural environment is highly complex; thus, biological sensory system is equipped to sense this complex environment with multiple modalities. When an event occurs, more than one sensor detects the events, generating redundant neural signals. This underlies multi-sensory processing, that is of substantial adaptive value and has been extensively examined in the cerebral cortex of mammals (Stein and Meredith, 1993; Stein *et al.*, 2014). Intraparietal Sulcus (IPS) and Superior Temporal Sulcus (STS) are the two cortical regions in human brain that is responsible for processing of multi-sensory information in biological brain. Both IPS and STS receive convergent inputs from visual, auditory and somatosensory areas. The human sensory system utilizes two-streams for neural processing of vision and hearing, namely, Ventral and Dorsal streams. Ventral stream is involved with visual identification based on structural information, and dorsal stream is involved in processing relative location and motion information. These two streams converge at the STS cortical area in order to provide the overall representation of the stimuli received from the natural environment.

Similar to the multi-sensory convergence in human brain, an equivalent mechanism is vital for its artificial counterpart to represent and perceive the natural environment. Due to the advent of sensing technologies, today, machines have the capability to represent the real-world in machines using multiple modalities and the development of computational processing methods allow to capture each modality in multiple characteristics. For example, advanced intelligent video surveillance systems capture surveillance contexts using multiple cameras, whereas it is possible to represent each video stream using multiple spatio-temporal characteristics, such as motion flow, gradient structure, colour distribution, depth information, etc. Combining this multitude of modalities and characteristics provide the means of optimally represent the input space without hindering the full potential of intra-modality and intra-characteristic patterns of the input streams.

Drawing from the multi-sensory information fusion phenomenon in biological brain, Simonyan and Zisserman (2014a) proposed a deep learning based approach based on two

separate recognition streams (spatial and temporal), which are then combined by late fusion. This architecture was demonstrated using human action recognition from video data. The spatial stream performs action recognition from still video frames, while the temporal stream performs on dense optical from consecutive video frames. Both these streams are implemented as Convolutional Neural Networks (CNN). However, due to the supervised nature of CNN architecture, the method requires large volumes of training data for successful performance.

Mici *et al.* (2018) presented a dual-stream self-organizing architecture for visual recognition of transitive actions comprising human-object interactions. This model composed of a hierarchy of Grow-When-Required (GWR) networks that learn prototypical representations of body motion patterns and objects, accounting for the development of action-object mappings in an unsupervised fashion. The evaluation of this methods using publicly available human action benchmark datasets are competitive with respect to supervised state-of-the-art approaches. Nonetheless, the network graph structure of the GWR depends on the locality of the input data. Therefore, the network can develop different dimensionality for different regions of the network, which can result in visualization difficulties and inability to focus on the representation space in different granular levels, i.e., zoom out of the network to visualize global patters and zoom in to visualize granular patters in the representation space, that are useful for data mining by identifying clusters in the data.

Inspired by this multi-sensory information fusion phenomenon in biological brain, we conceptualize, design and implement a general multi-stream architecture composed of growing self-organizing maps (RTGSOM) to process arrays of feature streams in order to utilize a multitude of modalities and characteristics of the natural environment. In our work, we attempt to address the limitations that exist in both supervised deep learning methods and unsupervised methods. The proposed multi-stream architecture comprises an array of hierarchically connected growing self-organizing structures (RTGSOM) to process a multitude of characteristic from input data streams. An overview of the proposed architecture is presented in Fig. 4.5. First, the multiple streams are extracted from the natural environment in *Feature Extractor* module. Second, each feature stream is presented to a particular feature stream (FS_i) in order to generate a representation of each stream based on structural adaptation. The primary structural adaptation is fed to the hierarchically connected self-organization structure in order to infer complex patterns from each respective feature streams. Ultimately, individual feature streams are fused to generate a holistic representation of the environment.

Based on the proposed multi-stream architecture, we conduct an experiment on processing video sensor data for human activity recognition. We emphasize the importance of using

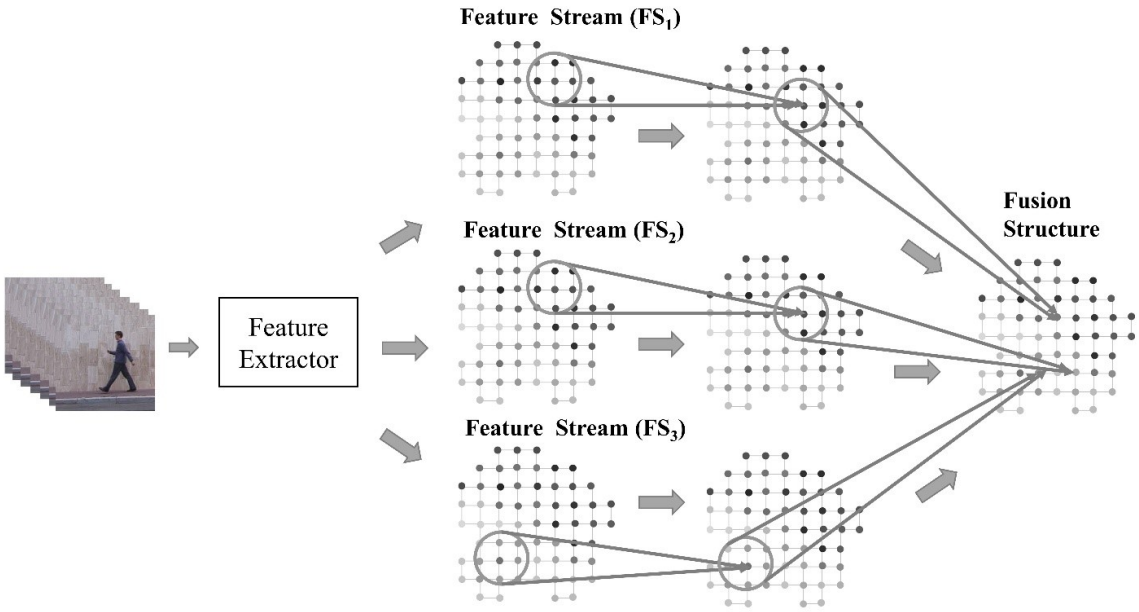


Fig. 4.5 Hierarchical multi-stream self-structuring Architecture

self-structuring as the foundation for learning that can adapt to the nature of the data and with the ability to learn from multi-modal unstructured data. We demonstrate the proposed architecture to confirm its validity and usability in real-world settings.

4.3.1 Self-Organization based Human Action Recognition

Human Activity Recognition (HAR) is a primary application area in video surveillance, which aims at identifying the activity of a person or a group based on visual and/or non-visual observation data from sensors such as cameras, accelerometers, as well the contextual knowledge on the background (Sargano *et al.*, 2017). HAR systems retrieve and process contextual data (environmental, spatial, temporal, etc.) to understand the human behavior, using the visual and non-visual data. In industry environments, HAR is predominant for development of smart homes, smart cities, transportation, healthcare, security and surveillance based applications (Basavaraj and Kusagur, 2017).

Current state-of-the-art visual analytics systems comprise of Convolutional Neural Networks (CNN), which have gained great success in visual analytics tasks such as object detection, scene and event classification, anomaly detection and HAR (Simonyan and Zisserman, 2014a). The deep CNN for HAR comprise of feature abstraction using hierarchically connected convolutional layers and multi-stream learning for spatial and temporal streams of training video data. These core concepts of CNN provide capability to learn descriptive

representations from raw visual data in large-scale supervised datasets (e.g., YouTube-8M). However, supervised learning techniques such as end-to-end deep convolutional neural networks learn from the examples in a training dataset, i.e., the learning algorithm is trained using a labeled dataset where each record is labeled with a known outcome. Due to the unavailability of labeled data and fast-growing nature of these high-dimensional video streams, the traditional approaches in pattern recognition and video processing face many challenges which in turn provides opportunities to advance machine learning applications which are unsupervised, scalable, self-structuring and are more robust. In this light, AI systems of the future is expected to incorporate higher degrees of unsupervised learning in order to generate value from unlabeled data.

To this end, we propose a HAR model that is able to capture multiple feature streams from videos and produce a unified stream of insights. We believe that this would create a better representation of the input video data, enabling a more accurate, robust processing of streams of video data. Most importantly, we demonstrate the newly introduced transience property (Section 4.1.1) and recurrent learning capability (Section 4.2.2) of the proposed RTGSOM algorithm in this intelligent HAR approach.

Two-Stream Self-Organization for HAR

In order to experiment the proposed method for HAR, we design the multi-stream architecture to accommodate information streams of video data. This study explores two-streams, i.e., spatial and motion characteristics, for activity recognition. Representation of spatial characteristics of video streams can be made using domain expert designed descriptors and detectors such as Hessian3D, Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) and Enhanced Speeded-Up Robust Features (ESURF) (Sargano *et al.*, 2017). Motion characteristics can be described using optical flow based features, which have been more frequently used in recent research (Zha *et al.*, 2015). In this work, we utilize histogram of oriented gradients (HOG) (Dalal and Triggs, 2005b) to represent spatial characteristics and histogram of oriented optical flow (HOOF) (Chaudhry *et al.*, 2009) to represent motion characteristics of action video, given their success that have been proven in recent attempts (Hasan *et al.*, 2016). The two-streams are eventually fused in order to provide a holistic latent representation of the human activity space. Subsequent to the generation of the fused self-organized latent structure, frequency-based label distribution is associated for activity classification.

Adopted from the hierarchical multi-stream self-structuring architecture proposed in Section 4.3, we illustrate the model architecture for the proposed HAR approach in Fig. 4.6.

It functions in three phases; 1) Action Representation, 2) Self-Organization, and 3) Action Classification. Each phase is explicated in following subsections.

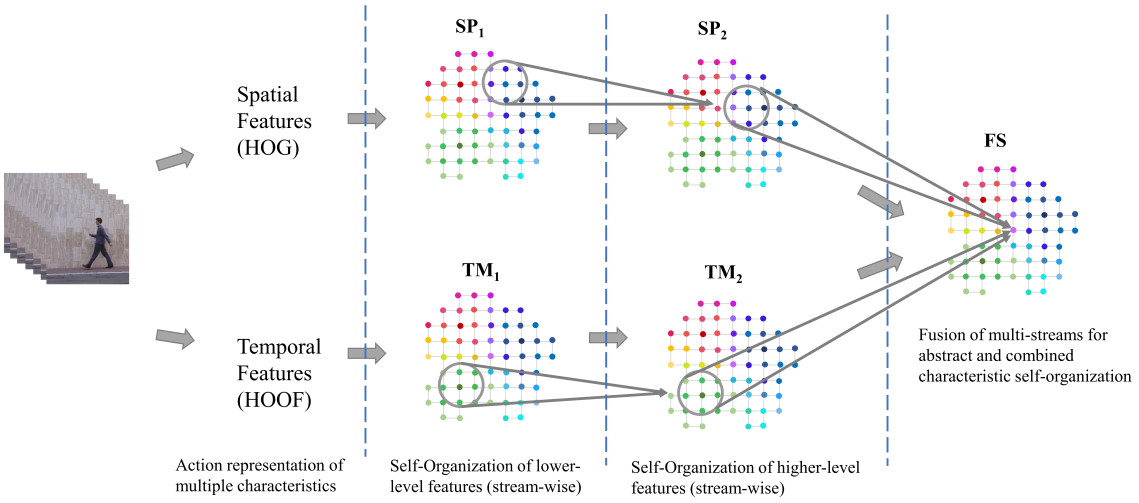


Fig. 4.6 Hierachical Two-stream Growing Self-Organizing Map Structure for Human Activity Recognition. SP_i: Spatial self-organization structure, TM_i: Temporal self-oranization structure, FS: Holistic latent representation generated by fusion of multiple streams.

Action Representation

To represent action videos, we utilize two primary characteristics of the action video: 1) Spatial structure of video frames, and 2) Motion of the actions occur in the video. The spatial structure is represented using histogram of oriented gradients (HOG), in which, the local object appearance and shape of an image is described by the distribution of intensity gradients (i.e., edge directions). The video frames are divided into $n \times m$ non-overlapping regions and gradient is calculated for each pixel. The calculated gradients are normalized using the intensity across the block. This gradient calculation is used to compile the histogram of gradient directions.

Subsequently, the motion of the actions in the video are represented using histogram of oriented optical flow (HOOF). First, the optical flow is calculated for neighboring video frames to estimate magnitude and direction of each individual pixel. Then the optical flow magnitude is voted into k directions by the direction of optical flow in order to obtain a k-bin histogram as HOOF, which represents the temporal characteristic of action videos. The HOOF estimation approach is illustrated in Fig. 4.7.

Self-Organization

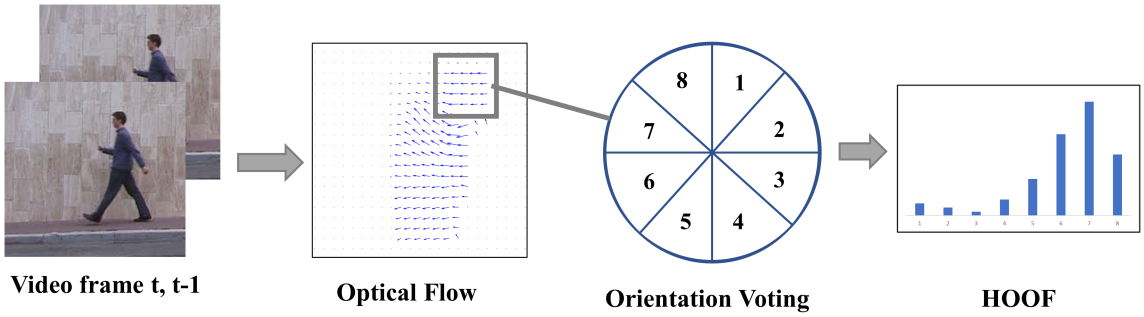


Fig. 4.7 Histogram of oriented optical flow estimation approach, k=8

In the two-stream growing self-organization architecture, we model RTGSOM as connected hierarchical layers in order to capture increased complex patterns with higher abstraction across the layers from activity video footages. However, increasing the number of layers has an adverse effect on the computational complexity of self-organization. Thus, in designing the number of hierarchical layers, the trade-off between computational complexity and performance should be taken into consideration. In this work, we model our network as two connected layers to capture the pattern abstractions without hindering the computational complexity. As illustrated in Fig. 4.6, the second layers of RTGSOM (SP_2 and TM_2) will capture increased complex patterns from the activation patterns learned in the first RTGSOM layers (SP_1 and TM_1) to learn the spatial and temporal structure respectively, from action videos.

From the data stream perspective, in the first layer of our architecture, RTGSOM networks are structurally adapted with respect to each stream. Once the first layer organization is completed, it is then used to compose the dataset for the second self-structuring layer. Therefore, after the first layer, we evaluate the latent representation for each input $x(t)$ in the dataset (X). The dataset for the second layer is composed as winner nodes (BMU) for each input $x_i \in X$, as shown in equation 4.14.

$$X_2 = \{w_{BMU(x_1)}, w_{BMU(x_2)}, ..., w_{BMU(x_3)}\} \tag{4.14}$$

The activation patterns from the two streams are fused into a single latent representation at the third RTGSOM layer (fusion layer), and then instantiate the self-organization to represent a holistic human action representation. This multi-stream hierarchical network structure enables multiple characteristics of activity videos to be exploited to derive complex activation patterns for the creation of a holistic representation, which can be used for visual exploratory analytics and activity classification.

At the fusion layer, the data stream is transformed based on the previous layer, where the winner node is selected for each data sample for each stream using equation 4.14. Then the training dataset (X_F) for final fusion layer is given by the horizontal concatenation of the multiple streams, as shown in equation 4.15, where $s_i(i \in [1,r])$ indicates the number of feature streams. In this case, $r = 2$ as we have two streams for spatial and temporal.

$$X_F = \{X_2^{s1} \cup X_2^{s2} \cup ... \cup X_2^{sr}\} \tag{4.15}$$

Action Classification

Human activity classification aims to predict the activity label of unseen activity video samples. Such classification will provide means in automated systems for assisted living in smart homes, healthcare monitoring applications, monitoring and surveillance systems. For this purpose, we extend the proposed approach by adopting the minimal-distance labelling function designed for Growing Neural Gas based learning (Parisi *et al.*, 2015).

The neurons in the FS layer is assigned with a distribution function for available action label at the learning phase. Given that L action labels are available in the activity video dataset, we associate the frequency of each label $l(l \in L)$, which is associated with neuron u_i in the neural map. For the action classification, we extend the Algorithm 2, where at step 27 in the last smoothing iteration, we associate a class label l for each neuron based on a minimal-distance strategy. Thus, subsequent to the selection of BMU at step 27, the action label of the current input sample $x(t)$ is associated with the best matching unit. It is to be noted that the self-organization process is completely unsupervised, thus, no labeled data in required. However, in order to classify the actions into human interpretable classes such as walk, skip, jump, run, etc., and to evaluate the classification accuracy, we adapt the labelling function.

In the activity classification phase, unseen activity video samples are processed by the two-stream hierarchical self-organization network by selecting the BMU at each layer using equation 4.12 and 4.13. For each action video, the neuronal weight of each layer is presented as the input for next layer. The activity label associated with the BMU at the FS layer is selected as the class label for each unseen action video sample.

4.3.2 Experiment on Human Activity Recognition

The proposed self-structuring multi-stream video analytics approach is evaluated using a real-life application scenario of human activity recognition (HAR). We utilize three bench-

mark video datasets that encompass diverse human activities for the evaluation. With the experiments, we demonstrate three-fold capabilities of our approach.

1. Demonstration of the newly introduced transience property with TGSOM algorithm to improve plasticity and reduce the likelihood of overfitting and the influence of outdated information during knowledge acquisition.

2. Demonstration of the RTGSOM that accounts for sequence information using recurring learning concepts.

3. Demonstrate the proposed hierarchical two-stream growing self-organizing approach for HAR to accommodate learning from unlabelled video data and diverse characteristics.

Datasets and Feature Extraction

The proposed hierarchical two-stream growing self-organizing approach for HAR was evaluated using three benchmark datasets: Weizmann (Blank *et al.*, 2005), KTH (Schuldt *et al.*, 2004) and UCF11 (Liu *et al.*, 2009) human activity datasets. Weizmann dataset contains human actions of 9 different subjects, where each clip lasts for approximately 2 seconds at 25Hz. KTH dataset contains human actions of 25 different subjects, where each subject has approximately 24 segments (total of 599 action video segments). Each video is sampled at 25Hz and lasts between 10 to 15 seconds. Both datasets contain actions such as boxing, hand clapping, hand waving, running, walking, jogging, bending, etc. captured in indoor and outdoor environments. The sample video snaps from the two datasets are illustrated in Fig. 4.8 and 4.9.

The UCF11 (YouTube Actions) dataset is considerably challenging due to large variations in camera motion, illumination, viewpoint and cluttered background. This contains 11 complex human action classes inclusive of shooting a basketball, cycling, diving, swinging a golf club, horseback riding, juggling a soccer ball, swinging motions in tennis, jumping on a trampoline, spiking a volleyball, and walking a dog. Snapshots from the UCF11 YouTube action dataset is illustrated in Fig. 4.10.

For action representation, we extracted raw action video frames, convert into gray-scale and resize to 32 x 32 pixels. The processed frames were normalized to range 0-1 and subtracted by mean in order to centre the input video data. The HOG and HOOF features are extracted from the pre-processed video frames as 8-bin histograms.

Self-Learning Parameters

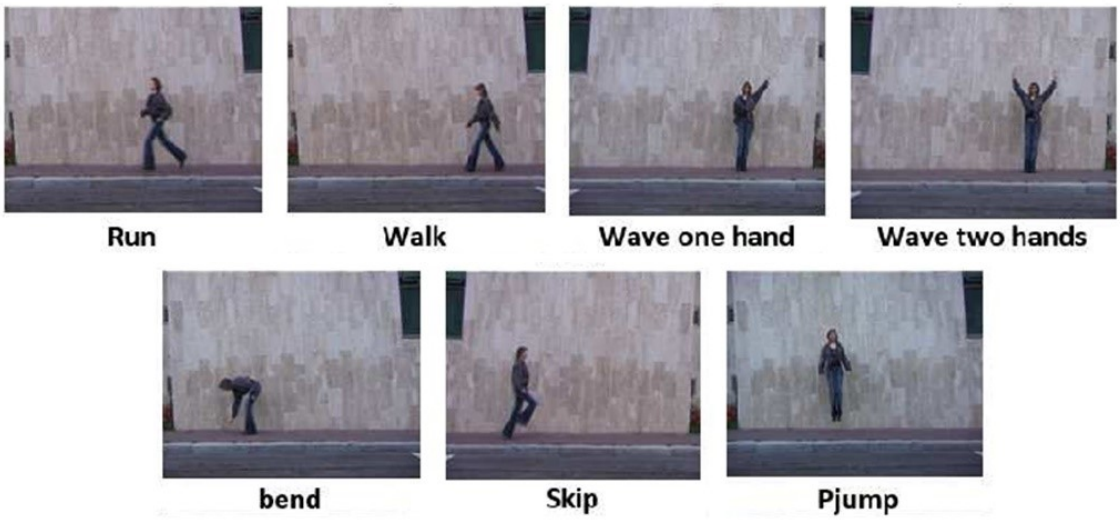


Fig. 4.8 Sample Video Snaps from Weizmann Dataset

In the growing phase, extracted HOG and HOOF features are fed to spatial stream and temporal stream respectively. The spread factors (SF) for spatial stream layers are selected as, $SP^1 : 0.8$, $SP^2 : 0.6$, for temporal streams, $TM^1 : 0.6$, $TM^2 : 0.5$, and for the fusion layer, $FS : 0.4$. The factor of distribution $FD = 0.1$, both learning and smoothing iterations are set to 50, $P^{BMN} = 1$, $P^{Neigh} = 0.5$, $R = 3.8$. The transience threshold (M) is setup as $M^{(Layer1)} = 3$, $M^{(Layer2)} = 2$ and $M^{(FusionLayer)} = 2$. The optimal values for SF were selected using a random grid search method (Bergstra and Bengio, 2012). The higher values for SF lead to a greater number of neurons while representing the input data space effectively yet resulted in higher time complexity. For instance, selecting $SF = 0.9$ for the first layers (i.e., SP^1 and TM^1) resulted the first layer GSOM representations larger networks with higher computation time for training. On the contrary, a lower value of SF ($SF = 0.3$) resulted the first layer with a lower number of neurons that were densely clustered and the actions were overlapped restricting an accurate classification. Thereby, considering both accuracy and computation complexity, optimal values for SF were selected.

The experiments were carried out on a multi-core CPU at 2.4 GHz with 16 GB memory and GPU of NVIDIA GeForce GTX 950M. Results for these experimental evaluations are prompted in subsequent sub-sections.

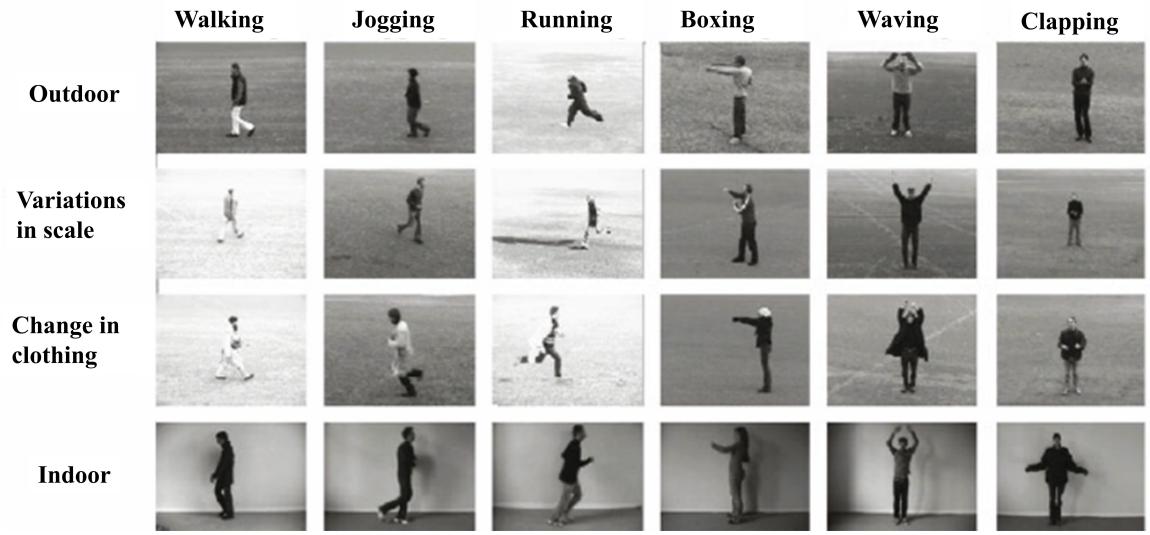


Fig. 4.9 Sample Video Snaps from KTH Actions Dataset

Representation of Video Actions in Latent Representation

The self-structuring of multiple feature streams, and that the capability to infer different patterns unique to each characteristic stream are demonstrated in this section. The representation outcomes are illustrated in Fig. 4.11.

In the first layer, i.e., TM_1, the curve A distinguish moving activities such as run, walk and skip with respect to stationary activities such as jump, bend, wave, etc. Within the stationary activities, the circled region B encompasses full body activities such jump and bend, and are distinguishable from hand-only activities such as single-hand wave and two-hands wave. The second layer, TM_2, further improved the activity separation in the neural map. In TM_2, activities: jump, two-hand wave and bend can clearly be distinguished by circled regions C, D and E respectively. This improvement of activity disambiguation confirms the robustness of the self-organization approach and the ability to capture increasingly complex patterns when the hierarchy deepens.

In contrast to conventional self-organizing maps, the novel transience property has enabled the new self-organization process to remove unused neurons (i.e., obsolete knowledge). This is evident by the clear separation of moving and stationary actions at TM_1, where the unused nodes in the separation line A have been removed. Further, the separation lines of different actions in TM_2 shows partial node removal making the separation clear and concise. This experiment proves that beside the inherent capability of self-organization in RTGSOM, the novel transience property enables clustering within the self-organization process itself, which can ideally be used for visual exploratory analytics.

Fig. 4.10 Sample Video Snaps from YouTube Actions Dataset

Hierarchical Representation of Increasingly Complex Patterns

We evaluated the RTGSOM map of each hierarchical layer in both the spatial and temporal streams in order to identify their representation capabilities. For this ablation evaluation, each RTGSOM network is assigned with action labels for activity classification on Weizmann and KTH video datasets. To compute the classification accuracy, we split both datasets on 70% for self-organization (action labels are ignored in learning) and 30% for classification (action labels are used for evaluation). The classification is measured using precision (p), recall (r) and F1-Score (F). Precision is the fraction of correctly classified activities among all the positively classified activities, whereas recall is the fraction of correctly classified activities that have been classified over all the relevant activities. F1-Score is the harmonic mean of precision and recall as defined in equation 4.16. The ablation analysis results are presented in Table 4.2.

$$F1 = 2 \times \frac{p \times r}{p + r} \tag{4.16}$$

In the activity classification of the Weizmann dataset, F1-Score of the spatial stream is improved from 0.43 to 0.59 and the temporal stream is improved from 0.71 to 0.85. Similarly, for the KTH dataset, spatial stream is improved from 0.44 to 0.66 and temporal stream is improved from 0.65 to 0.76. This improvement of the F1-Score signifies that when the GSOM hierarchy moves to deeper layers, the representational capability is enhanced. Thus, the

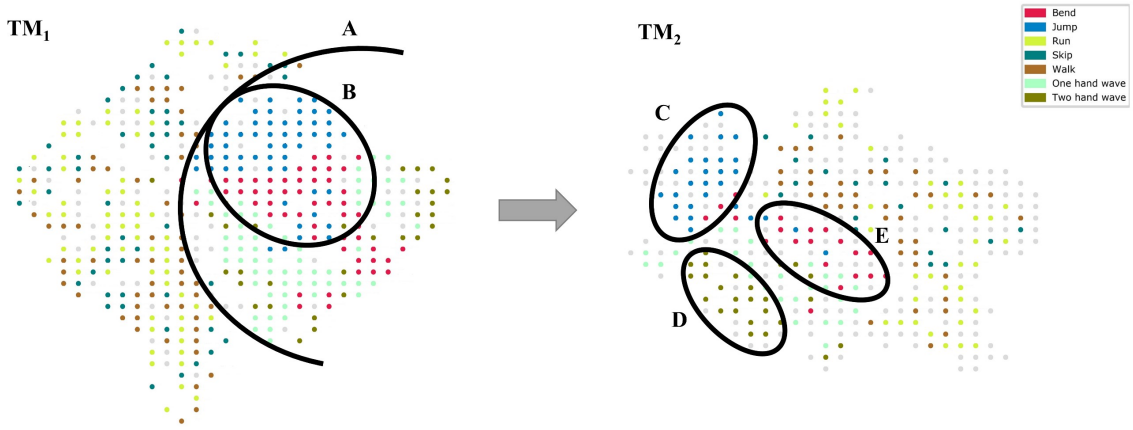


Fig. 4.11 Self-learning in the temporal stream for Weizmann dataset

Table 4.2 Ablation Analysis

Dataset	Layer	Precision	Recall	F1-Score
Weizmann	SP1	0.47	0.40	0.43
	SP2	0.60	0.58	0.59
	TL1	0.80	0.63	0.71
	TL2	086	0.83	0.85
	FS	**0.95**	**0.94**	**0.95**
KTH	SP1	0.45	0.44	0.44
	SP2	0.69	0.64	0.66
	TL1	0.65	0.64	0.65
	TL2	0.81	0.72	0.76
	FS	**0.92**	**0.88**	**0.90**

evaluation results confirm the hierarchical GSOM structure can capture increased complex patterns with higher abstraction when the hierarchy deepens.

In the proposed model, the activation patterns from both spatial and temporal streams are fused at the fusion layer (FS). For both datasets, the activity classification has improved significantly in the FS layer, respectively Weizmann dataset and KTH dataset achieving a F1-Score of 0.95 and 0.90. This justifies the use of multiple characteristics from the input video data, process as individual streams and combine ultimately in order to provide a holistic representation. Thus, the activity classification has resulted in a better fused self-organization layer.

Human Action Classification

The human action classification results are evaluated across the three benchmark datasets. The confusion matrix for the selected datasets are shown in Fig. 4.12. It can be observed that among the misclassified actions, mostly the labels of moving activities such as walk, run, jog and skip have been interchanged. Among the stationary actions, boxing, clapping and hand-waving have been interchanged. The misclassifications were mostly caused by the similar motion flow of the said actions.

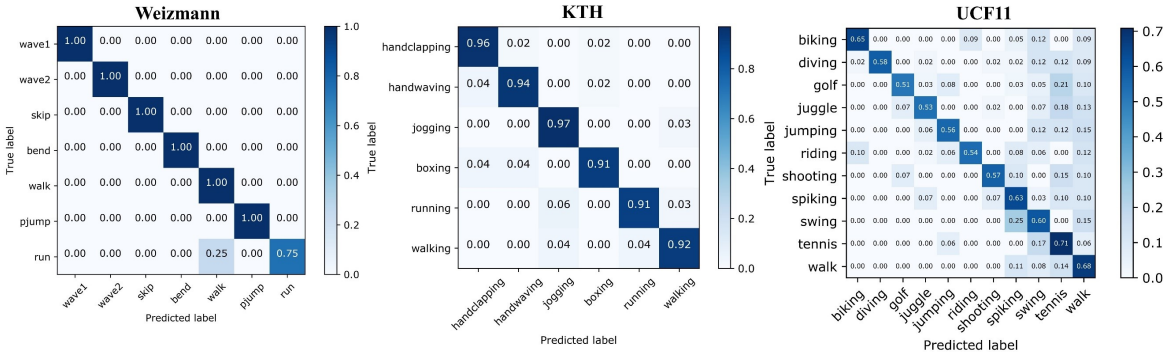


Fig. 4.12 Confusion Matrix for Activity Classification

We evaluate the accuracy of human action classifications with respect to seven benchmark HAR methods. First, an unsupervised activity and gesture recognition approach by Yang *et al.* (2012) that automatically discover vocabulary of actions using raw optical flow. Second, a semi-binary video descriptor in the context of activity recognition proposed by Umakanthan (2016), which use motion and appearance modelling for supervised classification. Third, a hierarchical clustering approach for joint human action grouping and recognition proposed by Liu *et al.* (2016). Fourth, an unsupervised video action clustering approach using spatial and temporal interaction constraints proposed by Peng *et al.* (2018).

Due to the complex characteristics of the UCF11 human action dataset, mostly it has been used with supervised machine learning. Therefore, we selected two supervised methods and one semi-supervised method to compare the proposed method. Hasan and Roy-Chowdhury (2014) method on incremental activity modelling in video streams and Kim *et al.* (2017) method on memory-based embedding for action representation, used as supervised methods, whereas, Zou *et al.* (2018) method on hierarchical temporal memory based model for one-shot distance learning that can use with limited training samples, selected as the semi-supervised HAR method.

The comparison is presented in Table 4.3, where the results appear as reported by respective authors. Among the compared methods, our approach outperformed the selected

Table 4.3 Classification Results

Proposed Method	Classification Approach	Accuracy (%)		
		Weizmann	KTH	UCF11
Yang *et al.* (2012)	Unsupervised	91.0	91.0	-
Umakanthan (2016)	Supervised	-	91.2	-
Liu *et al.* (2016)	Unsupervised	-	**94.3**	-
Peng *et al.* (2018)	Unsupervised	77.8	83.4	-
Hasan and Roy-Chowdhury (2014)	Supervised	-	91.0	54.5
Zou *et al.* (2018)	Semi-supervised	-	-	**60.3**
Kim *et al.* (2017)	Supervised	-	-	50.8
Ours	Unsupervised	**96.4**	93.6	59.7

approaches by achieving 96.4% accuracy in Weizmann dataset. For the action classification in KTH dataset, our approach obtained on-par results with the state-of-the-art by achieving 93.6%, where Liu *et al.* (2016) superseded the accuracy only by 0.7%. However, as Liu *et al.* (2016) used BoW model to concatenate spatial and temporal characteristics into a single feature vector, it will hinder the capabilities to discriminate characteristic dependent patterns, which limits the capabilities to conduct comprehensive visual exploratory analytics on the activity video streams.

Obtaining an accuracy of 59.7%, our proposed model shows on-par with the Zou *et al.* (2018) (a semi-supervised method) which only lead by 0.6%, in the evaluation of complex UCF11 dataset. With this evaluation, our proposed model presents promising results even by using unlabeled data, which justifies the usability of unsupervised self-organization for HAR.

A runtime analysis was conducted on the human activity classification to evaluate the computational overhead of the activity classification workflow. We calculated the runtime for each process: activity representation, self-organization and activity classification in terms of seconds per frame, as presented in Table 4.4. The average time complexity for a single frame to be processed and classified has accumulated to 221.9 milliseconds, thus, the proposed approach is able to achieve approximately 4.5 frames per second (FPS), enabling near real-time human activity classification. From the analysis, it is evident that the maximum computation occurs at the parsing the activity representation through the self-organization architecture. Thus, by speeding up this process using paralleled implementation, it would be possible to further enhance runtime efficiency.

Discussion

We have introduced a new unsupervised machine learning approach to process video sensor data and extract useful insights. By combining the underlying concepts of self-structuring,

Table 4.4 Runtime Analysis

Process	**Runtime (ms)**
Action Representation	12
Self-Organization	144.7
Action Classification	65.2
Total	221.9

transience and recurrent learning, we proposed "Hierarchical Multi-Stream Recurrent Growing Self Organizing Maps with Transience", which can be utilized for video processing accommodating fast-growing high-dimensional unlabeled video streams. The proposed model acquires knowledge from the unlabeled human activity data and process through hierarchical two-stream pathways.

The proposed approach addresses the key limitations posed by self-organization, 1) inability to accommodate the temporal nature of the input data, specifically the video data, 2) inability to accommodate multiple streams/channels of input data, and 3) overfitting and the influence of outdated information on the acquired knowledge. We address the inability to accommodate the temporal nature of the input data by implementing recurrent leaky integration of time-dependent sequential input information from the natural environment. Hierarchical connections provide the capability to learn higher abstractions from input data, whereas the two streams enable the algorithm to separately learn from complementary characteristics of input data, providing a holistic representation. We address the overfitting and the influence of outdated information on the acquired knowledge by implementing a transience property in the algorithm.

We position the proposed extension to core GSOM algorithm in order to incorporate the biological bases through the MSKRF as presented in Fig. 4.13, where the uncovered regions show the components introduced in Chapter 4. Surveillance cameras (CCTV) act as the sensor in the experimental context where we have implemented Latent Representation (LR) and Cognitive Representation (CR) using SSAI as the foundation. We used GSOM as the base SSAI algorithm, in which we cater each of the biological bases as denoted in red colour font. i.e.,

- *Sequential information storage* and *auto-associative recall of information* is implemented using recurrent learning mechanism in GSOM as presented in Section 4.2,
- *Hierarchical abstraction* is implemented as hierarchically connected RTGSOM layers as presented in Section 4.3,

- *Persistence and transience* are implemented as a strategic partial forgetting function in RTGSOM algorithm as presented in Section 4.1,
- *Invariant memory representation* has been an inherent capability associated with GSOM algorithm,
- *Multi-modal information fusion* is enabled through multi-stream RTGSOM architecture as presented in Section 4.3.

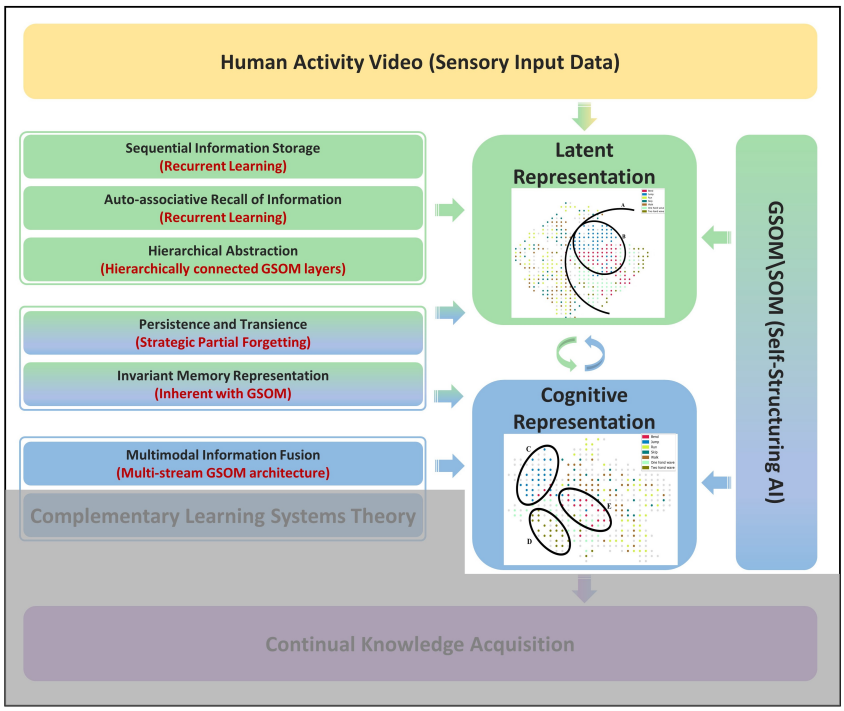


Fig. 4.13 Positioning the experiments in MSKRF.

The aforementioned capabilities were demonstrated using a human activity recognition video dataset by predicting the activity of previously unseen videos. Results from experiments conducted on three benchmark datasets demonstrate the human activity classification accuracy, robustness and visual exploratory analytics capability of the proposed approach, confirming its wide applicability in industrial and urban settings. We believe that these applications will have significant importance for IoT based data processing where a plethora of data are generated with previously unseen structures and nature. The ability to capture temporal relationships will be particularly important for fast-growing streams of data, specifically when considering real-time applications.

4.4 Summary and Research Questions Revisited

The intermediate knowledge representation mechanism of the MSKRF framework lies in the two hierarchically connected layers: latent (LR) and cognitive (CR) representations, which provides the foundation for learning from input by representing sensometry input stimuli from artificial somatosensory in a digital form. Chapter 3 conducted an exploration to identify options available in existing computational paradigms to select the most suitable computational model to facilitate the two representation modules, which resulted in the introduction of Self-Structuring Artificial Intelligence (SSAI) with an unsupervised learning paradigm as the solution for this. The Growing Self-Organizing Map (GSOM) was selected as the candidate algorithm to lay the foundation for these representation layers, as it has inherent SSAI capabilities.

The current chapter (Chapter 4) advanced the GSOM algorithm by combining the underlying concepts of self-structuring, transience and recurrent learning. Further, the modified algorithm was used to design and develop a multi-modal stream architecture named "Hierarchical Multi-Stream Recurrent Growing Self Organizing Maps with Transience", which was utilized for video processing accommodating fast-growing high-dimensional unlabeled video streams.

Three publications were originated from this chapter. First, the evaluation of topology preservation for the new GSOM with transience (TGSOM) algorithm was presented in the conference article entitled *HT-GSOM: dynamic self-organizing map with transience for human activity recognition* (Nawaratne *et al.*, 2019b). Second, the extended RTGSOM algorithm was presented in the journal article entitled *Recurrent Self-Structuring Machine Learning for Video Processing using Multi-Stream Hierarchical Growing Self-Organizing Maps* (Nawaratne *et al.*, 2020a). Third, the proposed multi-stream self-organization architecture was presented in the journal article entitled *Hierarchical Two-Stream Growing Self-Organizing Maps with Transience for Human Activity Recognition* (Nawaratne *et al.*, 2019a).

With that, the current chapter partially addressed the second research question (RQ2), that stated **"What are the computational and machine learning constituents of continuous lifelong learning for materializing the proposed conceptual framework?"**. The RQ2 consists of 5 sub-questions, where the current chapter successfully addressed sub-questions RQ 2.3 and RQ 2.4 as follows.

RQ 2.3) How can the knowledge embedded in computational models preserve stability and plasticity when introduced to continuous data streams?

To address RQ 2.3, we first discussed the importance of transience as a mandatory capability in a successful memory system to retain memory plasticity. In section 4.1, we provided a detailed review on how the existing computational models attempted to achieve plasticity without compromising the stability of the memory system. Thereby, we extended the GSOM algorithm, which we selected as the base framework for self-structuring to develop the representation module in MSKRF framework, by incorporating transience capability.

Followed by the algorithmic extension for transience, we evaluated the performance of TGSOM (transience GSOM) using a suite of synthetic datasets that consist of several datasets with varying difficulties and properties resembling problems in real world. The experimental evaluation demonstrated that the TGSOM has the capability to facilitate to encapsulate plasticity in the neuronal latent representation without the loss of stability. Further, the experiment demonstrated how the self-organization with transience will discard the outdated information and overfitting knowledge in its knowledge acquisition, without the loss of stability.

Incremental knowledge acquisition is an essential characteristic for a self-organizing network due to the heavy focus of both spatial and temporal structure of natural data streams. Section 4.2 provided a review of existing methods that incorporate temporal information processing, and lead to the development of recurrent self-structuring mechanism based on TGSOM algorithm.

RQ 2.4) With multiple facets and characteristics of data being captured to represent actions, events and situations, how can a comprehensive representation be developed in digital environments?

In section 4.3 we proposed a hierarchical multi-stream architecture that is able to capture multiple feature streams from videos and produce a unified stream of insights. The proposed architecture is based on RTGSOM that introduced transience by drawing parallels between neurophysiology and computational mechanisms in order to implement strategic forgetting in the proposed intelligent HAR approach, and recurrent learning behavior for the GSOM where it accounts leanings from previous knowledge captured by the GSOM. We demonstrated the proposed model using three benchmark video datasets and the results confirm its validity and usability for human activity recognition.

In despite the strengths and power of RTGSOM to represent the natural environment for both temporal and spatial streams, the evolving nature poses challenges in adapting the current computational models to represent the data. This is primarily due to the fact that not all the possibilities in data are presented at once but comes to light in different time

periods. For instance, in a smart city video surveillance system, what is considered as normal behaviour evolves over time making current knowledge incomplete and/or obsolete. When offenders become aware of detected anomalies, they can maliciously adapt behaviours so that subsequent anomalies are difficult to detect. Or in a separate instance, given a medical diagnosis system, the diseases known to practitioners might be only a small subset of all the possible diseases. Thus, the system should evolve based on novel diagnosis made by practitioners and medical researchers. Thereby, the digital representation ecosystem is in need for the capability to continuously adapt its knowledge representation to accommodate the frequent changes appear in the natural environment, i.e., continual lifelong learning. The next chapter of the thesis attempts to address these novel demands in computation to accommodate continual lifelong learning.

Chapter 5

A Continuous Lifelong Learning Approach for a New Digital World

This thesis introduced an overarching conceptual framework for an AI system, *Multi-layered Self-structuring Knowledge Representation Framework* (MSKRF) that brought together the neurophysiological inspiration, the features of the big data and digital environment to capture continuously evolving environmental stimulus and adapt its knowledge representation accordingly to achieve the overall objective of continuous lifelong learning. An in-depth exploration of existing work on knowledge representation in AI systems resulted in identifying Self-Structuring Artificial Intelligence (SSAI) with an unsupervised learning paradigm to be a highly viable algorithmic foundation for MSKRF. Chapter 4 proposed algorithmic modifications to the Growing Self-Organizing Map (GSOM) algorithm by combining the underlying concepts of self-structuring, transience, recurrent learning and multi-stream information fusion as computational constructs for core components of MSKRF. The current chapter aims to complete the materialization of the MSKRF by providing continual knowledge acquisition capability, which in turn will incorporate the seventh biological base, complementary learning systems theory.

In today's world, most industries adapt AI under different settings. For instance, transport industry has extensively adapted AI mechanism such as vehicle detection, anomaly detection, traffic propagation forecasting and anomaly detection to enhance the safety, security and convenience of travelers. In healthcare, patient scheduling in clinics to diagnosis of medical conditions based on medical imaging (X-rays, Computer tomography scans, MRI, ultrasound, etc.) to recommendation of medicine prescriptions have been automated using novel and efficient AI techniques. In fact, current AI solutions represented by state-of-the-art machine learning models are able to learn and even outperform human performance in individual tasks, as in Atari games (Oh *et al.*, 2015) and object recognition (Russakovsky *et al.*, 2015).

The typical sequence of these AI adaptations is to gather data, learn the underlying structure of the data, develop predictive/diagnostic models to conduct specific tasks and deploy the model to systematically perform these specific tasks. Gathering, preparing, and enriching the right data is essential and remains a key bottleneck among these many industries wanting to use AI. This phenomenon emphasizes the classical AI paradigm, known as Narrow Artificial Intelligence (ANI), which performs in isolation. As such, traditional AI systems have been perfected to wrangle with discrete and stationary data modalities, that has resulted in an asymmetric learning paradigm with historic data being used to solve current tasks. This asymmetric learning paradigm is also known as isolated learning because it does not consider any other related and background information or take in to account any current knowledge (Hong *et al.*, 2020). The fundamental problem with this isolated learning paradigm is that it does not retain and accumulate knowledge learned in the past and use it in future learning. This is in contrast to the human learning, where humans never learn in isolation but always retain the knowledge learned in the past and use it to guide future learning and problem solving. Thereby, whenever we encounter a new situation or problem, we may notice that many aspects of it are not really new because we have seen them in the past in some other contexts (Chen and Liu, 2016). Without the ability to accumulate knowledge, an AI system typically needs a large number of training examples in order to learn effectively. Labelling of training data is often done manually, which is highly labour-intensive and time-consuming. The world is highly complex with many possible tasks, making it almost impossible to label a large number of examples for every possible task for an AI algorithm to learn.

Adding further complications, the environment changes constantly, and any labelling thus needs to be done frequently and regularly to be useful, making it a daunting task for humans. In today's digital environment, where IoT is geared to generate non-stationary and high-frequent continuous data volumes, such practice becomes intractable. This would deem inefficient and unrealistic in most of the real-world scenarios, where the streaming data might disappear after a given period and/or not allowed to be stored at all due to storage or privacy constraints (Aljundi, 2019). This has generated the need for a symmetric learning paradigm where the AI system is not only geared to learn from past data to solve the current task but to harness the non-stationary and continuous data streams to acquire knowledge for past and future tasks.

In contrast, human knowledge acquisition process is quite different. Humans accumulate and maintain the knowledge learned from previous tasks and use it seamlessly in learning new tasks and solving new problems. Over time we learn more and more and become more and more knowledgeable, and more and more effective at learning. The key characteristic

of human learning (occurs in the brain) is the continuous learning and adaptation to new environments. Humans accumulate knowledge gained over a lifetime and use this knowledge to assist future learning and decision making with possible adaptations. Thus, humans can discover new tasks and learn while performing the tasks in open environments in a self-supervised manner. For instance, assume a scenario where a human drive a vehicle in an unknown narrow road in a residential area and hits a speed bumper. The next time on a similar road on a residential area the driver will most likely look out for speed bumpers.

Continuous Lifelong learning (CLL) aims to mimic this human learning process in AI. As such, AI systems are designed to learn from continuous streams of data adapting to the external environment and associated with different tasks with the goal of augmenting the acquired knowledge for problem solving and future learning. CLL is also known as lifelong machine learning, continual learning, continuous learning bearing a resemblance to lifelong learning of humans whose learning system is perfected in harvesting non-stationary and continuous data streams to acquire knowledge to perform past and future tasks. CLL accommodates to smoothly update AI systems to consider different tasks and data distributions while still being able to re-use and retain useful knowledge and skills learned previously. CLL is a learning paradigm that focuses on a higher and realistic timescale where data and tasks become available real-time and the access to previous data are limited. On this premise, this chapter intends to design, develop and evaluate two approaches for CLL addressing different challenges posed in learning continuously.

The subdivision of the chapter is presented in Fig. 5.1. The need for CLL is introduced and challenges to achieve it are presented in Section 5.1. The first section introduces the tasks associated with CLL that needs to be addressed in order to achieve CLL in computational models. These challenges are two-fold; the first relates to the evolving nature of input stimuli (data) while the second relates to the evolving nature of tasks computational models target to achieve. Section 5.2 intends to address the challenges associated with evolving nature of distribution of data, followed by the proposal of a new unsupervised deep learning based active learning approach (Section 5.2.1) and its experimental evaluation in the context of anomaly detection using surveillance video (Section 5.2.2).

The evolving nature of tasks in continuous learning is addressed in Section 5.3, followed by an in-depth review (Section 5.3.1) on existing complementary learning systems (CLS) based computational models, both supervised and unsupervised, leading to the design and development of a new self-organization based CLL approach (LifeNet) by incorporating constituents of CLS theory. LifeNet is designed using an architecture of RTGSOM adapted from the developments in Chapter 4 and incorporates the remaining biological base, *complementary learning system inspiration*. The LifeNet completes the materialization of the

proposed MSKRF, which will be evaluated using a series of benchmark datasets on object recognition and human activity recognition tasks.

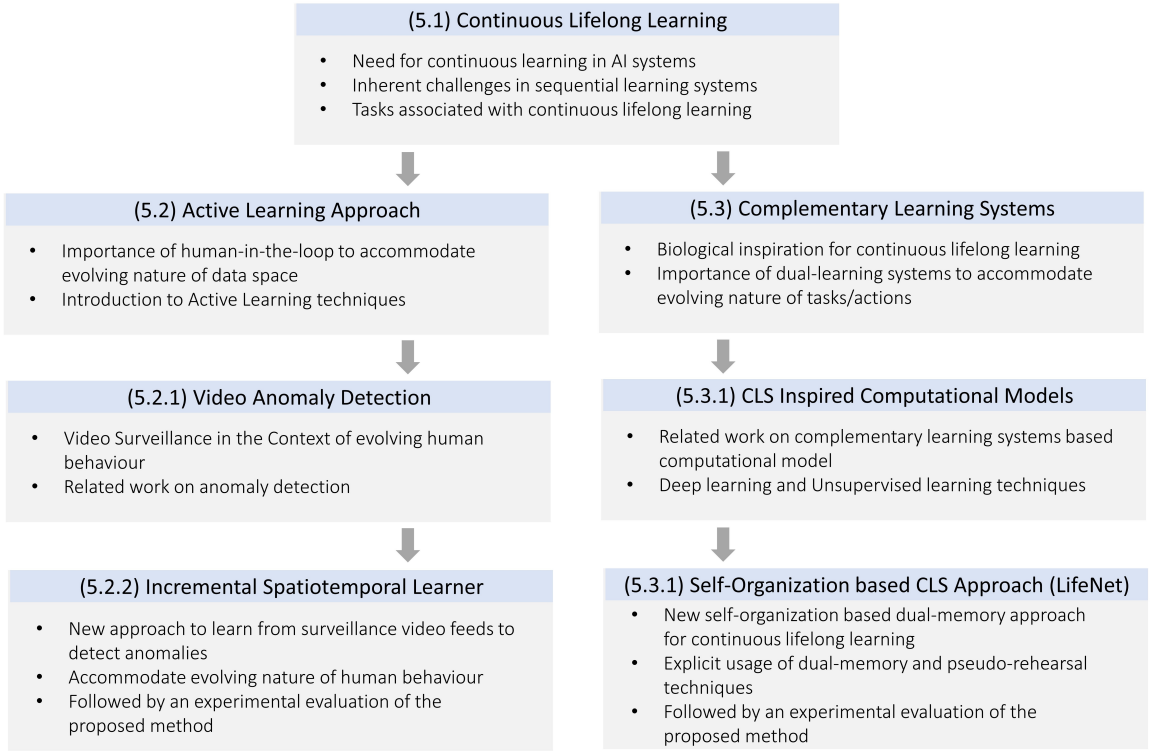


Fig. 5.1 Chapter Overview

5.1 Continuous Lifelong Learning

Continuous Lifelong Learning (CLL), studies the problem of learning from continuous streams of data adapting to the external environment and associated with different tasks with the goal of augmenting the acquired knowledge in problem solving and future learning (Aljundi, 2019). This section provides an extensive review of CLL and inherent challenges in achieving CLL in computational models relating to the module (5.1) *Continuous Lifelong Learning* as presented in chapter overview in Fig. 5.1.

In the context of AI systems, CLL accommodates to smoothly update the prediction model by taking new tasks and data distributions into account while still being able to reuse and retain useful knowledge and skills learned previously. CLL is a learning paradigm that focus on a higher and realistic time-scale where data and tasks become available during online and the access to previous data are limited. In general, CLL can be identified as a

sequential learning process where: (i) only a small portion of the input data, or (ii) a subset of the tasks is available at a time. The first case occurs when there is a larger distribution (variety) in data space, but at the training time, the learning model is only exposed to a subset of data. For instance, consider a scenario where an AI based system is implemented to diagnose diseases. For a disease there could by a variety of symptoms. However, at the initial phase, only a subset of symptoms is used to train the model because there are limited number of patients diagnosed with the said disease. Thus, new symptoms are needed to be accounted in the learning model when they are identified and known to the practitioners.

The second case occurs when all the possible outcomes are not known at the model development time. For instance, assume the medical research team is not aware of all the possible diseases that could occur in humans. As such, the AI model needs to be developed for the subset of diseases that are already known and along the way the model should be updated to accommodate new diseases without wiping out the knowledge to diagnose previously known diseases.

The main drawback of current connectionist models is that they are prone to catastrophic interference, i.e., new information used for train the computational model severely disrupts the existing knowledge. We discussed catastrophic interference related to representation learning in computational models relating to stability-plasticity dilemma in section 3.2.4. This phenomenon typically leads to an abrupt performance decrease or, in the worst case, to the old knowledge being completely overwritten by the new one. In efforts to overcome catastrophic interference, learning systems should be equipped to acquire new knowledge to update and augment existing knowledge based on continuous environmental stimuli, while preventing the novel input stimuli from significantly interfering with existing knowledge.

Early approaches of CLL consisted of memory systems that stored past data (used to train the model) and regularly replay these past data interleaved with samples drawn from new data (Robins, 1993). A major drawback of storing previous data throughout the lifetime of learning models is that they require explicit storage, leading to larger memory requirements. In addition, due to limited fixed number of neural resources in Connectionist models, special mechanism should be used to consolidate knowledge from being overwritten and maintain the same model performance level for different data distributions (Parisi *et al.*, 2019).

Allocation of additional neural resources for new knowledge in connectionist models have been attempted in recent work (Rusu *et al.*, 2016). For instance, additional neurons are added to a neural network architecture in subsequent learning steps when the model is exposed to new data. Generally, in a lifelong learning scenario, the number of tasks to be performed is not known at the beginning, constraining the learning model with a pre-defined amount of neural resources may compress the knowledge leading to degradation of performance.

However, this approach may lead to scalability issues when the neural architecture becomes extremely large requiring increased computational efforts. In contrast, humans are exposed to a dynamic world with a multitude of experience, and incrementally the human will acquire knowledge about the environment. At birth, the neuronal connections in biological brain starts at a relatively limited capacity and incrementally develop the capacity while the human ages (Shatz, 1992). Thereby, advancing the development of CLL computation mechanisms using the inspiration from biological brain has the potential to result in AI with the ability to identify and similarity and deviation of current occurrences with the past with better pattern recognition capability.

In essence, CLL systems should be developed for scenarios where: (i) only a small portion of the input data, or (ii) a subset of the tasks is available at once. Thereby, this thesis proposes two approaches to achieve CLL: (i) active learning approach to acquire knowledge on evolving nature of data distribution, and (ii) a self-organization based complementary learning system to learn incrementally updating tasks with associated data. The former presents a solution to address lifelong learning scenarios where only a small portion of the input data is available at once, and the latter addresses both the problem where either only a subset of the tasks is available or subset of data available at once.

5.1.1 Relation to other Machine Learning Paradigms

The concepts of multi-task learning, transfer learning and online learning are outlined in this section. These fields of machine learning have been developed in isolation and in connectivity, however, can draw clear differences with respect to CLL.

Multi Task Learning. Multi-task learning (MTL) is a concept of machine learning in which multiple learning tasks are solved simultaneously, while exploiting commonalities and differences across tasks (Zhang and Yang, 2017; Aljundi, 2019). This aims at better generalization and less overfitting using the shared knowledge across the related learning tasks. MTL can result in improved learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately. In contrast to CLL, MTL does not involve any continuous improvement or incremental knowledge adaptation post model deployment.

Transfer learning. Transfer learning (TL) is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned (Pan and Yang, 2009). In practical sense, TL is used as when a model is developed for a task and is reused as the starting point for a model on a second task. However, in comparison with CLL, no continual improvements to the deployed model are made post deployment.

Online Learning. Online learning (OL) is the paradigm in machine learning where the machine learning model is updated to best predict with data becomes available in a sequential order (Shalev-Shwartz *et al.*, 2011). This is as opposed to batch learning (or offline learning) which develops the model by using the entire data set. The OL approach is relatable to CLL, as both CLL and OL do incremental knowledge adaptation to the learned model in production. However, the learning in OL is just for a single task as opposed to CLL, where in CLL both data and tasks (classes to be predicted) can be updated.

5.2 Active Learning Approach for Continual Learning

This section intends to address the challenges associated with evolving nature of data as presented in the module (5.2) *Active Learning Approach* in chapter overview (Fig. 5.1). Fully automated deep learning has become the state-of-the-art (SOTA) for many tasks including data acquisition, analysis and interpretation. For instance, ImageNet (Russakovsky *et al.*, 2015) composites a large visual database designed for visual object recognition with more than 14 million hand-annotated images. ImageNet is organized according to the WordNet hierarchy (Miller, 1995), where each meaningful concept in WordNet is described by multiple words or word phrases, called a "synset". ImageNet aims to provide an average of 1000 images to illustrate each synset. Based on the ImageNet, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is annually conducted to foster the development and bench-marking of SOTA algorithms based on a subset of the images in ImageNet. Fully automated deep learning algorithms have been the front-runner of this competition, and as of March 2020, FixEfficientNet developed by Facebook AI research team (Touvron *et al.*, 2020) has become the current front-runner with an accuracy of 88.5%. However, it must be noted that this competition, similar to many automated deep learning based algorithms focus on data (or datasets) that are static, and their distribution do not evolve over time. However, the reality is different as it drastically evolves over time making current knowledge incomplete or obsolete for future decision making (Nawaratne *et al.*, 2019c). In this light, we investigate the role that humans might need to play in the development and deployment of learning models to accommodate this continuously evolving nature providing continual lifelong learning for the learning models.

The branch of AI that leverages machine intelligence with human-in-the-loop to develop CLL models is known as *Active Learning*. In traditional active learning methods, human involvement was limited to selection of the best descriptive data to annotate for optimal model performance, irrespective of the time-dependent evolution of physical world that inevitably influence the data to evolve. Therefore, an increased interest is directed to use active learning

for model interpretation and refinement using human-in-the-loop in production of industrial AI systems. That is to use iterative feedback to steer learning models to optimal prediction capacity, offering meaningful ways to interpret and respond to predictions (Budd *et al.*, 2019).

Active learning with human-in-the-loop is an area that we see as increasingly important in future research due to the evolving nature of data and the safety-critical nature of working in diverse domains such as healthcare, safety and security. In this section, we propose an active learning approach to acquire knowledge on evolving data distribution. This study intends to present the active learning approach from a context of real-time video surveillance. The proposed active learning approach consists of a deep learning model that learns spatiotemporal patterns from surveillance video streams. The learning is modulated with an active learning approach enabling continuous learning.

On this basis, the next section introduces to video surveillance in the context evolving behaviour emphasizing the current limitations in AI systems. Following sections will present a background on existing methods to overcome these limitations and present the novel approach for continuous active learning. We demonstrate and validate the proposed approach using three benchmark datasets.

5.2.1 Video Surveillance in the Context of Evolving Human Behaviour

Video surveillance is a predominant consideration in the development, operation and sustainability of modern industrial and urban environments. Video surveillance contributes towards efficiency, safety, security and optimality of the locality, infrastructure, individuals, operations and activities. Industrial environments are transitioning towards autonomous machinery, cyber-physical systems and energy efficient layouts. Urban environments are becoming densely populated, with high usage of multi-level buildings, increased vehicular, pedestrian and crowd movements. This vertical and horizontal expansion of asset and area utilisation in both industrial and urban environments have eventuated an exponential increase in the deployment of Closed-circuit television (CCTV) camera systems. However, it is unrealistic and infeasible for human observers to monitor and analyse every video stream with high precision. AI techniques for autonomous video surveillance reported in current literature can be categorised into video summarisation (Kosmopoulos *et al.*, 2012), object detection and re-identification (García *et al.*, 2014), activity/behaviour detection (Sargano *et al.*, 2017), and anomaly detection (Kiran *et al.*, 2018).

Anomaly detection is a constitutive task in autonomous video surveillance as it contributes to the success of the other categories noted above. It is also a complex task as the anomalies to be detected are not known prior, imposing difficulties even for a human observer.

A general definition for anomaly detection is the identification of behaviours that do not conform to expected and accepted behaviour (i.e., normal behaviour) (Chandola *et al.*, 2009). In the context of autonomous video surveillance, anomaly detection is impacted by three primary challenges. First, the computational complexity and cost of video data processing due to spatial and temporal dimensional structure combined with non-local temporal variations across video frames (Kiran *et al.*, 2018). As an example, anomalous objects such as vehicles/bicycles in a pedestrian walk must be identified using spatial processing whereas anomalous behaviour such as jaywalking must be determined using temporal variations across video frames. Second, the anomaly itself is ill-defined, the boundary between normal behaviour and anomalies is often imprecise, and anomalies are highly contextual (Chandola *et al.*, 2009). For example, industrial machinery operating at low power can be either normal or anomalous depending on the operational circumstances. Third, what is considered as normal behaviour evolves over time making current knowledge incomplete and/or obsolete (Najafabadi *et al.*, 2015). For instance, when offenders become aware of detected anomalies, they can maliciously adapt behaviours so that subsequent anomalies are difficult to detect.

Existing literature attempts to address the first and second challenges (i.e., computational complexity and identifying contextual anomalies). The third challenge, the evolving nature of normal behaviour over time, remains unaddressed, and this makes current knowledge of normality incomplete. In this study, we propose the Incremental Spatio-Temporal Learner (ISTL), to address the aforementioned challenges and limitations. ISTL is a new anomaly detection approach for real-time video surveillance that actively learns spatiotemporal patterns of normal behaviour as it evolves over time. ISTL is inspired by continuous learning process in human cognition and the paradigm of active learning. Inspired by the human brain, ISTL begins by developing a basic understanding from immediately available information to distinguish between normal (safe) and anomalous (unsafe) behaviours, and continuously refines this understanding as the surroundings change and new information becomes available (De Silva and Alahakoon, 2010). Active learning is primarily used for refinement and validation in ISTL, where a human observer contributes to the learning process for improved learning outcomes across iterations. The paradigm of active learning has been widely used in industrial image and video analysis applications such as character reading, facial recognition, autonomous vehicles and e-commerce (Liu, 2018; Blog, 2017).

Related work on Anomaly Detection

Techniques and approaches for intelligent video surveillance in current literature broadly range across two areas of research, hand-crafted video features and learned-representations

based on deep learning architectures (Nawaratne *et al.*, 2019c). In techniques that utilize handcrafted features, trajectories and spatiotemporal changes are extracted as input/output features for computational and AI modelling. For instance, Xie and Guan (2016) proposed a motion instability based anomaly detection framework that discriminates anomalous behaviour based on the direction randomness and motion intensity, whereas Wu *et al.* (2010) proposed an approach in which objects are classified as anomalous based on how they follow the learned normal trajectory. These trajectory-based methods define normal behaviour based on previously observed motion patterns. However, such trajectory-based methods fail to detect anomalous behaviour based on the appearance of entities in the surveillance video stream and computationally expensive for crowded scenes. State-of-the-art handcrafted feature extraction methods describe video events ranging from pixel-level to 3-dimensional cuboid. For instance, Zhao *et al.* (2011) utilize histogram of gradient (HoG) and histogram of optical flow (HooF) along spatial and temporal dimensions to encode an event and learn the normality upon dynamic sparse coding, whereas, Zaharescu and Wildes (2010) models the normal behaviour based on distributions of spatiotemporal oriented energy. These handcrafted feature based techniques can accurately model both spatial and temporal dynamics, however, they require prior knowledge for the design of effective features, and are time consuming to extract, thereby impractical to use in real-time anomaly detection.

With the advancements of deep learning, recent work has utilized convolution neural networks (CNN), autoencoders and recurrent neural networks (RNN) for video anomaly detection (Nawaratne *et al.*, 2019c). Xu *et al.* (2017) proposed Appearance and Motion DeepNet (AMDN) that utilizes an autoencoder to automatically learn feature representation from the surveillance video, use a double fusion framework and support vector machine (SVM) models to predict the irregularity of an event. The AMDN model results in the state-of-the-art accuracy, however, its processing time is in the order of 10,000 milliseconds, which makes it impractical to use in online anomaly detection. Hasan *et al.* (2016) approached the problem of anomaly detection by learning a generative model for regular motion patterns. The approach achieved positive results using a 10-layered fully convolutional feed-forward autoencoder to reconstruct input video, then detect anomalies based on its reconstruction cost analysis. Luo *et al.* (2017) attempted to detect anomalies by leveraging a CNN for appearance encoding and a convolutional long-short-term memory (ConvLSTM) for remembering history of the motion information. Recently, Vu (2017) proposed an anomaly detection approach using a deep generative network in which normality is modelled by an unsupervised probabilistic framework. With these advancements, it is evident that learned-representations based on deep learning architectures can distinguish anomalies from normal behaviours by processing high-dimensional surveillance video streams. However, existing deep learning approaches

for anomaly detection are highly dependent on a known dataset validated as normal, for training and constrained by sparse evaluation based only on reconstruction error, without consideration for surveillance context.

In summary, current literature is mostly limited to addressing the computational complexity of processing high dimensional video data and identifying contextual anomalies from surveillance video streams. To the best of our knowledge, the challenge of capturing evolving nature of events over time, remains unaddressed which makes current knowledge of normality incomplete.

5.2.2 Incremental Spatio-Temporal Learner Approach

A high-level overview of ISTL is illustrated in Fig. 5.2. First, the live video surveillance feed is presented as input to spatiotemporal model training of normal behaviour (from time t_0 to t_u). Second, the trained model is utilized for anomaly detection and localization within the time interval t_u to t_v. Third, the detected anomalies are validated by human observer and the validation input is used to construct updated normal behaviour using fuzzy aggregation. This updated normal behaviour is fed back into the ISTL learning model for continuously learning. This overview is expanded into a functional view and illustrated in Fig. 5.3.

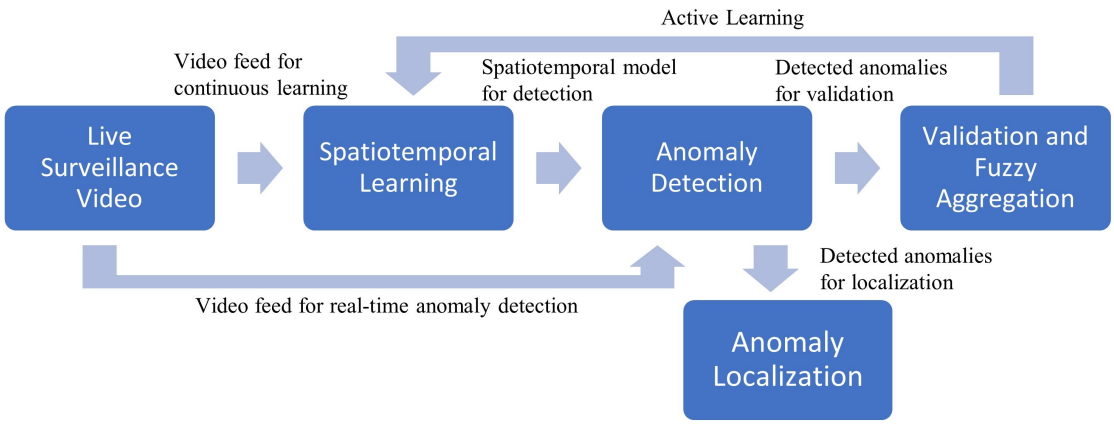


Fig. 5.2 Overview of the proposed ISTL approach

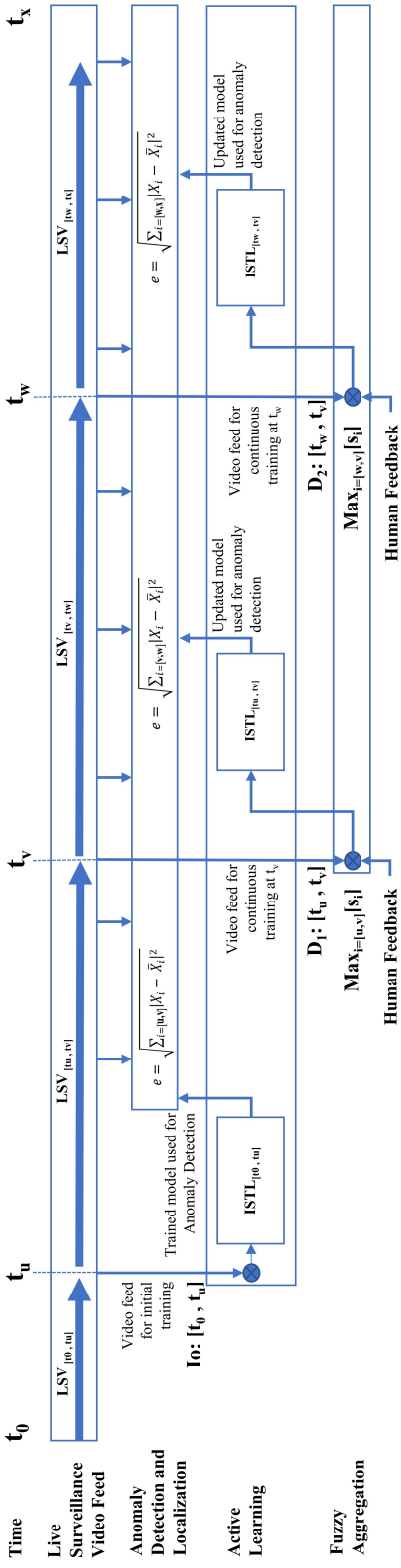


Fig. 5.3 Overview of the proposed ISTL approach

The computational formulation of anomaly detection in video surveillance is presented as follows. The training video stream (X_{train}) composed of a sequence of frames of height h and width w, $X_{train} \subset IR$, that only contains video frames exhibiting normal behaviour in a given camera view. R indicates all the video frames of the camera view in real world. In the testing phase, a video stream (X_{test}) is employed, where $X_{test} \subset IR$ contains video frames of both normal and anomalous behaviour. The goal is to learn a representation (Ω) of normal behaviour from X_{train} which is subsequently validated with X_{test} to distinguish anomalies. In contrast to previous work (Hasan *et al.*, 2016; Xu *et al.*, 2017) that require a complete training dataset of normal behaviour, the ISTL approach will actively update previously learned knowledge (Ω) based on (i) spatiotemporal information from continuously received video streams, and (ii) active human observer feedback on detected anomalies.

The three phases of ISTL, 1) Spatiotemporal Learning, 2) Anomaly Detection and Localization, 3) Active Learning with Fuzzy Aggregation, are explicated in following subsections.

Spatiotemporal Learning

Spatiotemporal representation of normal behaviour is learned from X_{train} as expected and acceptable behaviour for the video surveillance application. The ISTL model is composed of a spatiotemporal autoencoder to learn the appearance and motion representation from video inputs. The autoencoder is an unsupervised learning algorithm that employs backpropagation to set the target values to be equal to the inputs by minimizing the reconstruction error (Baldi, 2012). In the proposed architecture, the spatiotemporal autoencoder consists of a series of CNN layers to learn the spatial representation and a series of ConvLSTM layers to learn the temporal representation. The input data layer and feature transformation layers of the autoencoder are described in following sections.

Input Data Layer

The raw video data are pre-processed to enhance the learning capacity of the spatiotemporal autoencoder model. At first, the video data are extracted as consecutive frames, convert into grayscale to reduce the dimensions, resize to 224×224 pixels and normalize pixel values by scaling between 0 and 1. The input to the spatiotemporal autoencoder model is a temporal cuboid of video frames, which will be extracted using a sliding window of length T without any feature transformation. The consecutive frames of length T are stacked together to construct the input temporal cuboid. Increased length of this temporal window (T) will enable to incorporate motion of longer length, however, the larger the T, the model convergence will take exponential time (Hasan *et al.*, 2016).

Table 5.1 Spatiotemporal Autoencoder Architecture

ID	Input Tensor	Operation	Output Tensor
C1	T x 224 x 224 x 1	CV; F: 128; K: 27 x 27; S: 4	T x 56 x 56 x 128
C2	T x 56 x 56 x 128	CV; F: 64; K: 13 x 13; S: 2	T x 28 x 28 x 64
CL1	T x 28 x 28 x 64	CL; F: 64; K: 3 x 3	T x 28 x 28 x 64
CL2	T x 28 x 28 x 64	CL; F: 32; K: 3 x 3	T x 28 x 28 x 32
DCL1	T x 28 x 28 x 32	CL; F: 64; K: 3 x 3	T x 28 x 28 x 64
DC1	T x 28 x 28 x 64	DCV; F: 64; K: 13 x 13; S: 2	T x 56 x 56 x 128
DC2	T x 56 x 56 x 128	DCV; F: 128; K: 27 x 27; S: 4	T x 224 x 224 x 1

Convolution Layers (CNN)

CNNs have been inspired from biological processes resembling the organization of the animal visual cortex (Matsugu *et al.*, 2003). The connectivity of the neurons in the convolution layers are designed in a manner similar to animal vision system such that an individual cortical neuron responds to stimuli only in a confined region of the input frame, i.e., the receptive field. In video analysis, the convolution layers can preserve the spatial relationship within the input frames by learning feature representations using filters, whose values are learned during the training process. The ISTL model consists of two convolution layers and two de-convolution layers, whose filters and kernel sizes are specified in the Table 5.1, where [CV] refers to Convolution, [CL] refers to Convolution LSTM, [DCV] refers to Deconvolution, [T] refers to the depth of temporal cuboid, [F] refers to the number of filters, [K] refers to the kernel size and [S] for strides.

Convolutional LSTM Layers (ConvLSTM)

Recurrent neural network (RNN) captures the dynamic temporal behaviour of a time-sequence input data by employing an internal memory to process the input sequences. Long short-term memory (LSTM) units are an advancement of generic building blocks of the RNN. The LSTM unit is composed of an input gate, an output gate, a forget gate and a cell. The input gate defines the extent the input value moves into the unit. The forget gate controls the extent the values from the previous time steps remain in the unit and the output gate controls to which extent the current input value is used for the computation for the activation of the unit. The cell remembers values over arbitrary time intervals.

As LSTM is primarily developed and utilized for modeling long-range temporal correlations, it has a drawback in handling spatial data as spatial information is not encoded in its state transition. However, it is essential to learn the temporal regularity from the surveillance video stream while preserving the spatial structure, particularly for anomaly detection. Therefore, we utilize an extension to LSTM, convolutional LSTM (ConvLSTM), in which

both the input-to-state and state-to-state transitions have convolution structures (Xingjian *et al.*, 2015). The ConvLSTM overcome this drawback by designing its inputs, hidden states, gates and cell outputs as 3D tensors, whose last dimension is the spatial dimension. Further, the matrix operations in its inputs and gates are replaced with convolution operator. With these modifications, the ConvLSTM is able to capture the spatiotemporal features from the input frame sequences. The ConvLSTM model is represented in the equations 5.1-5.5.

$$i_t = \alpha(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \tag{5.1}$$

$$f_t = \alpha(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \tag{5.2}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{5.3}$$

$$o_t = \alpha(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_{t-1} + b_o) \tag{5.4}$$

$$H_t = o_t \circ \tanh(C_t) \tag{5.5}$$

In the equations, $*$ and $\circ$ represents convolution operation and the Hadamard product respectively. Inputs are represented by $X_i, \ldots, X_t$, the cell states are represented by $C_i, \ldots, C_t$, the hidden states are represented by $H_i, \ldots, H_t$, and the gates it, f_t and o_t are all 3D tensors. α is the sigmoid function and, $W_{x^{\sim}}$ and $W_{h^{\sim}}$ are 2D convolution kernels in the ConvLSTM. The ISTL model consists of three ConvLSTM layers. The spatiotemporal autoencoder architecture is illustrated in Fig. 5.4 and its composition further elaborated in Table 5.1.

Anomaly Detection and Localization

The ISTL model can be used to obtain a reconstruction of the normality of the input video at pixel-level precision. However, the trained autoencoder does not have the ability to accurately reconstruct the anomalous or unseen scenes, due to the fact that, such scenes have not been presented in the training phase. This phenomenon is used to evaluate and detect anomalies from the input video. We obtain the reconstruction error (E) as the square root of the sum of the squared vector values, as represented in equation 5.6 and 5.7, where X is the input temporal cuboid, $\overline{X}$ is the reconstructed temporal cuboid, T is the time window, w is the width and h is the height of the video frame.

$$\varphi(i,j,k) = |X_{(i,j,k)} - \overline{X}_{(i,j,k)}|^2 \tag{5.6}$$

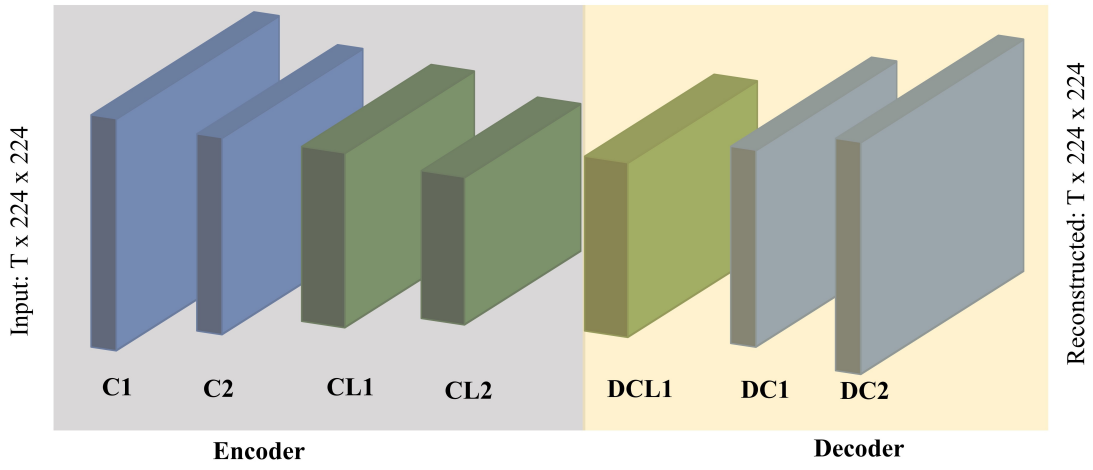


Fig. 5.4 Spatiotemporal Autoencoder Architecture. Layer IDs are referred from Table 5.1

$$E = \left(\sum_{i=1}^{T} \sum_{j=1}^{w} \sum_{k=1}^{h} \varphi(i,j,k) \right)^{\frac{1}{2}} \tag{5.7}$$

The reconstruction error represents the score for each temporal cuboid defining the anomaly. We define a reconstruction error threshold to distinguish between normal behaviour and anomalies, named anomaly threshold (μ). In practical video surveillance applications, the human observer can select a value for μ based on the sensitivity required for the surveillance application. A low μ would result in higher sensitivity to the surveillance arena, resulting in higher number of alerts. A high μ would result in lesser sensitivity that could lead to miss sensitive anomalies in the surveillance arena.

Additionally, we introduce the temporal threshold (λ), which we define as the number of video frames that should be higher than the μ to recognize an event as an anomaly. λ is employed to reduce the false-positive anomaly alerts due to sudden variations of the surveillance video stream due to occlusion, motion blur and high-intense lighting conditions. Fig. 5.5 illustrates anomaly detection approach based on the reconstruction error.

Anomaly localization locates the specific area of the video frame, where an anomaly has occurred. Subsequent to detecting a segment of the video as anomalous, we localize the anomalies by calculating reconstruction error (E_c) over non-overlapping spatiotemporal local cuboid windows, where m and n are the width and height respectively, and T is the depth (i.e., number of frames in the cuboid). Equation 5.7 is used to calculate the E_c for local cuboids.

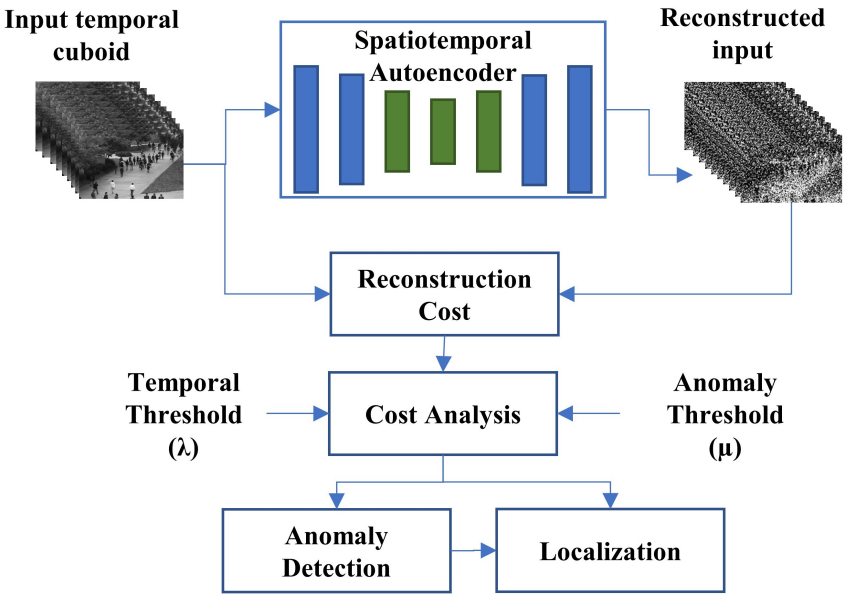


Fig. 5.5 Anomaly detection and localization

Active Learning with Fuzzy Aggregation

The purpose of the active learning in practical video surveillance context is to enable anomaly detection of dynamically evolving environments. By automating the anomaly detection using the deep learning model, we train the learning model to identify accepted normal behaviour provided at the beginning. However, in dynamic environments comprising of new normal behaviour that have not anticipated and/or existing behaviour that considered abnormal reformed to normal, it is important that the detection system evolves with capabilities for detecting such new scenarios. ISTL addresses this challenge by adopting an active learning approach using fuzzy aggregation to continuously train the learning model with unknown/new normal behaviour specific to the corresponding surveillance context. This approach is inspired by the human brain's ability to develop a basic understanding which is continuously refined as new information becomes available.

ISTL is initially trained with a pre-identified normal behaviour in the surveillance context and used for anomaly detection. If a video frame is detected as an anomaly, i.e., the reconstruction error of the input cuboid is above the anomaly threshold, the input cuboid is classified as an anomaly. The classified frames are then sent to a human observer for verification. The objective of human observer feedback is to actively feed the learning model with dynamically evolving normality behaviour. Therefore, if a detected video frame is an

incorrect detection (false positive), then the human observer can mark the video frame as 'normal', which will be used in the continuous learning phase.

After human observer feedback, the video frames that were marked as normal will be used to continuously train the ISTL model, updating its knowledge of the notion of normality. As shown in Fig. 5.6, continuous update of the ISTL model is conducted using (i) spatiotemporal information from the continuously received surveillance video stream, and (ii) active human observer feedback on detected anomalies.

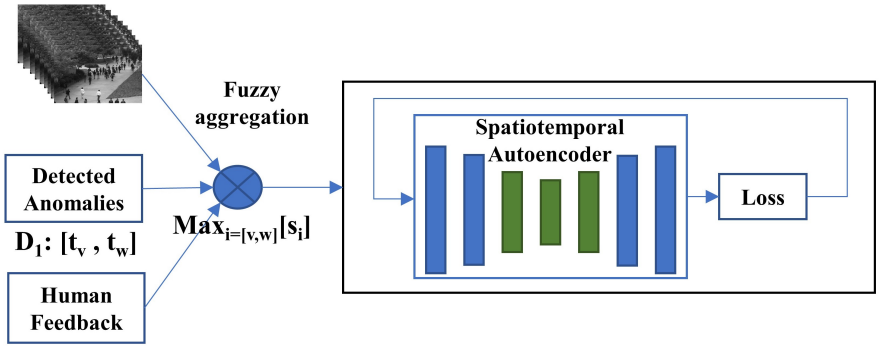


Fig. 5.6 Anomaly detection and localization

The continuous learning of the ISTL model is enriched by fuzzy aggregation of video frames, in order to retain stability across iterations of learning. At the detection phase, all the video frames being evaluated are tagged with a fuzzy measure g_λ based on its reconstruction error and grouped into finite number (n) of sets based on g_λ. Subsequently, in the continuous learning phase, the algorithm will select the k video frame cuboids that contain highest g_λ from each set of fuzzy measures (S) to train the ISTL model. The parameters k and n are defined at initiation based on the duration of video surveillance stream employed for continuous learning. The scene selection for continuous training is defined by the equation 5.8; $\forall s \in S$, where, $S = s_1, s_2, \ldots s_n$ and d is the indexes of the selected temporal cuboids that will be included in the continuous training dataset.

$$d = \sum_{i=1}^{n} \max_{j=[1,k]}(S_i) \tag{5.8}$$

The dataset for continuous training iteration is now composed of (i) false positive detection verified by the human observer, and (ii) temporal cuboids selected across normal behaviour using the fuzzy aggregation. This will ensure the continuous training will update the detection model's capability to capture novel normal behaviour while remaining stable for previously known normal behaviour. This fuzzy aggregation approach has been suc-

cessfully demonstrated to maintain stability-plasticity in continuous learning for IoT stream mining (De Silva *et al.*, 2011), text mining (De Silva and Alahakoon, 2010) and video stream mining (Nawaratne *et al.*, 2017).

After the scene selection, the ISTL model will be continuously trained upon the selected representation from input video data, which is the updated expected and acceptable behaviour form the surveillance arena. Thenceforth, the updated ISTL model will be re-employed for anomaly detection.

5.2.3 Evaluation of ISTL Approach

The proposed approach, ISTL, is evaluated using three benchmark datasets, CUHK Avenue dataset (Lu *et al.*, 2013b), UCSD Ped 1 and UCSD Ped 2 datasets (Mahadevan *et al.*, 2010b). With this empirical evaluation, we demonstrate the capability of the ISTL to detect and localize anomalies in near real-time and that the ISTL model performs on par with state-of-the-art anomaly detection methods proposed in the current literature. ISTL was implemented in Python with TensorFlow framework (Abadi *et al.*, 2016) for enhanced capabilities in deep learning and GPU utilization. ISTL was trained on a high-performance computing specification, 36-core CPU 2.3GHz with 128 GB memory and dual NVIDIA Quadro of 24 GB GPU units. Evaluation of ISTL was conducted on a typical personal computer configuration, a 4-core CPU 2.6 GHz with 24GB memory and GPU of NVIDIA GeForce GTX 970M, in order to ensure that the proposed ISTL model can be realistically deployed in an industrial setting.

Datasets

The CUHK Avenue dataset (Lu *et al.*, 2013b) was acquired using a stationary video camera with a resolution of 640×360 pixels, recording street activity at City University, Hong Kong. This dataset has 16 train video samples that contain normal human behaviour and 21 test video samples that contain unusual events and human actions. The normal behaviour are pedestrians on the sidewalk and groups of pedestrians congregating on the sidewalk, whereas the anomalous events are people littering/discarding items, loitering, walking towards the camera, walking on the grass and abandoned objects.

The UCSD pedestrian Dataset (Mahadevan *et al.*, 2010b) was captured by a stationary video camera with a resolution of 238×158 pixels, focusing on two pedestrian walkways. This includes two datasets, ped 1 and ped 2, capturing different crowd scenes, ranging from sparse to dense. The normal behaviours of the train video samples contain only scenarios of pedestrians walking on the pathway, whereas the test video samples contain anomalous

pedestrian movement patterns such as walking across the sidewalk or walking on the grass, unexpected behaviour such as skateboarding, cycling, and vehicular movement. Ped 1 dataset has 34 train video samples and 36 test video samples, whereas Ped 2 dataset has 16 train video samples and 12 test video samples. Both the selected datasets were captured at a frame rate of 26 frames per second (FPS).

Experimental Setup

The experimental setup is fourfold:

1. First, the anomaly detection capabilities of the spatiotemporal autoencoder model is evaluated and compared with the state-of-the-art anomaly detection models based on the three benchmark datasets.

2. Second, the anomaly localization capability is evaluated using non-overlapping cuboids of $16 \times 16 \times T$ pixels. This size is selected for the input cuboid as it is small enough to capture the location of anomalies as well large enough to extract related appearance information, based on the video resolution of selected datasets.

3. Third, we evaluate the continuous learning capability of the ISTL model for UCSD Ped1 and Ped2 datasets, adapting a scenario as normal which was previously considered as an anomaly.

4. Fourth, we conduct a runtime analysis of our approach demonstrating the real-time processing capabilities of our algorithm.

As the video samples have different dimensionality, we pre-process the inputs by resizing the extracted frames to 224×224 pixels, and normalizing pixel values by scaling between 0 and 1. Based on the frame rate of the selected training data (26 FPS), we select the depth of temporal cuboid, $T = 8$ representing an approximate duration of one-third of a second. The selection of T is both dependent on maximising the motion to be captured within consecutive frames as well minimizing the convergence of the deep learning model due to large depth of input cuboids. In scenarios where the input surveillance data has lower frame rate, it is possible to capture longer motion with low temporal depths.

In this experiment, we trained the learning model using a learning rate of 0.01 and 1500 training epochs. Stochastic gradient descent algorithm is used to optimize the spatiotemporal autoencoder model and mean squared error is used as the cost function to calculate the reconstruction loss. In order to avoid overfitting of the model, we employed early stopping regularization technique where the training terminates when the loss has stopped improving.

Table 5.2 Selection of Anomaly Threshold and Temporal Threshold

Dataset	Optimal AUC/EER	Anomaly Threshold μ	Temporal Threshold λ
Ped 1	75.2/29.8	0.33	1
Ped 2	91.1/8.9	0.38	9
Avenue	76.8/29.2	0.29	6

The training was conducted for 3 continuous iterations by splitting the data set as 60% for the first iteration, and 20% each for second and third iterations (as elucidated in Fig. 5.3). The reconstruction error was used as the fuzzy measure in the active learning phase.

In the anomaly detection and localization phase, the two thresholds are the temporal threshold and the anomaly threshold. We evaluated a range of λ from 1 to 9 in order to select the optimal value for each test dataset. The evaluation is presented in Fig. 5.7. The optimal λ was different for the three datasets; minimum value of 1 (one-third of a second) for Ped 1 dataset whereas the maximum value of 9 (three seconds) for the Ped 2 dataset. This can be justified by the view point of the video samples as Ped 1 dataset has the farthest view making the pedestrian/object movement to be small, whereas the Ped 2 dataset contained a closer view of the pedestrians/objects which made video sample to capture movements a lengthier than the Ped 1 dataset. Avenue dataset captured the optimal anomalies with the λ of 6 (two seconds), which similarly can be justified by the camera view and the movement of people captured in the video sample. The optimal accuracy the ISTL model was able to achieve is presented in Table 5.2, with respective λ and μ values.

Results - Anomaly Detection

Anomaly detection was evaluated with three state-of-the-art handcrafted feature representation-based approaches and four state-of-the-art deep learning-based approaches. The selected handcrafted feature representation-based methods include, first, abnormal crowd behaviour detection using social force model (SF) by Mehran *et al.* (2009) which employs a grid of particles is placed over the video frame and the space-time average of optical flow to enforce the social force model. Second, we evaluate MPCCA model (Kim and Grauman, 2009) that utilizes space-time Markov random field and video optical flow for anomaly detection. Third, we evaluate MPCCA+SF model (Mahadevan *et al.*, 2010a), the original work of the UCSD ped 1 and ped 2 datasets. The anomaly detection of this approach is based on mixtures of dynamic textures, where the outliers under this model are labelled as anomalies.

The selected deep learning-based approaches for comparison are as follows. First, Conv-AE (Hasan *et al.*, 2016) is a deep convolution feed-forward autoencoder architecture that learns both local features and classifiers as an end-to-end learning framework. Second,

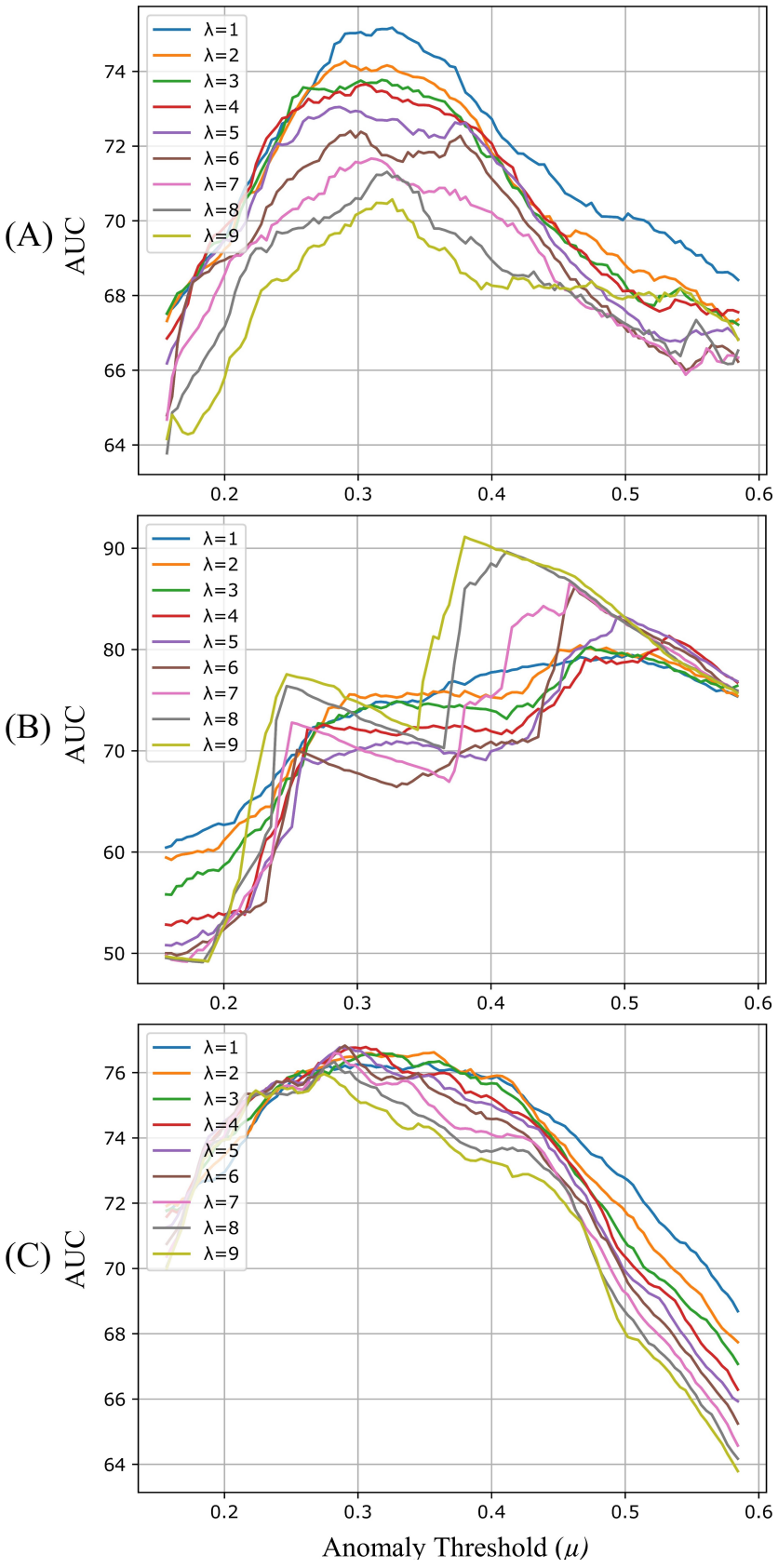


Fig. 5.7 Evaluation of optimal AUC with respect to μ based on different λ values. (A) UCSD Ped 1, (B) UCSD Ped 2, and (C) CUHK Avenue.

Table 5.3 Comparison of AUC / EER

Model	Ped 1	Ped 2	Avenue
SF (2009)	67.5/31.0	55.6/42.0	NA
MPCCA (2009)	66.8/40.0	69.3/30.0	NA
MPCCA + SF (2010)	74.2/32.0	61.3/36.0	NA
Conv-AE (2016)	**81.0/27.9**	90.0/21.7	70.2/25.1
S-RBM (2017)	70.3/35.4	86.4/16.5	78.8/27.2
ConvLSTM-AE (2017)	75.5/NA	88.1/NA	77.0/NA
Unmasking (2017)	68.4/NA	82.2/NA	**80.6/NA**
Ours (ISTL)	75.2/29.8	**91.1/8.9**	76.8/29.2

S-RBM (Vu, 2017) is an unsupervised probabilistic framework that models the normality and learn feature representations automatically. Third, ConvLSTM-AE (Luo *et al.*, 2017) is an integrated CNN and ConvLSTM autoencoder to encode spatial and temporal patterns in normal behaviour. Fourth, Unmasking-late-fusion (Tudor Ionescu *et al.*, 2017) is an anomaly detection approach based on unmasking technique. This method employs motion features captured from 3D gradients and appearance features from pre-trained CNN, specifically VGGNet (Chatfield *et al.*, 2014).

We compare the results of respective models using frame level ROC curves, the corresponding Area Under the Curve (AUC) and Equal Error Rate (EER). The comparison is presented in Table 5.3, where the results appear as reported by respective authors. Overall, our method outperforms all the handcrafted approaches whereas we obtain on-par results in comparison to deep learned representation-based methods with respect to Ped 1 and Avenue datasets. For the Ped 2 dataset, our proposed ISTL method outperforms all the compared models including the benchmark of Conv-AE (Hasan *et al.*, 2016) approach.

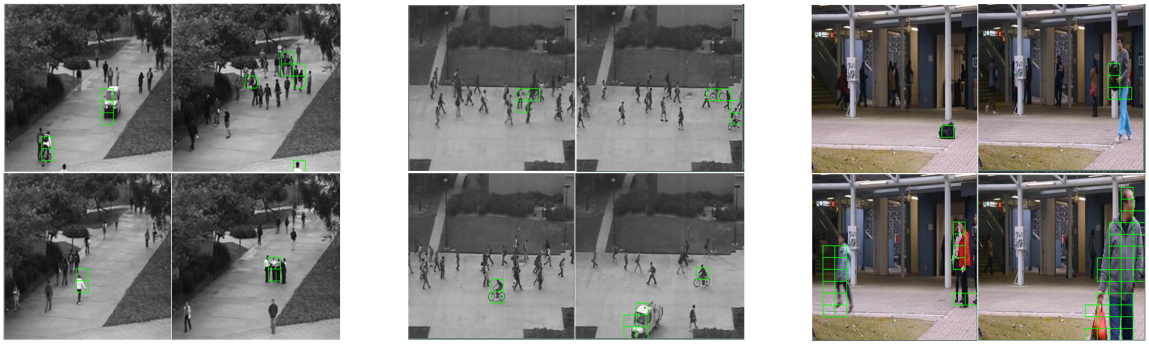
Results - Anomaly Localization

Qualitative analysis of the localized anomaly patches is presented in Fig. 5.8. It is shown that anomalies such as cyclists and vehicles on the pathways, pedestrians walking across the pathways, crowd loitering and pedestrians pushing carts are localized by ISTL in the UCSD ped 1 dataset. It is important to note that there were false negative detections with respect to skateboarding in ped 1 dataset (Fig. 5.8A). Out of the 12 test videos samples that contained people who skateboard, only 10 were detected by the ISTL model. However, in the ped 2 dataset, all the video samples that contained skateboarding were detected. This can be explained by the camera angle of ped 1 datasets where its elevation makes it difficult to differentiate between pedestrians and skateboarders by appearance.

Table 5.4 Anomaly Detection for Cycling Scenario

Dataset	Prior to active learning	After active learning
UCSD Ped 1	12 / 14	2 / 14
UCSD Ped 2	7 / 7	1 / 7

In UCSD ped 2 test samples, bicycles, vehicles and pedestrians walking in different directions are localized. The main anomaly in the ped 2 test samples was cyclists, in 11 out of the 12 instances. Anomalies such as an abandoned bag, a person throwing a bag, child playing in the surveillance area, people walking in wrong directions, people running are localized as anomalies by ISTL in the CUHK Avenue dataset.



(A) (B) (C)

Fig. 5.8 Localised anomalies. (A) UCSD Ped 1 Dataset, (B) UCSD Ped 2 Dataset, and (C) CUHK Avenue Dataset.

Results - Active Learning

In order to demonstrate the active learning capability of ISTL, we selected cycling on the pedestrian pathway scenarios of UCSD Ped 1 and Ped 2 datasets. Here we defined cycling on pedestrian pathways as a normal behaviour, thereby tagged all the anomaly detections from test samples of cyclists as normal. We employed 4 test samples containing cyclists from each Ped 1 and Ped 2 datasets to continuously train the ISTL model with human observer verification. Subsequent to the training phase, we evaluated the anomalies of the test samples excluding the 4 samples selected for continuous training. The anomaly detection ratio is presented in Table 5.4. In Ped 1 dataset evaluation, it was detected that 2 test samples that had cyclists were anomalous, because these were across sidewalk cycle movements.

To further evaluate utility of the active learning approach, we singled out two particular test scenarios that have been previously detected as anomalous; (A) cyclist only, and (B)

cyclist and a vehicle moving on the pedestrian pathway (as illustrated in Fig. 5.9). The evaluation resulted in test video A being detected as normal while test video B being detected as an anomaly. This localization confirms that the video B was detected as an anomaly due to the moving vehicle, whereas the cyclist was detected as normal.

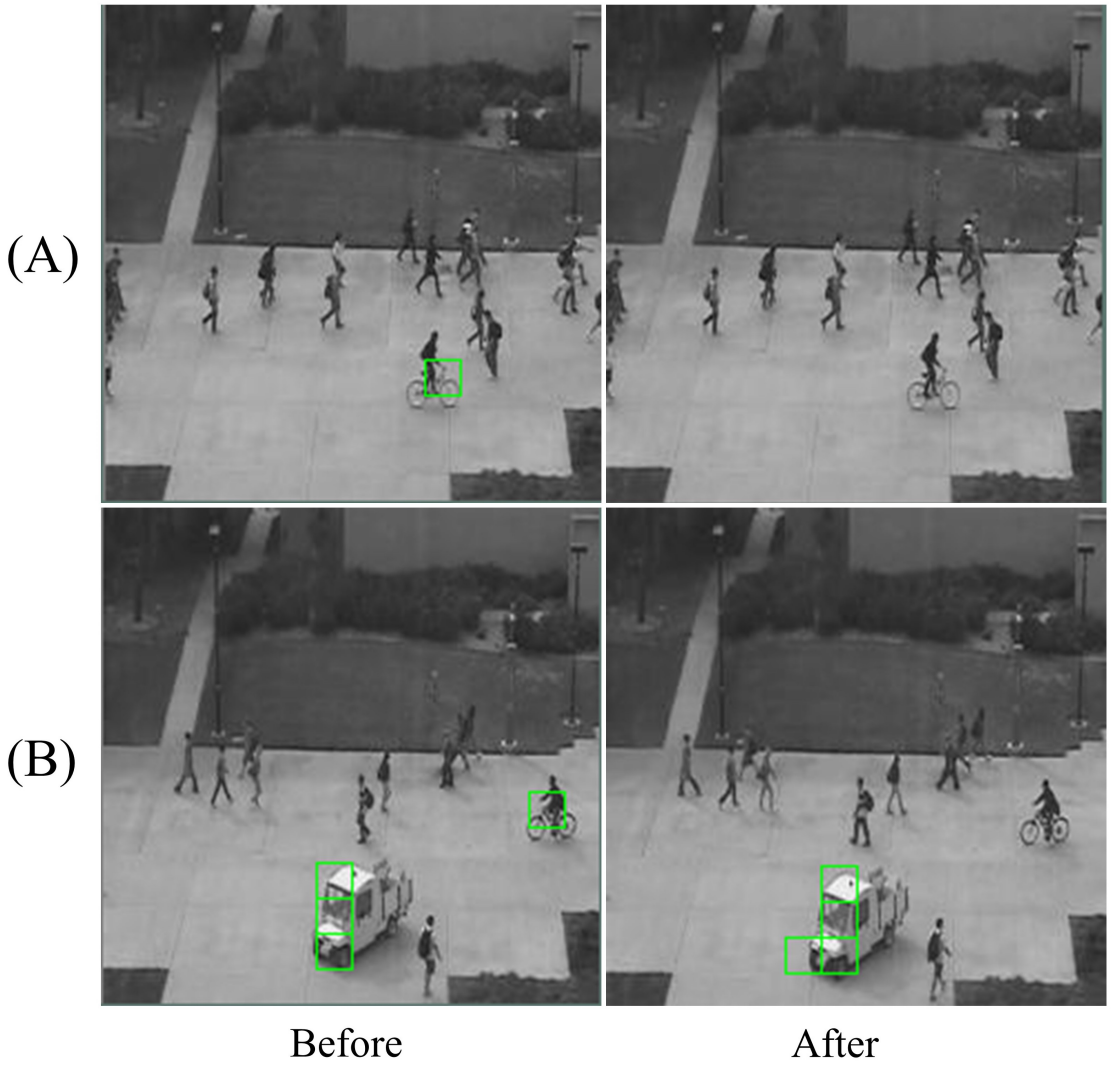


Fig. 5.9 Evaluation dataset from UCSD Ped 2: (A) person riding a bicycle, and (B) person riding a bicycle and a vehicle moving on the pedestrian walk.

Table 5.5 Processing Time Analysis (Seconds per frame)

Process	Ped 1	Ped 2	Avenue
Pre-processing	0.0012	0.0012	0.0012
Representation	0.0293	0.0292	0.0290
Detection	0.0019	0.0019	0.0018
Localization	0.0047	0.0049	0.0042
Total	0.0369	0.0371	0.0360
FPS	27	27	28

Results – Run-time Analysis

We evaluated the real-time video surveillance capability of our anomaly detection approach and the computational overheads for sequenced process of anomaly detection and localization. Table 5.5 presents an overview of the time analysis for our anomaly detection approach in the three datasets evaluated. The averaged processing time for anomaly detection and localization is 37 milliseconds. Achieving approximately 27 frames per second (FPS), ISTL demonstrated the capability for anomaly detection from video surveillance streams in real-time. It should be noted that the difference in processing time for datasets were due to their differences in original resolution as even though frames are resized for anomaly detection, localization is assessed for original frame resolution. For these experiments, ISTL was implemented in series. However, detection and localization can be paralleled, thereby further reducing run time to achieve a higher FPS rate.

5.2.4 Discussion

The Chapter 5 is focused on discovering the constituents of a lifelong learning system and capabilities such system should be composite of. Based on empirical findings, one major scenario where CLL is required when only a small portion of input data is available at the model development time. The reason being that the evolution of the natural environment is inevitable. Similarly, animals and humans do evolve in their behaviour, thus, the distribution of data evolves over time. Thereby, the knowledge in trained learning model becomes obsolete and/or incomplete to make decisions regarding the future. As a solution for this limitation, this thesis looks at utilizing active learning approach, in which we investigate the role that human might need to play in the development and deployment of learning models to accommodate this continuously evolving to provide continual lifelong learning for the learning models.

In this light, we proposed a new spatiotemporal active learning architecture for continual knowledge acquisition. We present our solution in the context of anomaly detection for video

surveillance. The proposed approach addresses three primary challenges of anomaly detection from surveillance video streams by, (i) handling high-dimensional video surveillance data streams in real-time, (ii) formulating the anomaly detection as to learn normality, and (iii) adapting to dynamically evolving normal behaviour with fuzzy aggregation and active learning. The proposed ISTL (Incremental Spatio-Temporal Learner) approach is based on a spatiotemporal autoencoder model consisting of convolution layers that learn spatial regularities and ConvLSTM layers that learn temporal regularities preserving the spatial structure of the video stream. ISTL incorporates a fuzzy aggregation of human observer feedback into continuous active learning process of unknown/new normalities to address the tightly-coupled dependence on a known normality training dataset. It uses two thresholds, anomaly threshold and temporal threshold, based on the context of the video surveillance feed, to overcome sparse evaluation which is based solely on reconstruction error.

Results from experiments conducted on three benchmark datasets demonstrate accuracy, robustness, low computational overhead as well as contextual indicators of the proposed approach, confirming its wide applicability in industrial and urban settings. From a practical perspective of video surveillance, ISTL ensures a human observer is not required to continuously monitor surveillance footage to determine anomalous behaviour. Human involvement is only required for verification of the detected anomalies in practical scenarios and refinement of the learning model.

The proposed ISTL approach, or active learning in general, are well suited for continuous machine learning scenarios where only a small portion of input data is available at the model development time. However, the second paradigm of continuous learning is when only a subset of tasks required are known at the beginning. For such scenarios using deep learning architectures such as CNN, ConvLSTM or autoencoders may not be the best option since the model architecture cannot be changed over time. Further, with utilizing deep learning architectures for continuous learning invokes a number of drawbacks inherent to deep learning. Mainly the requirement of a large volume of data samples for training. This prohibits the use of our approach in scenarios where number of training samples are limited. However, in the context of video surveillance we formulate the problem as a learning to detect normal behaviour, thereby, we were able to acquire enough sample data for training. In contrast, a problem such as object detection or activity recognition would require immense amount of labeled data where in most practical scenarios finding labeled data is unrealistic and generating such labeled data is time-consuming and expensive. In addition, deep learning architecture are complex with a large amount of hyperparameters to be set, and thus, requires a high computational cost.

Evidently an unsupervised deep learning approach such as autoencoder architecture has its own merits, however, with a number of inherent limitations from its deep learning ancestry. Recollecting the biological inspiration we discussed in Chapter 3 and Chapter 4, the proposed RTGSOM self-structuring architectures is capable of overcoming these limitations, however, not yet constructed to handle the two scenarios of CLL: (i) only a small portion of the input data, or (ii) a subset of the tasks is available at once. Thereby, attempts to utilize this self-structuring architecture for CLL with advancements may deem a promising approach to overcome the limitations and drawbacks of ISTL approach.

5.3 Complementary Learning Systems

A successful AI system that accommodates CLL should be able to smoothly update the learning model to consider different tasks and data distributions while still being able to re-use and retain useful knowledge and skills learned previously. Relating to the module 5.3 *Complementary Learning Systems* in chapter overview (Fig. 5.1), this section aims to develop CLL mechanism in computational models by adapting Complementary Learning Systems (CLS) theory that indicates intelligence must possess two learning mechanisms, based on biological discoveries where mammalian neocortex's learning mechanism involved two memory systems: neocortex and hippocampus. The prominent discovery of dual memory based learning for continuous learning was formalised by McClelland *et al.* (1995), in which the neocortex gradually acquires structured knowledge representations while the hippocampus quickly learns the specifics of individual experiences.

The CLS theory identifies that learning system should necessarily be slow for two main reasons. First is that each experience represents a single sensory sample from the environment. Given this, a small learning rate allows a more-accurate estimation of the underlying population statistics by effectively aggregating information over a larger number of samples. Second, the optimal weight of each connection in the memory depends on the values of all the other connections. Before these connections are exposed to experiences, the initial weights of these connections are noisy and weak, leading to a slower initial learning. This slow learning has been both theoretically and practically proven and particularly important in deep neural network architectures that consists of many layers (LeCun *et al.*, 2015).

Although there are advantages of gradual learning system (neocortex of human brain), it suffers from two drastic limitations: i) not being able to learn from an individual experience, and ii) new information severely disrupts the existing knowledge, i.e., catastrophic interference relating to stability-plasticity dilemma. To address both limitations, a second complementary learning system was introduced, enabling rapid and relatively individuated

storage of information about individual items or experiences. The CLS theory proposed that the hippocampus and related structures in the Medial Temporal Lobe (MTL) support the initial storage of experience-specific information. This proposal has been captured in models in which the role of hippocampus structure as recognition of memory for specific items and sensitivity to context and co-occurrence of items within the same event or experience.

Evidence from neuro-biological research demonstrates a replay of recent memories (events and experiences) which occur during offline periods of brain such as during sleep and rest (O'Neill *et al.*, 2010; Wikenheiser and Redish, 2015). According to CLS theory, this phenomenon occurs if the hippocampus can replay the contents of a novel experience back to the neocortex, interleaved with replay and/or ongoing exposure to other experience (Kumaran *et al.*, 2016). This makes new experiences become a part of the repository of experiences that govern the values of connections in the neocortical learning system (Winocur *et al.*, 2010). The previous experiences that are used to interleave with ongoing experience still remain unanswered, yet likely a schema where related experiences activated by the new experiences through dynamics of a recurrent learning mechanism could be used by the biological memory (Kumaran *et al.*, 2016). This seems sensible as the biological memory is structured to function the recall of memory in as auto-associative manner (as detailed in Section 2.3.3).

In this light, this section attempts to address CLL from continuous streams of data when not all the tasks or data are available at once. We conduct a comprehensive literature review on existing methods that adapt CLS theory in achieving CLL in order to identify limitations in the state-of-the-art, followed by a proposal for a novel CLL system based on RTGSOM self-structuring architectures.

5.3.1 CLS Inspired Computational Models

French (1999) presented an early adaption of CLS theory by developing a computational model with two separate memory centres: one for the long-term storage of older memories and another to quickly process new information as it comes in. This method is capable of consolidating memory from fast learning centre to long-term storage. The authors did not explicitly store previous experiences but drew from a probabilistic model. However, the probabilistic methods do not scale up to large-scale industrial datasets as well with images and videos (Parisi *et al.*, 2018).

One of the earliest methods for reducing catastrophic interference in CLL methods is to mix old experiences with new experiences, which is known as *rehearsal* (Hetherington, 1989). For instance, assume a CLL models is required to train to perform 10 tasks in the current session and afterwards another 5 tasks in a study later. One solution could

be to mix experiences from the first study into the later study. Rehearsal technique often uses an external memory system to store all the training examples. Rebuffi *et al.* (2017) proposed a rehearsal based practical strategy for class-incremental learning that learns classifiers and a feature representation simultaneously, named iCaRL (incremental classifier and representation learning). iCaRL consists of a nearest-mean-of-exemplars classifier that is robust against changes in the data representation while needing to store only a small number of exemplars per task. Nearest-mean-of-exemplars rule, which uses the average of extracted feature vectors as the class mean instead of true class mean. In addition, iCaRL uses distillation (Hinton *et al.*, 2015) to prevent information in the learning model deteriorate over time. Overall, iCaRL performs well for incremental class learning, but it still requires storing training examples for each task, making it challenging to scale in real-world settings.

Robins (1995) introduced the concept of *pseudo-rehearsal* arguing against the rehearsal method because the ineffective storage of training examples. Pseudo-rehearsal attempts to generate new examples for a given task, rather than store and replay past training data. In Robins (1995) method, pseudo-rehearsal was conducted by constructing random input vectors, have the model assign them a label and mix them with new training experiences. Recently, Draelos *et al.* (2017) revived the idea of generating experiences based on previous data distributions. The authors developed a generative autoencoder architecture to create pseudo-examples for unsupervised incremental learning. Wu *et al.* (2018) proposed a similar pseudo-rehearsal approach to generate task-specific data for previously encountered classes using a Generative Adversarial Network (GAN). These methods have improved performance of CLL models as the data generated by GAN are more likely to be close to real distribution than the sub-sampled real data. In addition, generating data resolves the privacy issues by not storing individual-specific data. One major drawback with GAN based approaches is the computation overhead by training and performing GAN based models (Choi *et al.*, 2019).

In recent work, FearNet was introduced by Kemker and Kanan (2017) in which the model was constructed as an deep neural network architecture that includes three neural networks: 1) hippocampul complex (HC) for recent memories, 2) medial prefrontal cortex (mPFC) for long-term storage, and 3) basolateral amygdala (BLA) to determine whether to use HC or mPFC for recall. These pseudo-rehearsal based methods are designed for the classification of static images by training examples in random order. However, the continuous streams of data in natural settings incorporate both temporal and spatial dimension thus FearNet lacks capability to process such spatiotemporal data. van de Ven *et al.* (2020) introduced a brain-inspired variant of generated replay in which hidden representations are replayed that are generated by the network's own, context-modulated feedback connections.

Aforementioned methods approach continuous knowledge acquisition from a supervised learning perspective in which the core computation algorithms are based on deep learning architectures. Gepperth and Karaoguz (2016) introduced a self-organization based approach (GeppNet) that reorganize inputs onto a two-dimensional lattice as an attempt to explore continuous learning in unsupervised machine learning paradigm. The SOM serves as the long-term memory followed by a linear regression layer for classification. Once the SOM is initiated, the SOM update occurs only if input is sufficiently novel than the previous data distribution. GeppNet consists of a memory that stores all previous training data for rehearsal. Due to this memory bottleneck, an extension was introduced, named GeppNet+STM, which used a fixed-size memory buffer to store novel examples. The STM (short-term-memory) uses a queue architecture with FCFS (first-come-first-serve) principle, where the STM replaces the oldest examples when the buffer is full. The STM is use to train the SOM in predefined intervals (consolidation phase). GeppNet+STM performs well in retaining base-knowledge as it is re-trained only during its consolidation phase. The GeppNet performs well to learn new data better as it updated the model on every novel input obtained. The major drawback with GeppNet and its variant is the use of limited memory to store all the previous input examples (i.e., entire dataset) throughout the model lifetime. In real-word settings where unlimited memory is not available GeppNet would not yield expected results. Further, GeppNet is not usable in scenarios where previous data is not allowed to store due to privacy regulations.

Addressing the unlimited memory assumption in unsupervised CLL techniques, Parisi *et al.* (2018) proposed a Growing Dual Memory (GDM) self-organization approach leveraging the idea of pseudo-rehearsal to learn new tasks in a short-term memory and progressively consolidate them in a long-term memory. GDM consists of two hierarchically connected recurrent variant of grow-when-required (GWR) (Marsland *et al.*, 2002) self-organizing networks with the complementary tasks of learning object instances and categories. In the absence of external sensory input, the two memory networks periodically replay trajectories of neural reactivations in order to consolidate knowledge, rather relying on an unlimited storage of previously encountered experiences. GDM demonstrates state-of-the-art results for continuous object recognition on benchmark datasets, however, the design is limited to represent a single sensory modality, i.e., Convolutional Feature Extractor (CFE) to extract spatial structure of input images. In despite, the ability of pre-trained CFE to extract representative features from images, it is insufficient to provide a sufficient representation of natural images continuing video. The reason being CFEs such as AlexNet (Krizhevsky *et al.*, 2012), VGGNet (Simonyan and Zisserman, 2014b) are originally designed to detect and classify a single object from ImageNet (Russakovsky *et al.*, 2015) images inclusive of only 1000 categories of objects. The temporal structure is represented through recurrent

information processing capability of recurrent GWR algorithm. In addition, the network graph structure of GWR depends on locality of input data. Therefore, the network can develop different dimensionality for different regions of the network, which can result in visualization difficulties and inability to focus on the representation space in different granular levels, that might be useful for data mining by identifying clusters in the data.

5.3.2 LifeNet: Self-Organization based Complementary Learning Approach

In this section, we propose a new continual lifelong learning architecture (LifeNet) to learn from continuous streams of data associated with different tasks with the goal of augmenting the acquired knowledge. The conceptualization and design of LifeNet is based on three important pillars inspired by the biological brain. They are; 1) multisensory information fusion, 2) dual-memory mechanism that enables continual learning, and 3) the replay of hippocampal memories and interleaved learning as theorized in CLS theory. An overview of LifeNet is presented in Fig. 5.10.

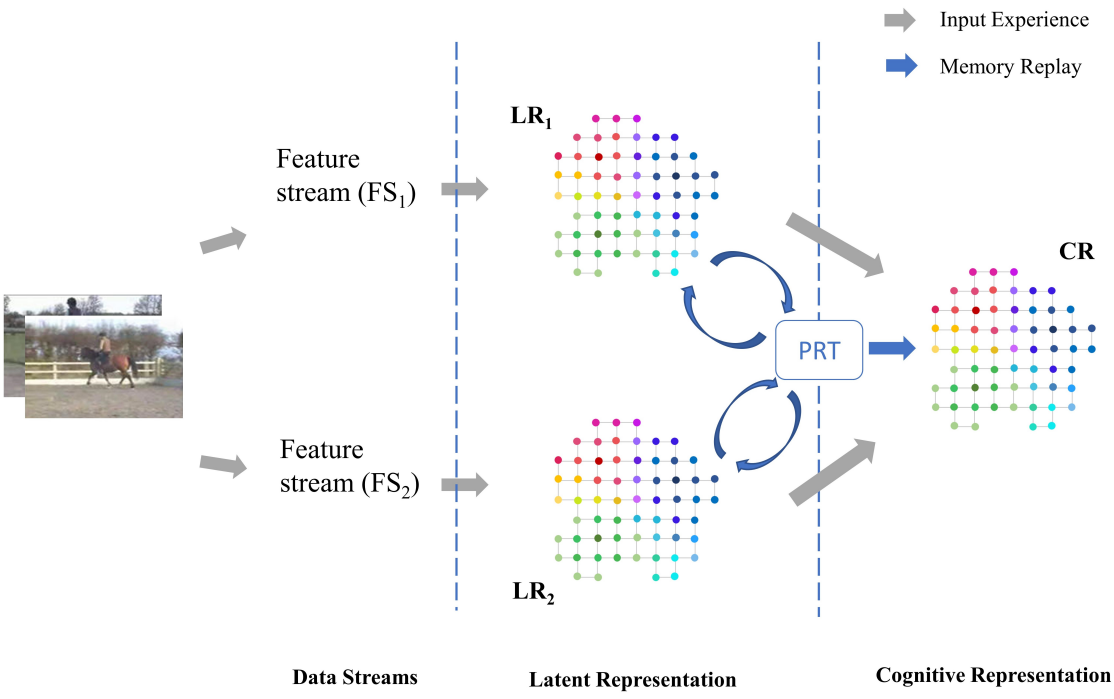


Fig. 5.10 Overview of LifeNet Architecture. PRT=Pseudo-Rehearsal Trajectories.

The natural environments can be presented to LifeNet through multiple modalities or multiple feature streams of a single modality. Each of these modalities and/or character streams are presented to the LifeNet through data streams layer. LifeNet generates a separate pathway for each of the presented feature stream as represented in data stream layer in Fig. 5.10. Each of the feature streams are processed using dual-memory system that are denoted as Latent Representation (LR) and Cognitive Representation (CR), in relation to MSKRF conceptual framework proposed in Chapter 2. Previously experienced memories are accumulated as self-organization trajectories in the Pseudo-Rehearsal Trajectories (PRT) module, which will be replayed to LR and CR in the absence of new experience. The inspiration behind each of these components of LifeNet and its algorithmic development is detailed in following sub-sections.

Multi-modal Data Streams

Similar to the multi-sensory convergence in biological brain (presented in Section 2.3.5), it is pertinent that artificial counterparts adapt an equivalent mechanism to represent and perceive the natural environment. With the advent of sensing technologies, machines have the capability to represent the real-world in machines using multiple modalities and the development of computational processing methods have enabled capture of each modality in multiple characteristics. Thereby, Section 4.3 proposed a multi-stream hierarchical self-organizing architecture to process arrays of feature streams in order to utilize a multitude of modalities and characteristics of the natural environment. We adopt the same multi-stream architecture for LifeNet.

Dual-Memory System

The CLS theory proposed that effective learning requires two complementary systems: one, located in the neocortex, serves as the basis for the gradual acquisition of structured knowledge about the environment, while the other, centered on the hippocampus, allows rapid learning of the specifics of individual items and experiences (McClelland *et al.*, 1995). The neocortex serves as the basis for the gradual acquisition of structured knowledge about the environment, characterized by a slow learning rate and builds overlapping representations of the learned knowledge. Conversely, the hippocampus allows rapid learning of the specifics of individual items and experiences, exhibiting short term adaptation of episodic memory (O'Reilly *et al.*, 2014).

Inspired by the dual-memory mechanism of biological brain, LifeNet is designed with two complementary memory systems resembling the neocortex structure and hippocampus

structure. The first memory module that resemble hippocampus structure is designed as a short-term memory that immediately learns new experiences with a rapid learning rate. The second memory module resembling the neocortex structure is designed to store remote memories for long-term recall with a slow learning rate. Relating to the MSKRF concept framework, we denote the two complementary memories as Latent Representation (LR) for hippocampus and Cognitive Representation (CR) for neocortex. Both of the memory modules are materialized using the proposed RTGSOM (Section 4.2.2) with different parameter settings to resemble continuous learning in the biological brain, where LR is designed with a high learning rate and a high spread factor while CR is designed with a slow learning rate and a low spread factor ($\eta_{CR} < \eta_{LR}$ *and* $SF_{CR} < SF_{LR}$).

From data streams perspective, pathways of LR networks are structurally adapted with respect to each stream. Once the LR acquired knowledge through self-organization from novel input experiences, it is then used to compose the dataset for CR (X_{CR}). Thereby, we evaluate LR for each input experience $x(t)$ in the dataset (X). The LR based experience for CR is composed as weight vectors of the winner nodes (BMU) for each input experience $x_i \in X$ for each stream k, as shown below.

$$X_{CR}^{k} = \{w_{BMU(x_1)}, w_{BMU(x_2)}, ..., w_{BMU(x_3)}\} \tag{5.9}$$

Then the multi-modal experiences are horizontally concatenated from the multiple streams, as shown in equation 5.10, where $k(k \in [1,R])$ is the feature streams and R indicates the number of feature streams. In the case of video inputs, $R = 2$ in order to accommodate spatial and temporal streams.

$$X_{CR} = \{X_{CR}^{1} \cup X_{CR}^{2} \cup ... \cup X_{CR}^{R}\} \tag{5.10}$$

Pseudo-Rehearsal for Memory Consolidation

The CLS theory states that the knowledge formed during hippocampal learning of an experience affords a way of allowing gradual integration of knowledge of experience into neocortical knowledge structures by replaying the experience back to the neocortex, interleaved with ongoing exposure to other experience McClelland *et al.* (1995). This will make the new experience becomes a part of the knowledge of experiences that govern the connections in the neocortical learning system (Frankland and Bontempi, 2005; Winocur *et al.*, 2010). However, which other memories to interleave with remains an open question, where early research assumed the memory replay constitutes of all other recent experiences still stored in hippocampus. In recent research on CLS, Kumaran *et al.* (2016) identified that

it is a possibility that new experiences to be interleaved with related experiences activated by the new experience. That is, instead of replaying all the previous experiences in hippocampus to neocortex, replay of recent experiences to be interleaved with activation of cortical activity patterns consistent with the structured knowledge implicit in the neocortical network (Tononi and Cirelli, 2014).

This process of memory replay in order to slowly integrate new experiences into neocortical representations is labelled 'systems level consolidation' (Frankland and Bontempi, 2005). The memory consolidation is crucial in concurrent learning of regularities (statistics of the environment) and specifics (episodic memories). The hippocampal replay further promotes preferential treatment to unusual/significant memories such as high reward, information content, novelty and surprise. The phenomenon of memory replay, i.e. memory consolidation, generally occurs when the living being is at rest, mostly during the rapid eye movement (REM) sleep (Taupin and Gage, 2002). Overall, the CLS theory holds the means for effectively generalizing across experiences while retaining specific memories in a lifelong manner.

As discussed in Section 5.3.1, early CLL techniques used *rehearsal* approach that mixed old experience with newer experiences (Hetherington, 1989). These techniques assumed an unlimited storage in order to store all previous experiences. The next generation of CLL methods were introduced with *Pseudo-rehearsal* technique that attempts to generate new examples for a given task from a generative model for memory consolidation, rather store and replay past training data (Robins, 1995). The pseudo-rehearsal techniques addressed the unlimited storage assumption for isolated tasks when dealing with stationary inputs (experiences) (Hadsell *et al.*, 2020). However, for the purpose of processing non-stationary experiences such as video data or continuous streams of data from IoT, it needs to account for temporal structure of experiences. On this account, Parisi *et al.* (2018) introduced pseudo-patterns in terms of temporally ordered trajectories of neural activations of a memory network to replay to the neocortical counterpart of artificial dual-memory network. This method is consistent with neurophysiological studies that propose hippocampal replay consists of replay of previously stored patterns of neural activity occurring after an experience (Karlsson and Frank, 2009).

Inspired by the evolution of pseudo-rehearsal techniques, LifeNet accumulates neuronal activation trajectories in Pseudo-Rehearsal Trajectories (PRT) module that consists of a series of activation trajectory matrices (PRT^k) for each feature stream (k). For a given feature stream k, when two neurons (i, j) are consecutively activated at times t and $t+1$ respectively, their activation $PRT^k_{(i,j)}$ is increased by 1. Using the activation trajectory matrices (PRT^k), we can identify the most likely next neuron activation v by:

$$v = argmax_{j \in LR_k/i}(PRT^k_{(i,j)}) \tag{5.11}$$

In pseudo-rehearsal, for each neuron j in LR_k, we generate activation trajectories (S_j) of length Υ as presented in Equation 5.12 where $w_j^{LR_k}$ is the weight vector of neuron j of feature stream k, $s(0) = j$ and $s(i)$ is provided in Equation 5.13 (Parisi *et al.*, 2018).

$$S_j = \{w^{LR_k}_{s(0)}, w^{LR_k}_{s(1)}, ..., w^{LR_k}_{s(\Upsilon)}\} \tag{5.12}$$

$$s(i) = argmax_{p \in LR_k/j}(PRT^k_{(p,s(i-1))}), i \in [1, \Upsilon] \tag{5.13}$$

The generated PRTs are interleaved with new input that becomes available consisting of new tasks are replayed to LR and CR. The replay will occur after each learning iteration over a batch of new sensory observations. This contrasts with storing previously encountered input experiences (or instances) and replay them to the training model (rehearsal approach). Instead, we generate PRTs periodically to replay interleaved with new data in order to overcome catastrophic forgetting.

5.3.3 Evaluation of LifeNet

The previous section introduced the LifeNet to learn from continuous streams of data associated with different tasks with the goal of augmenting acquired knowledge. The LifeNet architecture was materialised using a formation of RTGSOM self-structuring algorithm. In this section, we conduct a series of experiments to validate the proposed LifeNet and the selection of RTGSOM algorithm in order to validate our contribution and confirm its suitability in real-world settings. The series of experiments are three-fold;

1. First, a demonstration of the capability of LifeNet in lifelong scenarios where a subset of the tasks is available at once. We conduct a qualitative analysis on how the representation is self-structured when new tasks become available over time using MNIST hand-written digit benchmark dataset. This experiment validates the benefit of structural adaptation provided by RTGSOM in continuous learning.

2. Second, an analysis and evaluation of classification accuracy of the proposed LifeNet model with CIFAR-100 object dataset in a strictly class-incremental setup. This experiment validates the proposed method obtains comparable results with state-of-the-art deep learning methods for CLL.

3. Third, an exploration of the applicability of LifeNet using multi-stream input features with human activity videos with Weizmann and KTH datasets. To the best of our knowledge, human activity videos have not been explored in a strict class-incremental setup, thereby, this experiment provides a new direction to validate continuous learning capabilities of CLL methods.

Structural Adaptation of Continuous Learning

We demonstrate the capability of LifeNet to structurally adapt its representation when new tasks become available using the well-known MNIST handwritten digit recognition benchmark (LeCun *et al.*, 1998). The experiment represents a task of moderate difficulty and real-world relevance problem with 10 categories, i.e., digits 0 to 9. A subset of the MNIST dataset is depicted in Fig. 5.11. The dataset is split into 5 subsets with each subset containing 2 digits. LifeNet is trained using strictly subset-incremental setup, where we incrementally introduce a subset at each iteration. The LifeNet architecture for MNIST consists only a single feature stream, as the objective of this experiment is to demonstrate the qualitative effects of self-structuring.

Fig. 5.11 Subset of MNIST database of handwritten digits.

We compose subsets of data samples from MNIST as $(0,5),(1,6),(2,7),(3,8),(4,9)$. We use the following parameters for the LifeNet model trained for this experiment, where the spread factors (SF) for LR and CR are selected as, 0.8 and 0.6. The factor of distribution is set to 0.1 ($FD = 0.1$) and $R = 3.8$. Initial learning iteration count was set to 100 and forthcoming learning and smoothing iterations count were set to 50, in order to develop a sufficiently larger self-organized network at the inception in order to adapt effectively once

new classes are introduced in forthcoming steps. The transience threshold (M) is setup as $M^{LS} = 40$ and $M^{CR} = 45$. The optimal values for SF and M were selected using a random grid search method (Bergstra and Bengio, 2012). The input experiences in context (MNIST digits) are stationary, thus, we selected $\Upsilon = 1$ as the length of activation trajectories to for interleaved memory replay.

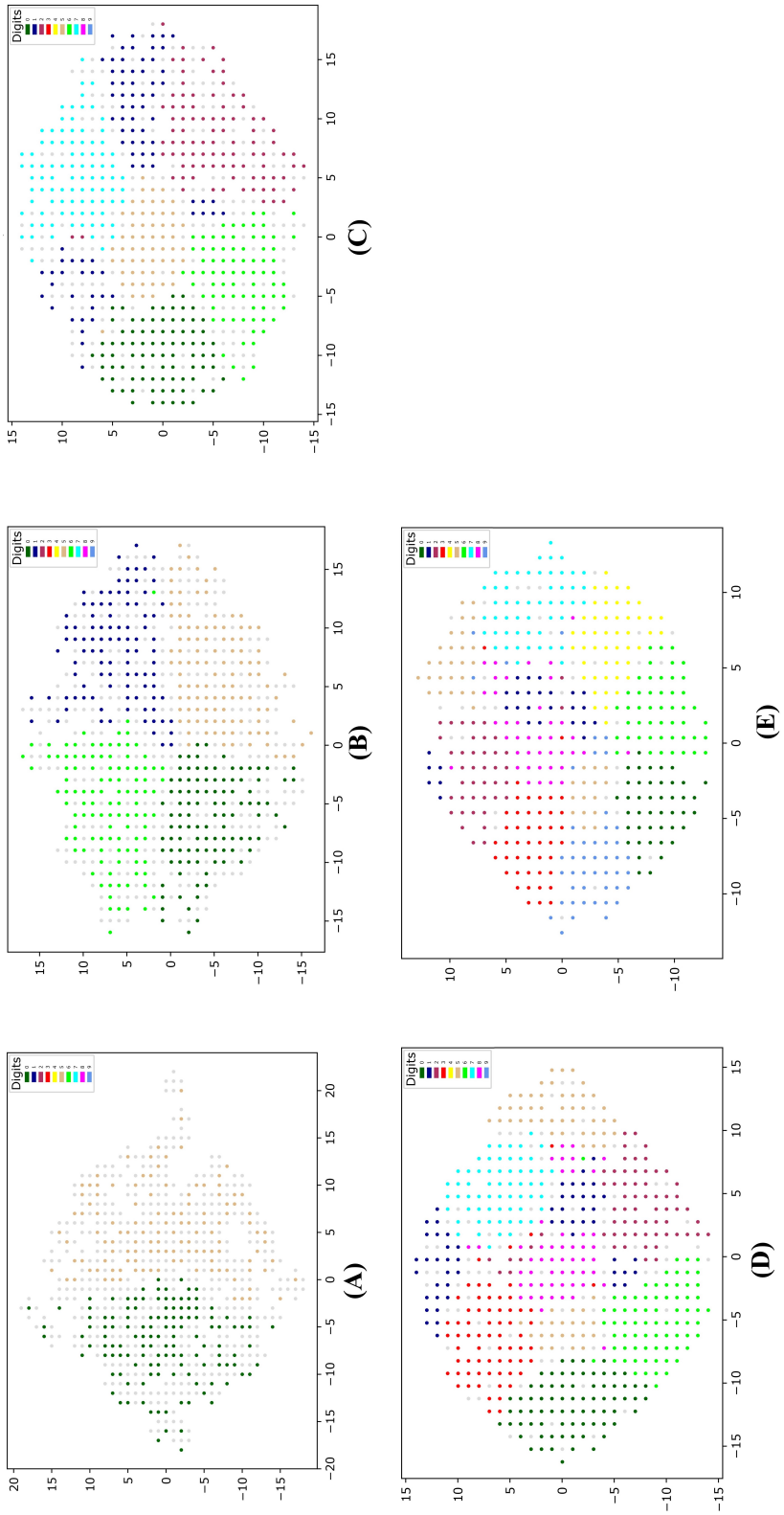


Fig. 5.12 Structural adaptation of MNIST dataset over 5 time steps. Incrementally presented subsets are represented in each visualization where $A = (0,5), B = (1,6), C = (2,7), D = (3,8), E = (4,9)$.

The cognitive representation of the experiment is visualized in Fig. 5.12, where incrementally presented digit subsets are presented in $A = (0,5), B = (1,6), C = (2,7), D = (3,8)$ and $E = (4,9)$. In the first step, instances of 0 and 5 were completely separated by the midpoint of network as shown in Fig. 5.12 (A). The introduction of digits classes 1 and 6 in Fig. 5.12 (B) has equally divided the self-organized network into 4 quadrants. If the interleaved memory replay has not taken in place, such an equal division of the memory to represent current classes is not possible.

Instances of class 2 and 7 were introduced in the next time step as illustrated in Fig. 5.12 (C), that shows a re-organization of classes making visually similar instances of digits come closer. For instance, 7 has organized to penetrate in between prototypes of digit 1, which can be understood by their similar visual appearance. As shown in Fig. 5.11 the input samples of class 1 are closely related to class 7 based on the how different people write the vertical line by hand. Similarly, Fig. 5.12 (D) and (E) incrementally learn the remaining classes.

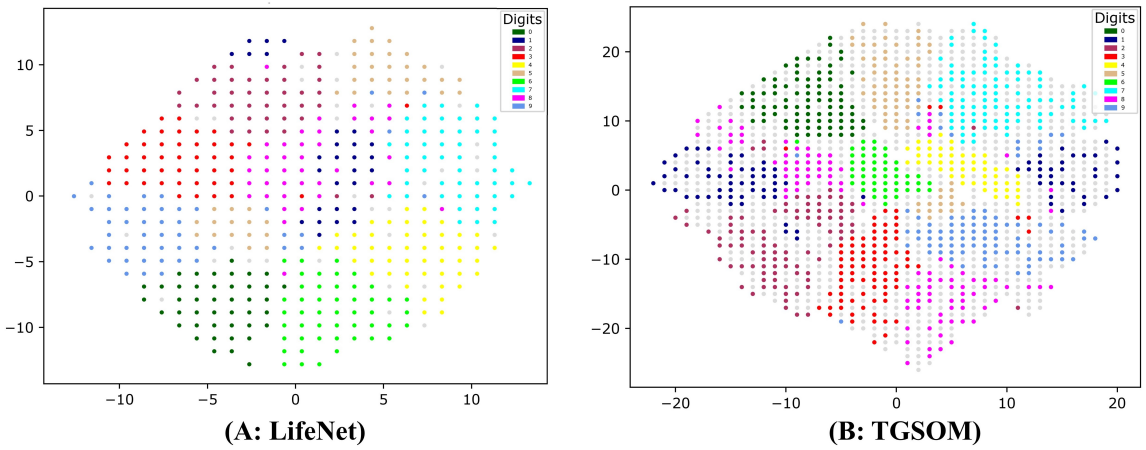


Fig. 5.13 Structural adaptation of LifeNet and TGSOM.

Fig. 5.13 illustrates a comparison of the structural adaptation of LifeNet with respect to TGSOM. The RTGSOM was trained with all the input instances (all classes) were available at once, while LifeNet utilized two memory networks with interleaved memory replay. TGSOM learns a raw representation of the input data where LifeNet learns more abstract representation due to the use of secondary memory connected hierarchically. The qualitative outcome clearly demonstrates that LifeNet has learnt a concise representation comparatively to TGSOM, which can be explained using hierarchical abstraction in biological memory storage that we adopt in LifeNet computational modelling (Section 2.3.4).

Multi-class Incremental Object Learning

The second experiment intends to analyze and evaluate classification accuracy of the proposed LifeNet model with multi-class incremental object learning scenarios. The experiment is setup in a manner that classes arranged in a fixed random order and LifeNet is trained in a class-incremental way on the available training data instances. After each batch of classes, the resulting classifier is evaluated on the test part data of the dataset, considering only those classes that have already been trained. Abiding Kemker and Kanan (2017) evaluation criterion, we select two measures to evaluate the accuracy; i) Ψ_{all} measures test accuracy accounted for how well the seen class were consolidated in memory, and ii) Ψ_{new} measures test accuracy on the most recently trained class. It is important to note that Ψ_{new} only accounts the test accuracy for newly introduced class regardless how the model performed on previously seen classes.

LifeNet model was experimented on CIFAR-100 (Krizhevsky *et al.*, 2014) object learning benchmark datasets. CIFAR-100 dataset consists of 100 classes containing 600 images each, where the 100 classes are grouped into 20 super-classes. It contains 500 training images and 100 testing images per class. A sample set of images are illustrated in Fig. 5.14. Features for CIFAR-100 dataset was extracted using the feature map after first 9 convolutional layers of VGG-19 (Simonyan and Zisserman, 2014b) pretrained on ImageNet similar to Choi *et al.* (2019).

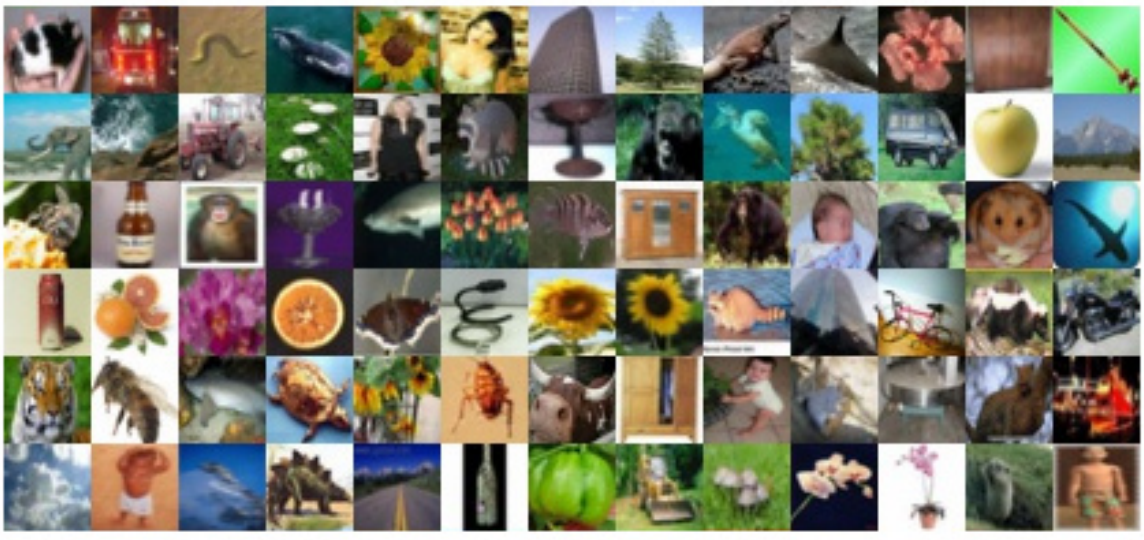
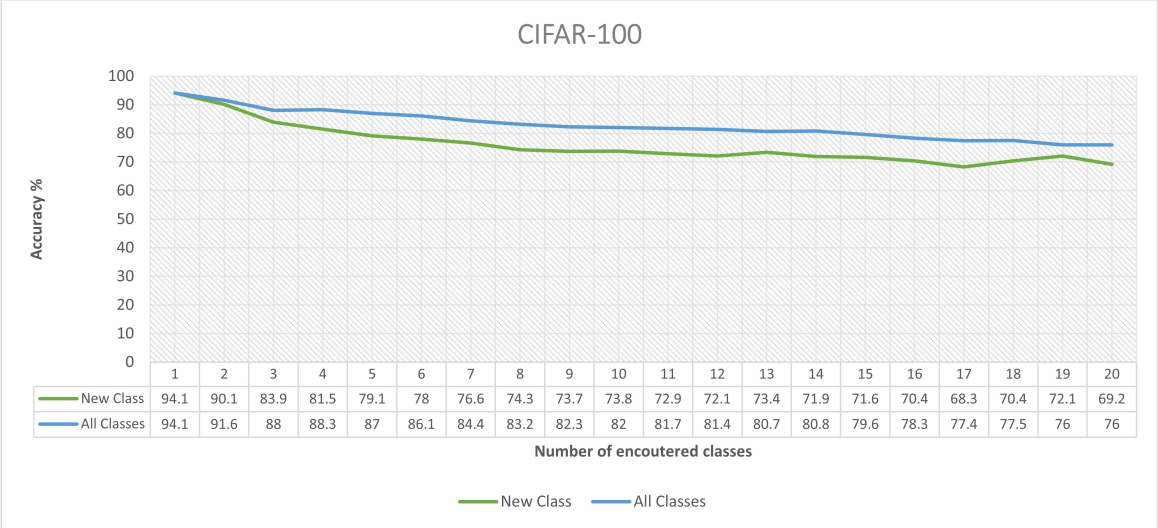


Fig. 5.14 Sample images of CIFAR-100 object dataset.

CIFAR-100 dataset was trained using LifeNet in a strictly class-incremental setup, i.e., training instances for a single class was made available for each batch of training. At each

step, test accuracy for Ψ_{new} and Ψ_{all} were calculated. The evaluation measure is the standard multi-class accuracy on the test set. Following parameters were used for the LifeNet model. The spread factors (SF) for LR and CR are selected as, 0.8 and 0.6. The factor of distribution is set to 0.1 ($FD = 0.1$) and $R = 3.8$. Initial learning iteration count was set to 100 and forthcoming learning and smoothing iterations count were set to 60. The transience threshold (M) is setup as $M^{LS} = 50$ and $M^{CR} = 55$. The optimal values for SF and M were selected using a random grid search method (Bergstra and Bengio, 2012). The input experiences in context are stationary objects, thus, we selected $\Upsilon = 1$ as the length of activation trajectories to for interleaved memory replay.

The continual class-incremental accuracy evolution is presented in Fig. 5.15. When new classes are introduced at the second step, it shows a drastic drop in accuracy. The classification accuracy of the newly introduced class is lower than the accuracy of all classes combined, which emphasizes that LifeNet has the capability to consolidate memory and to improve knowledge over time. This is excepted as the primary objective of interleaved memory replay is to consolidate memory by overcoming catastrophic forgetting.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
New Class	94.1	90.1	83.9	81.5	79.1	78	76.6	74.3	73.7	73.8	72.9	72.1	73.4	71.9	71.6	70.4	68.3	70.4	72.1	69.2
All Classes	94.1	91.6	88	88.3	87	86.1	84.4	83.2	82.3	82	81.7	81.4	80.7	80.8	79.6	78.3	77.4	77.5	76	76

Fig. 5.15 Class-incremental Multi-class classification of CIFAR-100.

In addition, we compare the overall accuracy of LifeNet with respect to four state-of-the-art CLL approaches. The methods selected for comparison include, i) SOM based GeppNet (Gepperth and Karaoguz, 2016), ii) GeppNet with Short term memory (STM) (Gepperth and Karaoguz, 2016), iii) Deep autoencoder based FearNet, and iv) Deep autoencoder based Choi *et al.* (2019) with Memory Aware Synapses (MAS). The final mean-class classification accuracy comparison is presented in Table 5.6. The results demonstrate the proposed

Table 5.6 CIFAR-100 Classification Results

Model	Ψ_{new}	Ψ_{all}
GeppNet	0.529	0.754
GeppNet+STM	0.408	0.800
FearNet	0.824	0.947
Choi (MAS)	0.641	0.850
LifeNet	0.759	0.828

LifeNet obtains comparable results with deep learning based models while outperforming existing unsupervised self-organization based approaches.

Multi-class Incremental Human Action Learning

The third experiment intends to explore the applicability of LifeNet in a natural setting using multi-stream input features with human activity videos. In general, human activity classification aims to predict the activity label of unseen activity video samples. Such classification will provide means in automated systems for assisted living in smart homes, healthcare monitoring applications, monitoring and surveillance systems. The class-incremental human activity learning experiment is setup in a manner that action classes arranged in a fixed random order and LifeNet is trained in a class-incremental way on the available training data instances. After each batch of classes, the resulting classifier is evaluated on the test part data of the dataset, considering only those classes that have already been trained. The accuracy of the model is evaluated using standard multi-class accuracy for i) Ψ_{all}, which measures test accuracy accounted for how well the seen class were consolidated in memory, and ii) Ψ_{new} measures test accuracy on the most recently trained.

LifeNet was evaluated for multi-class incremental video action learning using two benchmark datasets: Weizmann (Blank *et al.*, 2005) and KTH (Schuldt *et al.*, 2004) human activity datasets that we previously used to evaluate hierarchical two-stream growing self-organizing approach for HAR in Section 4.3.2. Weizmann dataset contains human actions of 9 different subjects in 10 action classes. KTH dataset contains human actions of 25 different subjects, where each subject has approximately 24 segments (total of 599 action video segments). KTH dataset contains 6 human actions. Comparatively action recognition in KTH dataset is difficult due to varying viewpoints with objects appear distant from the camera. Both datasets contain actions such as boxing, hand clapping, hand waving, running, walking, jogging, bending, etc. captured in indoor and outdoor environments.

For action representation, raw action video frames were extracted, converted into gray-scale and resized to 32×32 pixels. The processed frames were normalized to range 0-1 and

subtracted by the mean in order to centre the input video data. The HOG and HOOF features are extracted from the pre-processed video frames as 8-bin histograms. These feature streams were used as inputs to the LifeNet model that consists of two feature streams for HOF and HOOF.

Datasets were trained using LifeNet in a strictly class-incremental setup, i.e., training instances for a single class was made available for each batch of training. At each step, test accuracy for Ψ_{new} and Ψ_{all} were calculated. Following parameters were used for the LifeNet model. The spread factors (SF) for LR and CR are selected as, 0.8 and 0.6. The factor of distribution is set to 0.1 ($FD = 0.1$) and $R = 3.8$. Initial learning iteration count was set to 100 and forthcoming learning and smoothing iterations count were set to 60. The transience threshold (M) is setup as $M^{LS} = 50$ and $M^{CR} = 55$. The optimal values for SF and M were selected using a random grid search method (Bergstra and Bengio, 2012). The input experiences in context are non-stationary human action, thus, we selected $\Upsilon = 3$ as the length of activation trajectories to for interleaved memory replay.

The continual class-incremental accuracy evolution is presented in Fig. 5.16. In both the instances, when new classes are introduced the accuracy tends to drop. However, it is interesting to note that there is a steep accuracy drop when two similar actions are presented such as walk/run and jump/jump-jack. This can be expected because when similar activities are presented the algorithm will face difficulty in distinguishing between the two.

The authors were unable to find any existing work on CLL based methods for human activity recognition, thus, we compare the overall accuracy of LifeNet with respect to four non-CLL based models including the hierarchically connected multi-stream RTGSOM architecture proposed in Chapter 4.3. The results are presented in Table 5.7 that demonstrate LifeNet obtains comparatively lower accuracy (close to -18%) with respect to hierarchical multi-stream approach we proposed prior for both Weizmann and KTH datasets. That should be expected as the entire dataset of human actions were available to all the compared state-of-the-art HAR models at beginning, while LifeNet was trained using a strictly class-incremental manner which makes it susceptible to catastrophic forgetting. Despite above LifeNet evaluation on Weizmann data set outperformed Peng *et al.* (2018) HAR model, which indicates a promising direction for HAR in CLL paradigm.

In addition, to the best of our knowledge, human activity videos have not been explored in a strict class-incremental setup, thereby, this experiment provides a new direction to validate continuous learning capabilities of CLL methods.

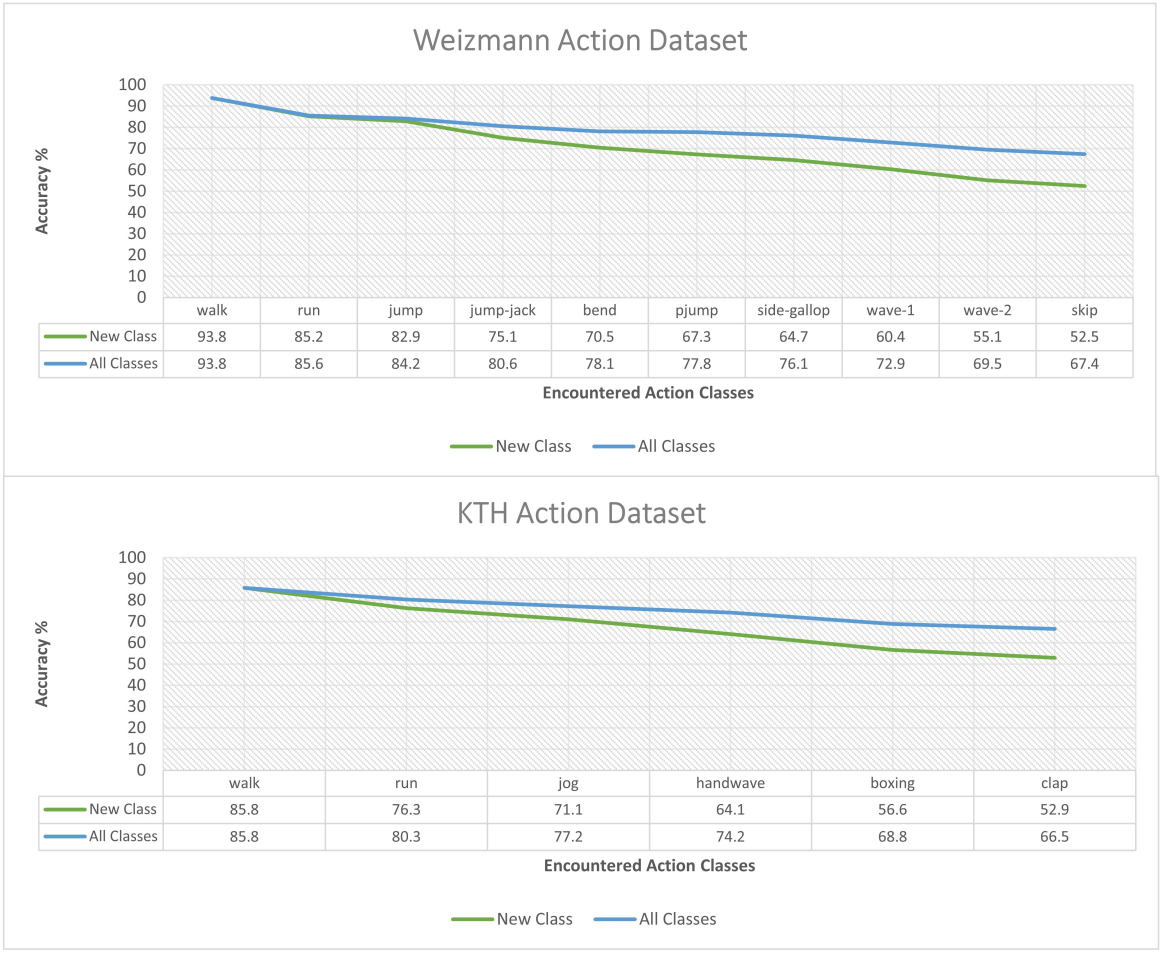


Fig. 5.16 Class-incremental Multi-class classification of Weizmann and KTH datasets.

Table 5.7 Classification Results

Proposed Method	Accuracy (%)	
	Weizmann	KTH
Yang *et al.* (2012)	91.0	91.0
Umakanthan (2016)	-	91.2
Liu *et al.* (2016)	-	**94.3**
Peng *et al.* (2018)	77.8	83.4
H-RTGSOM (Ours)	**96.4**	93.6
LifeNet	78.6	75.5

5.3.4 Discussion

In this section we developed a Continuous Lifelong Learning (CLL) computational model to learn from continuous streams of data adapting to the external environment and associate with different tasks with the goal of augmenting the acquired knowledge. Drawing on the Complementary Learning Systems (CLS) theory that indicates intelligence must possess two learning mechanisms: neocortex and hippocampus, along with the periodic interleaved memory replay from hippocampus to neocortex, we designed a new CLL architecture, LifeNet. The LifeNet architecture consists of 3 concept pillars, that are, 1) multi-sensory information fusion phenomenon in biological perception systems, 2) dual-memory mechanism of the biological brain that enables continual learning, and 3) the replay of hippocampal memories and interleaved learning in biological brain as theorized in CLS theory. The materialization of the proposed LifeNet model was based on RTGSOM core algorithm that we developed in Chapter 4.

We position the LifeNet architecture in the proposed MSKRF as presented in Fig. 5.17. The Latent Representation (LR) is designed as a series of RTGSOM neuronal networks to accommodate multiple feature streams originate from the environment, i.e., sensory input data module in MSKRF. The Cognitive Representation (CR) is designed using a single RTGSOM neuronal network that fuses information of the multiple LR networks. In addition, the connectivity of LR and CR is modulated by Pseudo-Rehearsal Trajectories (PRT) that resemble interleaved memory replay of hippocampus. The biological base (from Chapter 2) we materialize in this chapter is *Complementary Learning Systems* theory, which is inspired by the dual-memory module concept and hippocampal replay.

The proposed LifeNet model was validated and confirmed its suitability for intelligent video surveillance using a series of experiments. In the first experiment, we demonstrate the capability of LifeNet in lifelong scenarios where a subset of the tasks is available at once. We analyzed the qualitative effect on how the representation is self-structured when new tasks become available over time steps using MNIST hand-written digit benchmark dataset. This experiment validates the benefit of structural adaptation provided by RTGSOM in continuous learning. The second experiment is aimed at evaluating the classification accuracy in a continual learning scenario, where new tasks/classes become available progressively. The experiment was conducted using CIFAR-100 benchmark object dataset in a strictly class-incremental setup, and results validated LifeNet's capability to obtain comparable results with state-of-the-art deep learning methods for CLL. The objective of the third experiment was to explore the applicability of LifeNet using multi-stream input features. We used Human Action Recognition (HAR) as the case study (comparable to our approach in Section 4.3) that use spatial and temporal features as feature streams. To the best of our knowledge,

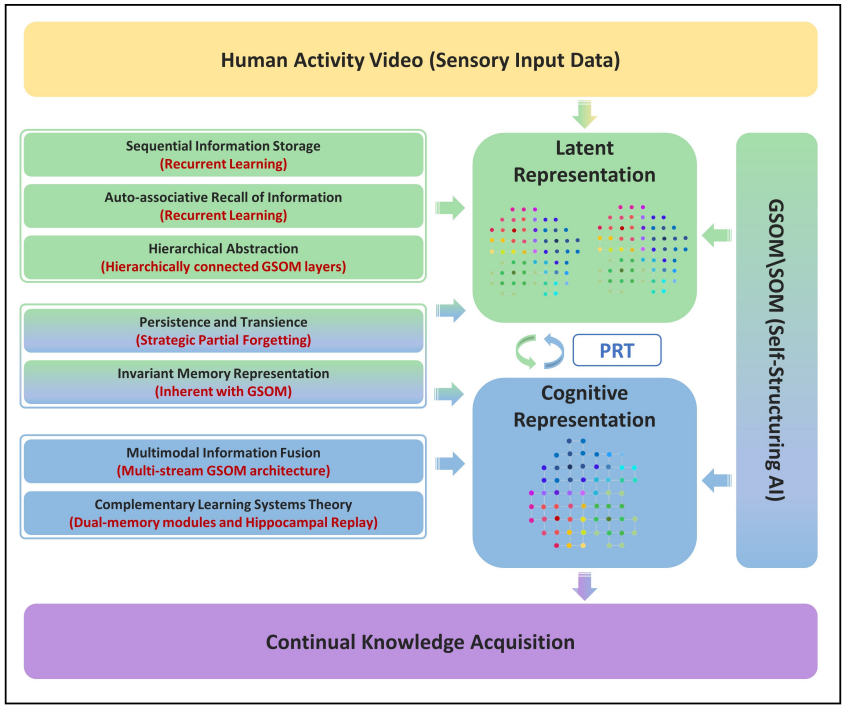


Fig. 5.17 Positioning the experiments in MSKRF.

human activity videos have not been explored in a strict class-incremental setup, thereby, this experiment provides a new direction to validate continuous learning capabilities of CLL methods.

5.4 Summary and Research Questions Revisited

In general, the typical sequence of industrial AI adaptations is to gather data, learn the underlying structure of the data, develop predictive/diagnostic models to conduct specific industry related task and deploy the model to systematically perform these specific tasks. Gathering, preparing, and enriching the right data is essential and remains a key bottleneck among these many industries wanting to use AI. This phenomenon emphasizes the classical AI paradigm, known as Narrow Artificial Intelligence (ANI), that perform in isolation. Given a dataset, a learning algorithm is applied to a dataset to produce a model without considering any previously learned knowledge. This paradigm needs many training examples and is only suitable for well-defined and narrow tasks in closed environments (Chen and Liu, 2018). Whenever new data are available, the training process of the AI system has to start all over again.

In contrast, the biological counterpart, the human brain, is capable of constantly adapting to and exploiting new information in this complex, constantly changing and evolving physical world. The key characteristic of *human learning* is the continual learning and adaptation to new environments. Humans accumulate knowledge gained over a lifetime and use this accumulated knowledge to assist future learning and decision making with possible adaptations. That is humans are able to discover new tasks and learn while performing the tasks in open environments in a self-supervised manner. The design of *human learning* provides a promising premise for AI systems to achieve true intelligence with continuous lifelong learning.

On this basis, we presented the need for CLL and a discussion of challenges in achieving CLL in Section 5.1. The first section introduced the tasks associated with CLL that needs to be addressed in order to achieve CLL in computational models. These challenges are two-fold where the first relates to the evolving nature of input stimuli (data) while the second relates to the evolving nature of tasks computational models attempts to achieve. Section 5.2 addressed the first challenge associated with evolving nature of data, followed by proposing a new unsupervised deep learning based active learning approach in Section 5.2.1 and its experimental evaluation in the context of intelligent video surveillance focused on anomaly detection in Section 5.2.2.

The second challenge in CLL, the evolving nature of tasks computational models focus on capturing is addressed in Section 5.3, followed by an in-depth review on existing complementary learning systems (CLS) based computational models, both supervised and unsupervised, is presented in Section 5.3.1. The analysis of the existing work has led to the design and development of a new self-organization based CLL approach, named LifeNet, by incorporating constituents of CLS theory. LifeNet is designed using an architecture of RTGSOM and incorporates the biological base: *complementary learning system inspiration*. The LifeNet completed the materialization of the proposed MSKRF, which was evaluated using a series of benchmark datasets on object recognition and human activity recognition in Section 5.3.3.

The journal article entitled *Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance* (Nawaratne *et al.*, 2019c) was originated from this Chapter.

Overall, this chapter addressed the second research questions (RQ2), stated in section 1.4: **What are the computational and machine learning constituents of continuous lifelong learning for materializing the proposed conceptual framework?** The RQ2 consists of 5 sub-questions, where the first four sub-questions were addressed in Chapter 3 and Chapter 4. The current chapter successfully addressed the fifth sub-question: **"What neurophysiological theories enable the development of a computationally plausible memory**

formulation to achieve continuous lifelong learning?", with the conceptualization, design and development of; i) the deep spatiotemporal autoencoder based active learning approach, and ii) the LifeNet architecture.

Therefore, by proposal, design and development of the seventh and final biological base, the computational modelling and implementation of MSKRF proposed in Chapter 2 has been successfully achieved.

Chapter 6

Self-Structuring AI to Empower Smart Cities and Digital Health

The thesis started with the proposal of an overarching conceptual framework, *Multi-layered Self-structuring Knowledge Representation Framework* (MSKRF) to facilitate development of a new breed of AI capable of continuous lifelong learning. The innovations were inspired by the neurophysiological functionality of human brain and catered to the needs of big data and digital environment. Identifying the pertinence of self-structuring topological representations and unsupervised learning to lay the foundation of this framework, Chapter 3 proposed Growing Self-Organizing Maps (GSOM) to be used as a viable algorithmic base for MSKRF. The representation capability of GSOM was widened in Chapter 4 to overcome catastrophic forgetting and to address stability and plasticity by proposing algorithmic modifications to develop: i) transience (forgetting) to improve plasticity without compromising stability of the representation, ii) recurrent learning to accommodate non-stationary continuous streams of data, and iii) multi-stream information fusion to accommodate multiple characteristic streams from data. A complementary memory formulation using the modified representation algorithm (RTGSOM) was proposed in Chapter 5 to achieve continuous lifelong learning that enable MSKRF to learn from continuous streams of data. This gave the underlying AI the ability of adapting to the external environment and associate with different tasks and new data distributions thus augmenting the acquired knowledge for problem solving and future learning.

This chapter demonstrates the novel computation models developed in this thesis using two key AI application areas: *Smart City* and *Digital Health*. The first case-study focuses on Smart Cities that endeavour to deliver safe, sustainable, effective asset utilization and service provision, amidst rapid urbanization. With the recent advent in deep learning, new technologies have been developed to address object detection and recognition capabilities

in Smart City surveillance video streams. However, several issues are unaddressed, such as sub-optimality, latency, predictive accuracy, and most importantly the contextualization of all detected salient objects for further decision-making. In Section 6.1, we propose a generative self-structuring and deep learning based approach to address these challenges, and demonstrate an adaptation for a license plate detection use-case. An evaluation of the proposed approach is conducted using a state-of-the-art benchmark dataset that captured using a single low-cost camera under different weather and recording conditions, in a realistic setting.

Section 6.2 presents the second case-study in the scope of Digital Health focused on neuroscience and mental health. The National Institute of Health Stroke Scale (NIHSS) is used worldwide to classify stroke severity, where recent evidence suggests that 50% of survivors have a 'mild' stroke based on this scale. Yet, survivors argue that the classification of 'mild' in contrast with their experience and they often have ongoing problems that impact their daily lives (Hand *et al.*, 2014). In this case-study, we systematically investigate different profiles of survivors classified as 'mild' based on tests measuring their stroke impairment and impact using self-structuring latent representations to uncover latent patterns from multiple measures, at different times in the recovery trajectory; with ongoing impairments and impact even 12-months post-stroke.

The following sections (Section 6.1 and Section 6.2) present the two case-studies in detail relating to how the novel computational developments introduced by this thesis enable them to successfully achieve each objective.

6.1 Smart-City: License Plate Recognition

Smart cities are an emerging paradigm, where rapid urbanization necessitates the enhancement of service delivery and asset utilization for increasing populations. Physical infrastructure, transport networks, power grids, as well as social and economic systems within an urban environment are witnessing an exponential increase of data. Video data streams generated by CCTV camera systems are a predominant consideration in the safety, efficacy and sustainability of such environments, where rapid developments in urbanization and mobility is inevitable with increased vehicular, pedestrian, as well as goods and services movement (Liu *et al.*, 2013).

This has led smart cities to become advanced digital ecosystems where many facets of these smart cities being represented via a number of sensors that capture the environment by generating large, multi-modal, multi-source, dense, high frequent and non-stationary datasets. With this proliferation of data generation, it is unrealistic and infeasible for human

observers to monitor and analyse every video stream with high precision. AI techniques can transform video data streams into actionable insights that provide the strategic advantage of safe, sustainable, effective operations and processes required for a smart city. For instance, face detection and face verification facilitates the localization and detection of persons of interest and perpetrators, vehicle speed measurement enables the identification of vehicles travelling above a speed limit, and detection of anomalies in industrial video streams enabling the detection of behaviors that do not conform to expected and accepted norms (Nawaratne *et al.*, 2019c).

Despite numerous approaches reported in the existing literature, several challenges remain unaddressed. First is the poor performance due to external factors related to recording and weather conditions such as level of lighting, motion blur, occlusion and varying levels of recording resolution due to low cost CCTV systems. Second is that none of the approaches report on the predictive accuracy in terms of the level of confidence of detection. Performance is generally evaluated using a test dataset to determine the detection accuracy, yet, not being able to provide an estimation of confidence for diverse objects as well as detection in different weather/recording conditions. Third is the computational complexity and cost of video data processing. Many existing approaches have difficulties in processing high-resolution imagery in real-time. In order to achieve real-time performance in practical scenarios, video frames of the surveillance video feed should be processed with minimal latency.

In this case-study, we propose a new generative self-structuring and deep learning based approach, Generative Latent Space (GenLS), to address the aforementioned challenges that affect real-time smart city surveillance. The development of GenLS is stimulated by the computation formulation of RTGSOM algorithm proposed in Section 4.2 and have been contributed by the theoretical formation of four biological bases, that are; (i) invariant representation of memory, (ii) transience of memory, (iii) sequential information storage, and (iv) auto-associative recall of memory, which have been introduced in Section 2.3. This development is integrated into the latent space of the GenLS, which is a learned abstraction of the input feature space. This learned abstraction encapsulates the interplay across the feature set and all input vectors on to a low dimensional topology (usually two-dimensions) that can be visualized, aggregated and synthesized for further exploration of patterns, outliers and feature importance. A latent space can accommodate varied forms of input features, such as raw data, pre-processed data as well as intermediate outcomes from other machine learning pipelines. The low dimensional topology itself provides a contextual mapping for such intermediate learning outcomes, supporting increased predictive accuracy. Latent spaces have been used for industrial smart city applications such as anomaly detection from surveillance video and human activity recognition (Nawaratne *et al.*, 2017, 2019a,b).

GenLS begins by detecting salient object regions and estimates a localized bounding box using a deep Convolution Neural Network (CNN). To generate the localized bounding boxes of salient object regions, the input video frames are divided into overlapping sub-regions using a region proposal network, followed by a CNN feature extraction network to learn a feature representation of salient objects to concisely distinguish with respect to the background. The feature representation of each sub-region is then evaluated for existence of the interested objects and bounding boxes are estimated to localize the objects. Next, the features representing the bounding box (i.e., Fully connected layer of the CNN) is input to a growing, self-organized latent space that is generated via unsupervised learning. Finally, this growing self-organized latent space generates a representation of detected salient objects with corresponding confidence and the contextualization of all objects for post-processing. In terms of predictive value for smart city scenarios, GenLS will generate a bounding box for each salient object region of interest from surveillance video alongside a confidence metric indicating the potentiality of the prediction in real-time, that can be used for contextualization of the entire video stream leading to further decision-making.

GenLS is applicable to many smart city surveillance scenarios that generate continuous video data streams, such as motor traffic regulation, crowd behavior classification, commodities movement, facilities provision, productivity improvements, and anomalous event detection. In practice, motor traffic regulation is more prevalent than other surveillance scenarios. For this reason, we selected the technically challenging use-case of License Plate Detection (LPD) within the motor traffic regulation scenario to demonstrate the application of GenLS to smart city surveillance. LPD is useful for all functions of safe, sustainable, effective asset utilization and service provision with implementations in citizen safety, efficient parking utilization and sustainable asset usage with intelligent toll and traffic management (Polishetty *et al.*, 2016). On that note, the research contributions of this case-study are as follows:

1. A generative latent space (GenLS) approach that consists of a Convolution Neural Network and self-structuring topological neural network that represents autonomously detected salient objects with corresponding confidence and the contextualization of all objects for post-processing and further decision-making.

2. A fine-tuned deep Convolution Neural Network based architecture for the use-case of detection and localization of license plates from continuously received surveillance video streams.

3. Evaluation of the proposed GenLS approach using a state-of-the-art benchmark dataset captured using a single low-cost camera under different weather and recording conditions, in a realistic setting.

The remainder of the case-study is organized as follows; Section 6.1.1 reports related work. Section 6.1.2 presents the proposed GenLS approach in terms of its five modules of operation. Section 6.1.3 presents the evaluation of the proposed system for accuracy, computational overhead and self-organizing capability for contextualization and confidence predictions. The case-study concludes in Section 6.1.4.

6.1.1 Related Work

A substantial body of research is reported for motor traffic regulation in smart cities, that expands across object detection, object tracking, image classification, scene labeling, video captioning and LPD (Wang and Sng, 2015). This case-study mainly focus the related work in LPD due to the direct relevance to the demonstrated use-case of the proposed GenLS approach in this case-study. All research on LPD can be classified into two categories, handcrafted features and deep learning architectures. Most LPD techniques are based on handcrafted representations that explore spatiotemporal features, interaction among low-level and high-level features, environmental conditions and temporal relationships among the identified features (Xie *et al.*, 2018; Ho *et al.*, 2009). However, handcrafted feature representations pose limitations in scalability, generalizability and computation efficiency. Handcrafted discriminators are selected for a particular objectivity of identification, thus, scaling up the recognition to complex and high-level features is infeasible as hand-engineered features do not relate to 'units of patterns' but are rather the result of convenient mathematical operations (Ordóñez and Roggen, 2016). Furthermore, as handcrafted features are extracted from each video frame, detection is computationally expensive for scenes that are highly congested.

More recently, many computer vision problems have been solved using deep learning techniques, particularly using Convolution Neural Networks (CNN). With the state-of-art performance for computer vision related tasks such as object detection and re-identification (Ren *et al.*, 2015), CNN is a natural fit for LPD. In line with this, Montazzolli Silva and Rosito Jung (2018) proposed a CNN based architecture capable of detecting and rectifying multiple distorted license plates in a single image. Kurpiel *et al.* (2017) presented a CNN architecture that detects the availability of a license plate for sub-regions of an image frame, allowing an estimation of the location of the license plate. However, it is important knowing the confidence of detections made by LPD algorithms in order to use them in vehicular surveillance for critical dynamics. A major drawback of existing deep learning based LPD approaches is the unavailability of a mechanism to provide an indication of the confidence for the license plate locations predicted by the algorithms.

In summary, LPD techniques based on handcrafted features can detect license plates and model a localization over the license plate, however, they require prior knowledge for the design of effective features, and are time consuming to extract, thereby impractical to use in real-time license plate detection. In contrast, learned-representations based on deep learning architectures can detect license plates accurately by processing high-dimensional surveillance video streams. With this context, let us now explore the proposed GenLS approach.

6.1.2 Proposed Approach

A high-level overview of the GenLS approach is illustrated in Fig. 6.1. It comprises of five modules: Region Proposals, Feature Extraction Network, Region Proposal Classification Network (RPCN), Object Localization Network (OLN) and the Latent Space Generator (LSG). First is the region proposal module that generates sub regions from input video frames that potentially contain complete or partial salient objects. In the second module, a CNN architecture extracts feature representation from the input frame sub-regions. The third module RPCN is a fully connected neural network that classifies the existence of objects within the input frame sub-regions. The fourth module OLN is a secondary fully connected neural network layer that estimates a bounding box to localize the objects. The RPCN and OLN modules share the fundamental CNN architecture of feature extraction module as end-to-end networks. The fifth module is the LSG that provides a confidence indication of the proposed bounding box coordination of the detect objects and a further contextualization of all detected salient objects. In the following sub-sections, each module is further delineated in terms of the technicalities of the demonstrative use-case of LPD.

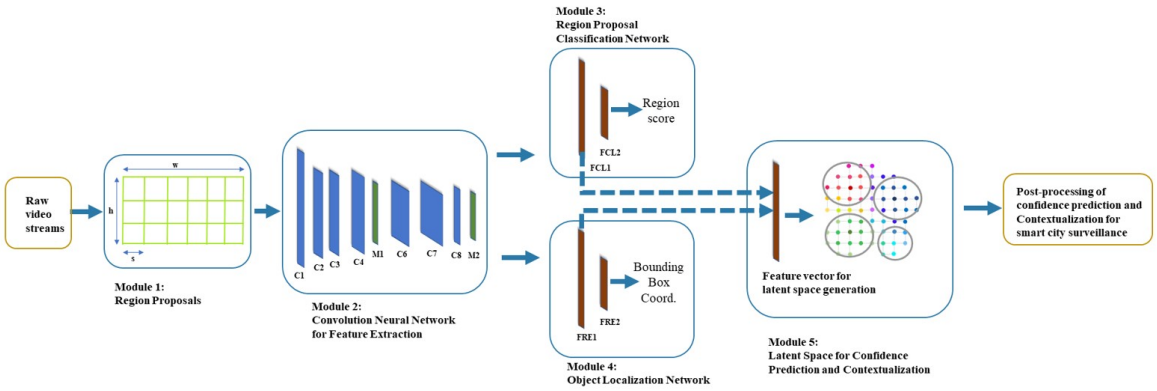


Fig. 6.1 Overview of the GenLS approach

Region Proposals

The raw RGB video data stream from surveillance cameras is input to the proposed GenLS approach. At first, the input images are divided into overlapping sub-regions using a sliding window approach, where the width and height of sub-regions are as w and h, and the stride of the sliding window is s. The dimensions of the sub-region should be selected in a manner such that the sub-region should be able to contain an entire salient object within its neighborhood, but not more than one license plate in a single sub-region. This will enable the CNN feature extractor to learn a feature representation of license plates to concisely distinguish with respect to the vehicle body, general text written in vehicle bodies and background. In selecting the sliding window stride (s), the horizontal stride should be smaller than $w/2$, and vertical stride should be smaller than $h/2$, enabling the sub-regions to create an overlapping. Further, the strides should be large enough, not to provide an extensive number of sub-regions driving the feature extraction computationally complex.

Feature Extraction Network

CNNs were inspired by biological processes resembling the organization of biological visual cortex. The connectivity of the neurons in the convolution layers are designed in a manner similar to animal vision system such that an individual cortical neuron responds to stimuli only in a confined region of the input frame, i.e., the receptive field. In processing images, the convolution layers can preserve the spatial relationship within the input images by learning feature representations using filters, whose values are learned during the training process.

In the proposed GenLS approach, we utilize a CNN architecture specifically designed to extract a feature representation from the input image sub-regions. The CNN architecture and its hyper-parameters were selected using a grid search optimization technique. This grid search optimization technique is a further novelty we report in this work, which leads to the development of a fine-tuned CNN architecture. Results of the grid search optimization are shown in Table 6.1 and illustrated in Fig. 6.1, where [CV] refers to Convolution operation, [MP] refers to Max Pooling operation, [H] are for Hidden Nodes, [F] is for the Number of filters, [K] is the Filter size, and [S] refers to the Strides. The ReLU activation function was utilized for convolution layers as it is less computationally expensive than tanh or sigmoid. In the design of CNN architecture, we deliberately kept it to the simplest form without trading off performance in order to optimize the computation time for real-time processing of surveillance streams.

Table 6.1 Deep Learning Architecture

Layer ID	Layer Type	Input Tensor	Hyperparameters	Output Tensor
C1	CV	180 x 120 x 3	F: 8; K: 5 x 5; S: 3	60 x 40 x 8
C2	CV	60 x 40 x 8	F: 12; K: 3 x 3; S: 3	20 x 14 x 12
C3	CV	20 x 14 x 12	F: 12; K: 3 x 3; S: 1	20 x 14 x 12
C4	CV	20 x 14 x 12	F: 16; K: 3 x 3; S: 1	20 x 14 x 16
M1	MP	20 x 14 x 16	K: 2 x 2	10 x 7 x 16
C5	CV	10 x 7 x 16	F: 20; K: 3 x 3; S: 1	10 x 7 x 20
C6	CV	10 x 7 x 20	F: 24; K: 3 x 3; S: 1	10 x 7 x 24
C7	CV	10 x 7 x 24	F: 4; K: 3 x 3; S: 1	10 x 7 x 4
M2	MP	10 x 7 x 4	K: 2 x 2	5 x 3 x 4
FCL1	FC	60	H: 32	32
FCL2	FC	32	H: 1	1
FRE1	FC	60	H: 32	32
FRE2	FC	32	H: 4	4

Region Proposal Classification and Integration

A fully connected neural network layer was designed on top of the convolution layers in order to classify the existence of a license plate in input image sub-regions. The network architecture is shown in Table I named as FCL1 and FCL2 layers, and illustrated in Fig. 6.1. A linear output function in the range $[0,1]$ is produced by the fully connected neural network, where the output is 1 if the license plate is completely contained within the sub-region, otherwise the output smoothly decrements to 0 until the license plate is completely out of the sub-region. In order to provide a formal definition of the linear output (scoring) function, we employ the geometric function defined by Kurpiel *et al.* (2017).

Here we consider both horizontal (x-axis) and vertical (y-axis) axes of the image sub-region separately. First, we derive the translation in x-axis as Δx (the pixel distance in x-axis between the centre of the license plate to the centre of sub-region) by keeping the translation in y-axis centered. The scoring function $f_x(\Delta x)$ can be defined as Equation 6.1:

$$f_x(\Delta x) = \begin{cases} 1, & \text{if } \Delta x \leq 0.25w \\ \frac{0.5w - |\Delta x|}{0.25w}, & \text{if } 0.25w < \Delta x \leq 0.5w \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

Similarly, $f_y(\Delta y)$ can be calculated by keeping the translation in x-axis centered. The combined function $f(\Delta x, \Delta y)$ can be calculated using the product as defined in Equation 6.2.

$$f(\Delta x, \Delta y) = f_x(\Delta x) \cdot f_y(\Delta y) \tag{6.2}$$

Based on the scoring function defined above, input image sub-regions are scored and used for training of the CNN feature extractor and fully connected layers. Once the sub-regions are scored, results should be integrated to offset the sub-region to completely comprehend the detected license plate. Thus, the sub-regions where scores above a pre-defined threshold and are local maxima, are considered as candidate sub-regions for license plates. Then we select its neighboring sub-regions for both horizontal and vertical direction. First, select the neighbor with the highest score from its left and right neighbors. Second, select the highest score from its neighboring sub-region above and below. Then offset the sub-region to horizontal and vertical direction based on Equation 6.3, where S_l and S_r are scores of left and right neighboring sub-regions respectively.

$$offset_x = \begin{cases} -\frac{S_l}{4} \cdot w, & \text{if } S_l > S_r \\ +\frac{S_r}{4} \cdot w, & \text{otherwise} \end{cases} \tag{6.3}$$

Similarly, the offset in vertical directions is to be calculated, considering the neighbors above and below using Equation 6.3. After calculating the offset, the sub-region is shifted according to the offset, and be used for localization.

License Plate Localization

A fully connected neural network layer on top of the pre-designed convolution layers localizes the license plates in the selected sub-regions by the RPCN. A tightened bounding box coordination is produced by the fully connected layers providing the width, height, x-coordination and y-coordination of the license plate.

In order to implement the License Plate Localization Network (OLN), we update the deep learning architecture by removing the two fully connected layers of the classification network and appending two new fully connected layers. i.e., Shown in Table I, named as FRE1 and FRE2 layers. The model is compiled using a custom loss (L) function using the Euclidean distance as defined in Equation 6.4, where N is the training sample size, $\hat{y}_i$ the prediction output for i^{th} input sample and y_i is the i^{th} input sample.

$$L = \sqrt{\sum_{i=0}^{N} (\hat{y}_i - y_i)^2} \tag{6.4}$$

Latent Space Generator

The LSG is the final module, which generates a latent space of all salient objects. This latent space provides contextualization of all detected objects as well as a confidence metric for

each. As mentioned prior, a latent space encapsulates the interplay across the feature set and all input vectors on to a low dimensional topology, and can accommodate varied forms of input features such as raw data, pre-processed data as well as intermediate outcomes from other machine learning pipelines. The LSG module receives the final feature representation layer of the RPCN and OLN as inputs. The latent space is generated using the RTGSOM algorithm, which has the capability to represent non-stationary input stimuli from natural environment in a latent representation space. This learned abstraction of the LSG module provides a contextualization of all salient objects and distinguishes different types of license plates (salient objects) as well as detection under varied weather and recording conditions. Each segment of the LSG output map was analyzed for accuracy scores of detected license plates in each segment to provide an estimation of the confidence.

The learning outcomes of the self-structuring process in RTGSOM are two-fold: i) the knowledge embodied in the weight vectors of the winning neuron, which represents well-defined features in the data space, and ii) the proportionate learning outcome, knowledge embodied in weight vectors of the winner's neighborhood neurons. The two learning outcomes are then combined into a single structure to produce a generalized representation. We have used fuzzy aggregation and a proximity matrix based on the closeness property (De Silva and Alahakoon, 2010) as means of generalizing learning outcomes into a single representation which has the same basic structure of a neuron. The generalization algorithm is detailed in Algorithm 3, where hit threshold (HT) is given by Equation 6.5 and γ is the hit threshold fraction. The generalized representation of RTGSOM can be used to recognize clusters, patterns and groupings of the data space. It is important to note that, the GSOM algorithm is initialized with random initialization for the self-organization to keep away from local minima and produce better quality clusters (Amarasiri *et al.*, 2005). In despite that the representation of the GSOM depends on the initial random initialization of the GSOM nodes, the generalization of the GSOM outcomes provide means to generate clusters that are independent of the initial weight initialization. Thereby, the final clusters/groupings depend only on the data space.

$$HT = length(X) \times \gamma \tag{6.5}$$

Model Training

To optimize the learning of RPCN and OLN, we used a variant of gradient descent with momentum, RMSprop optimizer (Ruder, 2016). RMSprop computes a dimension-wise learning rate by an exponentially decaying average of squared gradients, adapting the rate of

Algorithm 3: Generalized RTGSOM Algorithm

Data: X

Result: Generalized Representation (ϕ)

1 Identify hit neurons with respective to the neighborhood neurons in the RTGSOM using hit threshold (HT) calculated using Equation 6.5 ;

2 Define the generalized map (ϕ_t) by selecting only the hit neurons. ;

3 Calculate the proximity matrix, S, where S_{km} contains the proximity of the m^{th} neuron of the k^{th} neighbourhood to the corresponding hit neuron in ϕ, using Euclidean distance. Here, S_{km} will represent the fuzzy measure (g_α) of the m^{th} neuron of the k^{th} neighbourhood. ;

4 Using the pre-calculated fuzzy measure g_α, update the weights of the neurons in ϕ. ;

5 The neurons in ϕ represent the clusters of the original GSOM. ;

gradients by a function of all previous updates on each dimension. This is widely used for its strong theoretical guarantee of convergence and empirical successes.

The RPCN was trained using Mean Squared Error (MSE) loss function, whereas the OLN was trained using Euclidean loss function. Further, as the number of parameters in the CNN feature extractor is large, we augmented the license plate regions for the OLN using multiple scales and multiple sub-region locations. With these parameters, we trained the RPCN network and the OLN network with an early stopping callback function that will stop training when the monitored metric has stopped improving. While training the OLN, the convolution layers in the feature extractor kept frozen as it has already learned to extract license plate regions sufficiently.

6.1.3 Experiments

The GenLS approach was evaluated using a state-of-the-art benchmark dataset for LPD. The evaluation criteria were the performance of plate detection, plate localization and the self-organizing capability of the LSG module. GenLS was trained on a high-performance computing specification, 36-core CPU 2.3GHz with 128GB memory and dual NVIDIA Quadro of 24 GB GPU units. Evaluation of GenLS was intentionally conducted on a personal computer configuration, an Intel Core i7 CPU (2.6 GHz) with 16 GB memory and GPU of NVIDIA GeForce GTX 950M, in order to ensure that the proposed GenLS approach can be realistically deployed in a practical setting. The algorithms were implemented in Python using TensorFlow framework.

Dataset

Most of the experiments in LPD research are based on static vehicle captures, where the images are taken to contain vehicle license plates. However, for a realistic scenario, the vehicles are mobile and the environmental conditions would be volatile. To ensure such a realistic setting, we selected the state-of-the-art license plate recognition dataset composed by Luvizon *et al.* (2016). The vehicle dataset was captured using a low-cost 5-megapixel CMOS image sensor with frame resolution of 1920×1080, at 30.15 frames per second. Total of 1829 video frames (with 4070 license plates) are available in the dataset, where the videos were grouped into categories based on different weather and recording conditions: [H] high quality, [L] frames affected by severe lighting conditions, [N] frames affected by natural or artificial noise, [B] motion blur, and [R] rainy weather condition. The 5 categories of the dataset are composed of: Cat-1) [H], Cat-2) [L], Cat-3) [N], Cat-4) [N, R], and Cat-5) [L, B]. Each video frame is associated with a ground truth data containing bounding box coordination for each license plate.

Evaluation Metrics

Widely adopted evaluation metrics for LPD research includes object detection measures and Intersection over Union (IoU) measure. Thus, we adopt the three prime measures for object detection: precision (p), recall (r) and F1-Score (F1). Precision is the fraction of correctly detected license plates among all the positive detections, whereas recall is the fraction of correctly detected license plates that have been detected over all the relevant license plates. F1-Score is the harmonic mean of precision and recall.

IoU measures the overlap between the detected license plate region (Q) and the ground truth license plate region (R), where defined in Equation 6.6. In considering correct detection, we use the IoU as a threshold measure, where a detection is only correct if $IoU \geq 0.5$.

$$IoU = \frac{Q \cap R}{Q \cup R} \tag{6.6}$$

Evaluation and Comparison

The proposed GenLS approach is evaluated using five state-of-the-art LPD approaches. First, Stroke Width Transform (SWT) approach where license plate texts are attempted to detect by seeking to find the value of stroke width for each image pixel, proposed by Epshtein *et al.* (2010). Second, SnooperText (Minetto *et al.*, 2014) locates license plate regions using toggle-mapping image segmentation and character classification based on the histogram of gradients (HoG) shape descriptor. Third, Luvizon *et al.* (2016) utilize a

motion detector and a text detector to locate vehicle license plate in image regions containing motion. Fourth, Kurpiel *et al.* (2017) propose a convolution neural network based LPD approach that models a function that produces a potentiality score for each image sub-region to locate the license plate. Fifth, Selmi *et al.* (2017) proposed a CNN based LPD system that utilize extensive pre-processing steps such as contrast maximization, geometric filtering and adaptive thresholding to supplement the CNN feature extraction. Both the CNN based approaches are closer to our study as they have attempted to address similar problems in LPD, which are to detect license plate under diverse weather and recording conditions. The first three approaches are based on handcrafted feature representation based whereas the fourth uses a deep learning approach. The aforementioned four approaches were evaluated with respect to proposed GenLS approach, evaluating accuracy in terms of precision, recall and F1-Score and computational performance in terms of processing time, using 20% of the data in each category (weather and recording conditions). The classification results obtained for each approach is presented in Table 6.2, where the results appear as reported by the respective authors.

Table 6.2 License Plate Classification Results

Data	Epshtein *et al.* (2010)			(Minetto *et al.*, 2014)			Luvizon *et al.* (2016)			Kurpiel *et al.* (2017)			Selmi *et al.* (2017)			GenLS (Ours)		
Set	p	r	F	p	r	F	p	r	F	p	r	F	p	r	F	p	r	F
01	0.76	0.61	0.68	0.81	0.88	0.84	0.96	0.94	0.95	0.86	0.87	0.87	0.84	0.82	0.83	0.92	0.90	0.91
02	0.28	0.23	0.25	0.86	0.81	0.83	0.92	0.84	0.88	0.86	0.82	0.84	0.77	0.73	0.75	0.90	0.90	0.90
03	0.66	0.62	0.64	0.56	0.79	0.66	0.94	0.94	0.94	0.87	0.83	0.85	0.80	0.77	0.78	0.88	0.84	0.86
04	0.79	0.58	0.67	0.44	0.71	0.54	0.94	0.92	0.93	0.89	0.84	0.86	0.79	0.76	0.77	0.88	0.85	0.86
05	0.15	0.15	0.15	0.76	0.72	0.74	0.88	0.82	0.85	0.86	0.78	0.82	0.74	0.73	0.73	0.85	0.83	0.84
Overall	0.44	0.37	0.40	0.73	0.80	0.76	**0.93**	**0.90**	**0.90**	0.87	0.83	0.85	0.79	0.76	0.77	**0.89**	**0.86**	**0.88**

Overall, Luvizon *et al.* (2016) demonstrated the highest accuracy in terms of F-Score by achieving a score of 0.90, whereas the proposed GenLS scores 0.88 having the second best accuracy. From a runtime perspective GenLS performed in 154.4 milliseconds per frame, while Luvizon et al. method took 2320 milliseconds per frame in average using a 2.2 GHz Intel Core i7 with 12GB RAM. GenLS, with the support of GPU and automated feature extraction, achieved 10 times speedup comparatively to Luvizon et al. method. As such, GenLS makes it feasible to be used in real-time practical surveillance applications processing 6 frames per second (FPS). SWT approach by Epshtein *et al.* (2010) and (Minetto *et al.*, 2014) approach showed a low accuracy and weak in terms of adapting to different recording and weather conditions as the accuracies differ substantially depending on each test set category.

With respect to deep learning based approaches, Kurpiel *et al.* (2017) and Selmi *et al.* (2017) showed in par computational complexity in terms of processing time and robust accuracies for each category. However, the proposed GenLS outperforms Kurpiel *et al.* (2017) in terms of accuracy.

A sample of GenLS outcomes is presented in Fig. 6.2 for a qualitative analysis. The examples illustrate correctly detected license plates under diverse weather and recording conditions as well, incorrect and/or missed detections. Especially, Cat.1[B], Cat.3[B-C], Cat.4[B] and Cat.5[B], shows vehicles that contain advertisement texts on the vehicle body, which have correctly ignored by the detection system, proving the robustness of the proposed detection system. However, the outcomes: Cat.2[D] and Cat.4[D] contained two false negative detections that the system had missed to detect due to dark foreground of the license plates. A closer analysis confirms the difficulty even for a human analyst to recognize the presence of a license plate in both occurrences. As false positives, Cat.1[C-D], Cat.3[D], and Cat.5[D] contained a vehicle each, with a large amount of advertisement texts on the body, that caused incorrect detections.

Ablation Study

We performed several experiments on the proposed feature extraction network architecture in order to identify its best arrangement for LPD with respect to performance versus computation overhead trade-off. Multiple configurations of the CNN network design were experimented with two sub-region dimensions to be used as input region proposals. The results of the performance and runtime of the ablation study are presented in Table 6.3. We demonstrated CNN architectures with four configurations, i.e., 5-8 convolutional layers (CL). It can be seen that when the number of convolutional layers increases, the performance (F-Score) increases gradually to a limit of 0.88. Both network configurations with 7 layers and 8

Table 6.3 Ablation Analysis

Network Design	F-Score	Runtime (ms)
CL x 5	0.84	153.7
CL x 6	0.87	155.2
CL x 7	**0.88**	154.4
CL x 8	0.88	154.6

layers showed the similar highest F-Score. In contrast, the computation time is stable across different levels of model complexity. This indicates the CNN model with 7 convolutional layers provides sufficient ability to capture the features from the given surveillance context. Thereby, we selected the optimal configuration with 7 convolution layers for the proposed feature extraction network in order to maximize the performance while optimizing the computation overhead. Further, we evaluated the selected network configuration with two configurations for sub-region dimensions, i.e., 90×60 and 180×120. The sub-region dimension 180×120 resulted in the highest performance (F-Score = 0.88), while the F-Score for dimension 90×60 was 0.82. The computation time reached to 271 ms when the dimension is set to 90×60. Thereby, considering both the F-Score and runtime, we specified the sub-region dimension of 180×120 for the region proposal network.

Confidence Prediction Evaluation

The latent space of the LSG module was generated using the RTGSOM algorithm. We utilized the final feature representation layers of the RPCN and OLN, i.e., layer FCL1 and FRE1, with RTGSOM parameters: SF = 0.83, 100 learning iterations, and 50 smoothing iterations. Then the generalized representation was conducted using Algorithm 3 to cluster the data space. The generalized representation and the GSOM is illustrated in Fig. 6.3.

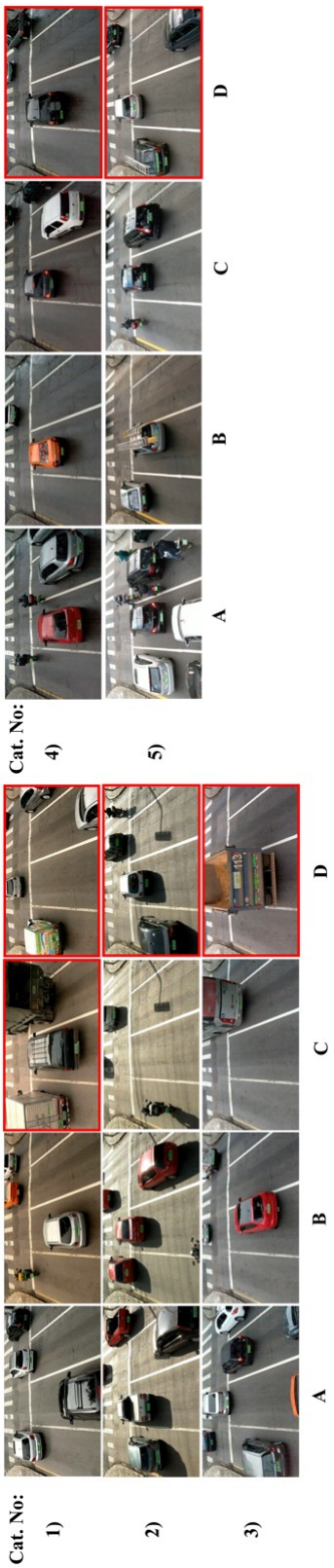


Fig. 6.2 A sample of GenLS outcomes in different recording and weather conditions. Cat. No. identifies the video frames category and red color outline box suggests incorrect and/or missed detections.

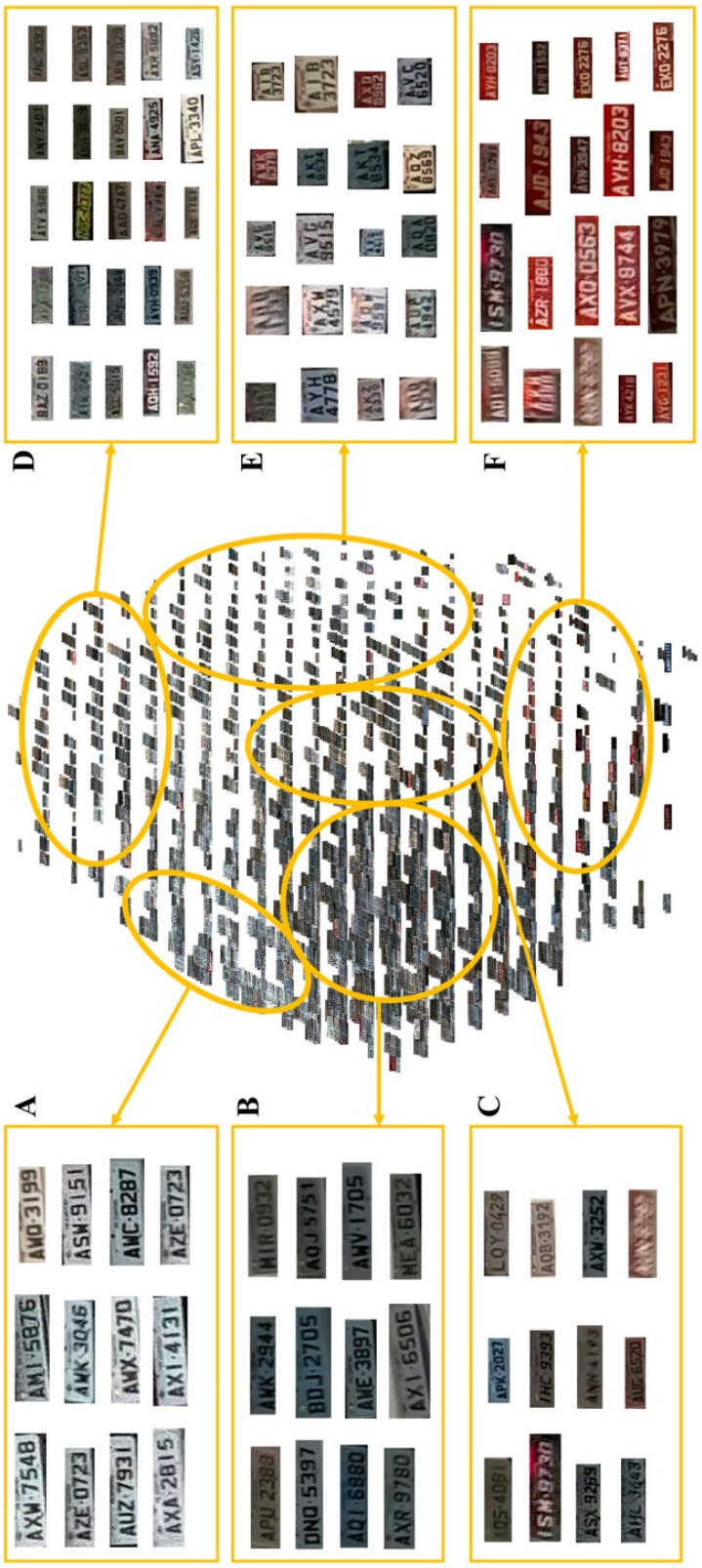


Fig. 6.3 Generalized cluster representation of the LSG.

A cluster analysis resulted in identifying six clusters (clusters A-F), as illustrated in Fig. 6.3. An overview of the self-organized map shows that when scanning through the map from west (left) to east (right), the quality (resolution) of the license plates deteriorate. Scanning the northeast and southeast regions of the map, mostly square-shaped license plates are contained where southeast focused on the license plates that are of motorcycles. Zooming into clusters identified by the generalization algorithm, cluster A composed of high-quality rectangular shaped license plates that are in clear white background. Cluster B composed of license plates that are in low lighting conditions, under rainy weather, and/or with a dark foreground. Cluster C composed the license plates contain a blur effect and a dark foreground. Cluster D contains the license plates that are in clear white background similar to cluster A, but with low-resolution. Cluster E contains square shaped license plates both from vehicles and motorcycles. The license plates that are from different color, e.g., blue, red, are grouped into cluster F. Additionally, the license plates that are in extremely poor quality and/or rare license plate designs are clustered outwards of the RTGSOM, as illustrated in Fig. 6.4.

Fig. 6.4 Outlier license plates identified from the LSG.

The identified clusters were further analyzed to verify the confidence of the predictions. Thus, we calculated the accuracy using F-Score for each of the clusters as well as outliers separately to provide a Confidence Identification Tag (CIT) for license plate detections. The results are presented in Table 6.4, where cluster A, B and D can be identified as the license plate groups with high confidence to be accurate. Clusters C, E, and F have a medium confidence as they achieve accuracies in the range of 0.84-0.86, which is below than the overall F-Score measure. The outliers have an average F-Score of 0.77 where we tag them as having a low confidence to be truly accurate.

In operation, the GenLS approach used license plate detections and confidence indication, and the example outcomes used in Fig. 6.2 were tagged with the respective CIT. The outcomes: Cat.1[C-D], Cat.3[D], and Cat.5[D] were tagged with a CIT = LOW as there were at least one detection that did not fall under any cluster, i.e., outliers. Thus, the aforementioned outcomes

Table 6.4 Cluster Accuracy and Confidence Identification

Cluster	Accuracy (F-Score)	Conf. Id. Tag (CIT)
A	0.94	HIGH
B	0.91	HIGH
C	0.84	MEDIUM
D	0.88	HIGH
E	0.86	MEDIUM
F	0.86	MEDIUM
Outliers	0.77	LOW

were analyzed by the GenLS approach to be incorrect. Further, outcomes Cat.2[A,B,D] and Cat.4[C-D] were grouped in cluster C, thus, was tagged with a CIT = MEDIUM, as the foreground of the license plates detected in the video frames were dark. Among these five video frames, Cat.2[D] and Cat.4[D] had a single vehicle each as false positive. This analysis of misclassifications by the proposed LPD in combination with the LSG, justifies the importance of confidence indication to provide an accountable and failsafe prediction outcome. Thus, in practice, the license plate detections are provided both with a predicted bounding box of the license plate and a CIT alongside proposing the confidence of the prediction. In the instances when the predictions are at low confidence, a human operator involvement can be requested.

To further validate the generative latent space developed by the RTGSOM algorithm, we utilized the T-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) for the same task. Similar to RTGSOM, we extracted a feature representation using the FRE1 layer of OLN and FCL1 layer of RPCN, to develop the t-SNE visualization, as illustrated in Fig. 6.5. The output generated by t-SNE neither provides a contextualization of the salient objects nor forms semantically meaningful clusters that can distinguish between different classes of license plates. Similar to t-SNE, PCA is also applied but shows no success in forming semantically meaningful clusters for confidence identification. This is further justification of the suitability of the GSOM algorithm for the GLS module in the GenLS approach.

6.1.4 Discussion

In this case-study, we proposed a new generative self-structuring and deep learning based approach, Generative Latent Space (GenLS), for smart city surveillance that generate continuous video data streams such as motor traffic regulation, crowd behavior classification, commodities movement, facilities provision, productivity improvements, and anomalous

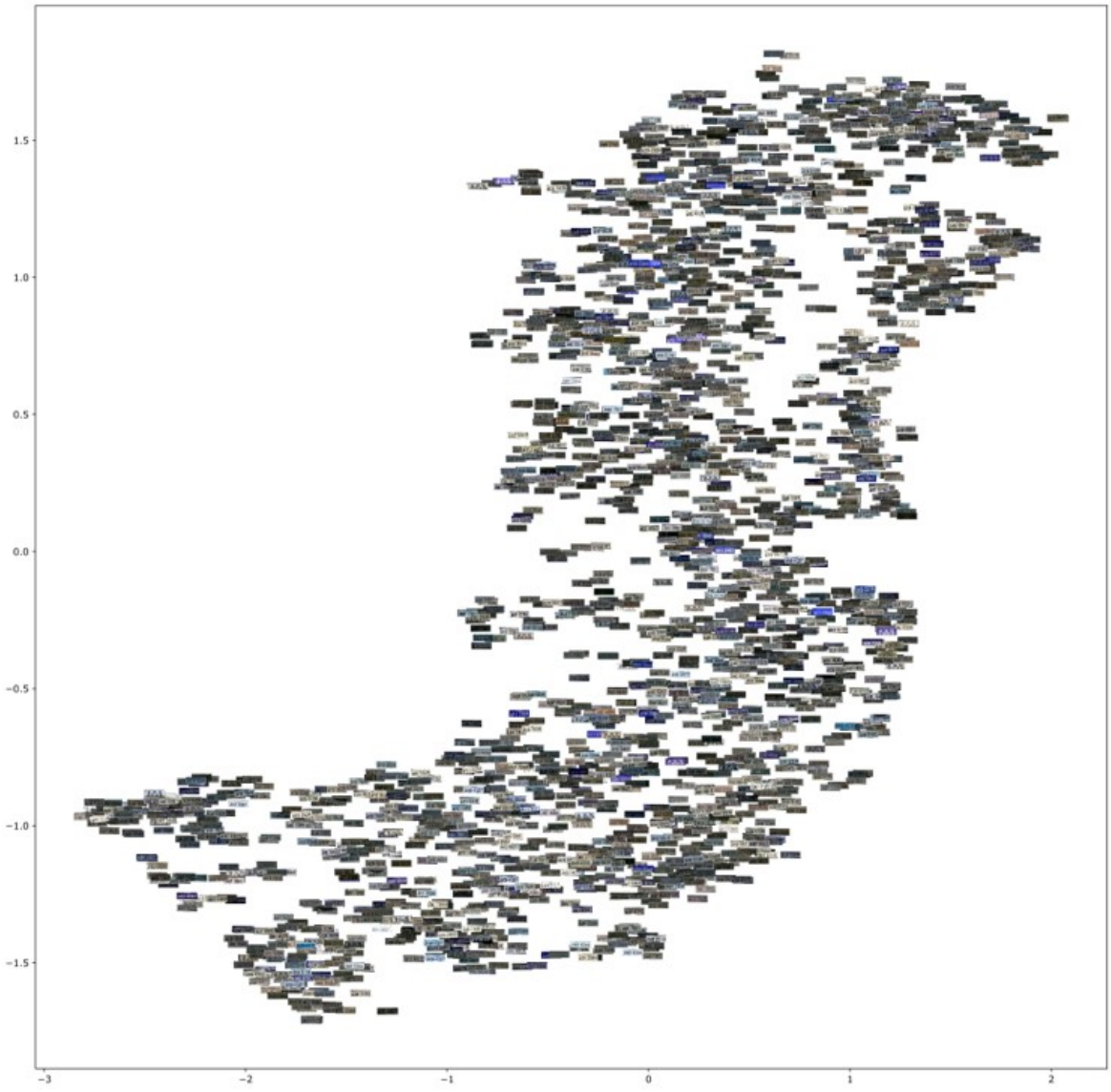


Fig. 6.5 t-SNE visualization of the detected license plates.

event detection. In practice, motor traffic regulation is more prevalent than other surveillance scenarios, thus we selected the technically challenging use-case of License Plate Detection (LPD) within the motor traffic regulation scenario to demonstrate the application of GenLS to smart city surveillance.

The proposed GenLS approach overcomes hitherto unaddressed challenges of sub-optimality, latency, predictive accuracy, and contextualization of all detected salient objects for further decision-making. In practice, GenLS can resolve complexities caused by diverse weather and recording conditions due to the use of low-cost CCTV cameras, and the computational cost of handling high-dimensional video surveillance data streams in real-time. Results from experiments conducted on a state-of-the-art benchmark vehicle surveillance dataset demonstrate the accuracy, robustness, low computations overhead and validity of the outcomes, confirming its wide applicability in smart city settings. The proposed GenLS approach can be extended for further applications in smart city domain such as intelligent video surveillance, content based medical image retrieval in Digital-Health applications, intelligent patient profiling in clinical settings, and even topics/concepts formulation from social media and online discussion forum data.

As future work, we intend to expand GenLS in terms of accuracy and reliability in order to be viable for deployment across all smart city surveillance scenarios. Furthermore, we intend to extend the post-processing potential of GenLS with optical character recognition (OCR) and cloud services to provide real-time insights for comprehensive motor traffic regulation across, safety, efficacy, and sustainability measures.

In conclusion, this case-study presented a successful implementation of the proposed MSKRF concept framework for a real-world situation, when an environment (smart city) is being represented in a digital form by using advanced data sensing technologies, generate large, multi-modal, multi-source, dense, high frequent and non-stationary datasets. In particular, RTGSOM algorithm conceptualised in Section 4.2 was demonstrated to be successful and robust in this practical case-study. Thereby, new self-structuring AI capabilities brought in to solve this particular problem was significant in uplifting the predictive modelling capabilities of the deep learning techniques, as well provided a contextualization of the digital environment leading to further decision-making, thereby making the new self-structuring AI the missing piece of a successful intelligent surveillance system.

6.2 Digital-Health: Stroke Patient Profiles and Trajectories

Digital-Health is the use of technological advancements to improve individual's health and wellness. It is an emerging area of research (and applications) that is likely to transform the biomedical world. From wearable devices to ingestible sensors, from mobile health apps to AI, from robotic carers to electronic records, all constituents of the domain of digital health enables healthcare environment to be represented in a digital ecosystem.

The objectives of digital health are diverse, including prevent diseases, help to monitor patients and manage chronic conditions, lower the cost of healthcare provision, and make medicine more tailored to individual needs. This has the potential to benefit both patients and healthcare providers. From patients perspective, gathering data from multitude of wearable devices, from activity level to blood pressure, compare the levels with the mass and derive insights that can transform into notifications, alerts and predictions will allow individual patients to improve their lifestyle and maintain a proper health (Best, 2019). From physician's perspective, an accumulation of granular details of patients will provide a thorough understanding of the patient, which will be enriched using AI based predictions and recommendations.

In this section, we present our second case-study within the scope of Digital Health focusing on neuroscience and mental health. The presented case-study analyzes clinical outcomes of a stroke survivor patient cohort (n=219) during three timepoints: 7-days, 3-months and 12-months post-stroke with the aim to identify distinct clinical profiles and recovery trajectories. Clinical outcomes were measured across physical, cognitive and mood domains, disability, stroke impact and work and social adjustment. The analysis of the various stroke survivors, generation of distinct clinical profiles and recovery trajectories were conducted using the new self-structuring based AI techniques developed in this thesis. Primarily, we use the fundamentals of MSKRF conceptual framework and its materialization, RTGSOM algorithm.

6.2.1 Introduction

Stroke is known as the leading cause of adult disability and the third most common cause of adult death in the industrialized world (Williams *et al.*, 1999). Due to the complexity of physiological, psychological and social burden associated with stroke, it is pertinent to measure and understand the holistic impact post-stroke (Doyle, 2002). Thus, separate measures and tests have been developed to measure multitude of aspects of stroke patients

and survivors such as cognitive, physical and social. Among these, one of the widely recognized measures is the National Institutes of Health Stroke Scale (NIHSS), which is a systematic assessment tool designed quantify stroke severity based on weighted evaluation findings (Ortiz and L. Sacco, 2014). NIHSS is employed to assess neurological impairment across consciousness, movement and language (Kasner, 2006). Assessment scores for each item are then tailored to derive an overall score of stroke severity (range 0-42). According to Spilker *et al.* (1997), the overall NIHSS score is interpreted to measure the severity of stroke as: $0-5$ mild; >6 moderate to severe. NIHSS is primarily used in the acute phase post-stroke and is recognized as a valid and reliable screening method that is widely used in clinical practices (Hinkle, 2014). However, it has been criticized that it does not measure all domains of function and thus the impact of stroke on survivors with impairment and associated consequences of stroke in particular domains may be unaccounted (Martin-Schild *et al.*, 2011).

This is particularly important for mild stroke survivors, as it is stated that, mild stroke survivors are often less investigated due to the assumption that they are expected to regain their premorbid functionality with less or no intervention (DeGraba *et al.*, 1999). The stroke patients and survivors that have been classified as 'mild' does often argue that the given classification does not correspond with their daily experiences, as some of them tend to report depression, difficulties in advanced physical and social activities indicating the diminished quality of life (Duncan *et al.*, 1997; Carlsson *et al.*, 2003). Further studies emphasize that attention needs to be prioritized for the needs beyond functional outcomes and the need to address hidden dysfunctions of mild stroke survivors (Green and King, 2010; Carlsson *et al.*, 2004). This leads to an opportunity for a granular investigation of the latent impairments associated with mild stroke survivors.

In this case-study, we aim to systematically investigate various profiles of mild stroke survivors based on the different tests measuring their stroke impairment. One of the key challenges in distinguishing different profiles is combining all stroke impairment measures together in order to uncover patterns. Therefore, we utilize latent representation from MSKRF to automatically discover variants of impairments in survivors classified as mild by NIHSS. We use the START dataset (Carey *et al.*, 2015) which consists of multiple test scores of stroke survivors at three time points (3-7 day; 90 days; 365-days post-stroke) in their stroke journey.

6.2.2 Study Design

Participants

The START study (STroke imAging pRevention and Treatment) is a longitudinal cohort study of 219 stroke survivors where they were investigated at baseline, 24h, 3-7 days, 90 days and 360 days post-stroke (Carey *et al.*, 2015). The stroke survivors were assessed for their functional outcomes, depression, physical activity, cognitive abilities and lifestyle at the aforementioned time points. The inclusion criterion of our study is to include only mild stroke survivors with complete information. There were 107 mild stroke survivors in the START dataset according to the baseline NIHSS, however, only 73 had complete information over a one-year time period. This yielded a study cohort of 73 stroke survivors for our study.

Post-stroke impairment measurements

START dataset consists of outcomes of multiple tests that are used to measure impairment in different aspects following stroke. We outline a description of each of the tests included in the analysis with the corresponding measurement criteria as follows.

- The **National Institute of Health Stroke Scale (NIHSS)** is a 15-item assessment used to evaluate the severity of stroke neurological deficits, including consciousness, language, neglect, visual-field loss, extraocular movement, motor strength, ataxia, dysarthria, and sensory loss (Brott *et al.*, 1989). The potential scores of NIHSS range from 0 to 42; higher scores represent more severe stroke deficits.

- The **Modified Rankin Scale (mRS)** is a widely used functional outcome measure in stroke. The mRS assesses an individual's degree of disability or dependence in the daily activities after stroke through a structured interview (Wilson *et al.*, 2002). Six levels are defined in the mRS scoring from 0 – 6: 0 for no symptoms at all, 5 for total dependence, and 6 for dead.

- The **Montgomery-Asberg Depression Rating Scale (MADRS)** measures a person's depressive symptoms (Montgomery and Åsberg, 1979). Using a structured interview, the MADRS was examined to have excellent inter-rater reliability. The MADRS has 10 items rated on a six-point Likert scale (0 – 6). The total score of MADRS is 60, a higher score indicates more severe depressive symptoms. A score of 18 or greater is suggestive of major depression (Williams and Kobak, 2008).

- The **Montreal Cognitive Assessment (MoCA)** is a sensitive and widely used screening tool to detect post-stroke vascular cognitive impairment (Nasreddine *et al.*, 2005).

The MoCA assess diverse cognitive domains, including visuospatial and executive functions, attention, memory, language, conceptual thinking, and orientation. The total score of MoCA is 30, with higher scores indicate better cognitive function. An extra 1 point is added to the total score if a person has less than 12 years of formal education.

- The **Stroke Impact Scale (SIS)** is a disease-specific, self-report questionnaire that evaluates disability and health-related quality of life after stroke (Duncan *et al.*, 1999). The SIS assesses the subjective impact of stroke in eight domains: strength, memory and thinking, emotion, communication, (instrumental) activities of daily living (ADL/IADL), mobility, hand function and participation. All items within each domain are scored on a 1 to 5-point Likert scale. The total scores of each domain range from 0 to 100. Higher item scores indicate a lower level of difficulty experienced. The reliability and validity of the SIS are excellent (Duncan *et al.*, 2003).

- The **Rapid Assessment of Physical Activity (RAPA)** is an easy-to-use, valid outcome measure of assessing levels of physical activity among adults (Topolski *et al.*, 2006). The RAPA has 9 items, including 7 items for aerobic activities and 2 items for strength training and flexibility.

- The **Work and Social Adjustment Scale (WSAS)** is a 5-item self-report scale of functional impairment resulting from a health problem (Mundt *et al.*, 2002). The five WSAS items determine the following impairment dimensions: (1) work; (2) home management; (3) social leisure activities; (4) private leisure activities; and (5) relationships with others. Each item is rated on a 0 to 8 scale: 0 indicates no impairment at all and 8 indicates very severe impairment. The maximum total score of the WSAS is 40. Good reliability and validity have been examined for the WSAS (Mundt *et al.*, 2002; Zahra *et al.*, 2014).

6.2.3 Analysis Approach

We propose an unsupervised machine learning based AI framework to identify latent behavioral patterns of patients who are undergoing stroke rehabilitation. The high-level architecture of the framework consists of a data processing module, a self-structuring based AI module and an insights module to analyse the outcomes based on visualizations and reporting generated from the analysis. We have designed this framework to accommodate the selected participants' data from the START longitudinal study. The Fig. 6.6 illustrates the proposed high-level framework.

Data Filtration

As the first step of data processing, we filter the START patient data sample based on the completeness of data. We evaluate the missing data for demographic details and selected test scores that we use for the analysis. The selected tests are NIHSS, MOCA, MADRS, RAPA, WSAS, mRS and SIS. As per the inclusion criterion, participants' records that do not contain data for these tests are removed from the dataset sample. This resulted in a study sample of 73 mild stroke survivors.

Data Standardization

The filtered dataset was then processed to standardize across all the tests. The aim of the standardization is to make the data internally consistent across the tests in order to make sure each test is comparable with each other. Therefore, an impairment index was introduced by scaling the selected test scores to the range from zero to one, where zero indicates no impairment and one indicates severe impairment. The standardized data is then transformed to construct a machine-understandable format for AI-based analytics. Thus, first, we used the technique binning to group continuous-valued columns into given intervals. The aim of this transformation technique is to reduce the effects of minor observation errors.

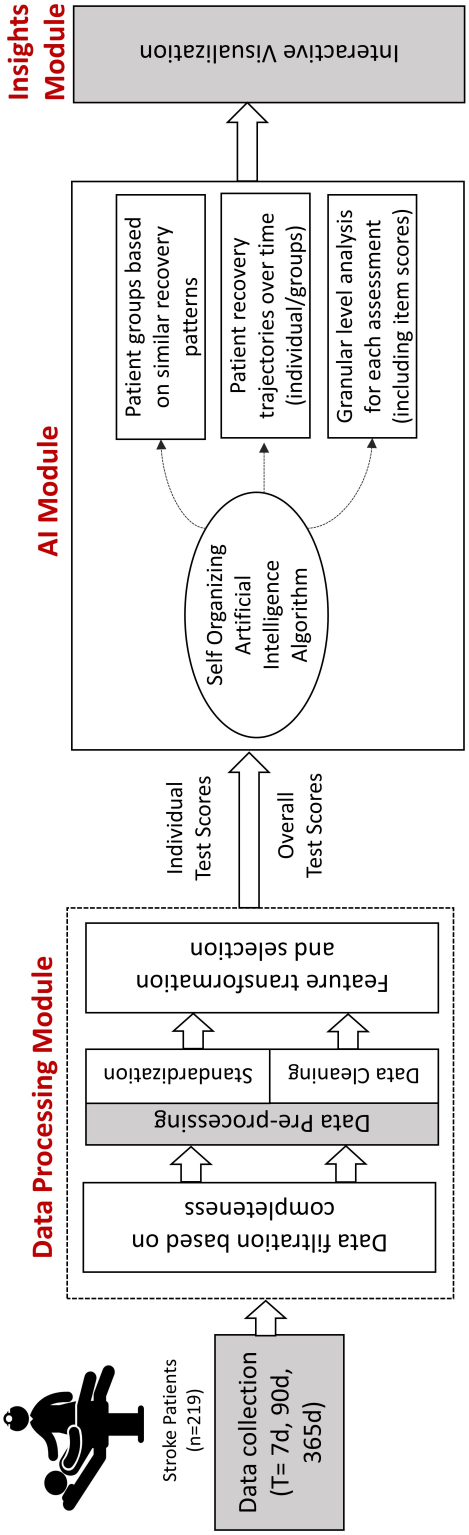


Fig. 6.6 The high-level architecture of the analysis framework.

Latent representation module to detect variants of impairment

The AI module of the proposed framework comprises of an unsupervised self-structuring algorithm which was developed to generate a latent representation of the natural environment in Chapter 4. Using RTGSOM algorithm, we conduct unsupervised clustering of multiple dimensions of participants' clinical data. One of the key reasons to incorporate RTGSOM was to discover latent representations of stroke impairments based on multiple stroke assessment measures, which are otherwise challenging to analyse manually. The ability of the RTGSOM algorithm to uncover patterns among non-stationary data along a temporal trajectory without prior training or supervision makes it possible to detect impairment across multiple domains that were not assessed in NIHSS screening. Thereby, in this case-study, we utilize the RTGSOM to derive a latent representation of mild stroke survivors at each time point.

Insights Module

In the third connected module of the proposed framework, we provide the capability to automatically detect important clusters. Once the RTGSOM latent representations are created for each time point, the representations are used to derive important patterns automatically. This functionality is built into the insight generation module of the framework. This module will first iterate through participant data in each node to examine clear patterns and mark the regions which have shown significant variations from the study sample at each time point.

This insight generation capacity is built into an interactive visualization platform that allows data pattern visualization from multiple dimensions such as cognitive ability, depression, physical activity and lifestyle adjustment. This provides a holistic view of the stroke survivors' experience by showing different variants of impairments. The visualization platform can be accessed by therapists or clinicians to support enhanced decision-making. Sample screenshots from the visualization tool are presented in the Appendix A.2.

6.2.4 Results

Analysis of patient profiles based on the NIH Stroke Scale

The NIHSS consists of 16 items that focus on different neurological aspects such as level of consciousness, horizontal eye movement, visual field test, facial palsy, motor arm, motor leg, sensory, speech, language and attention. Scores for each question item are aggregated to form the final NIHSS score which is used for the stroke severity classification. However, this aggregated score does not capture which aspects are more impaired or recovered. Therefore, we used the AI framework on the baseline NIHSS item-scores to detect different groupings

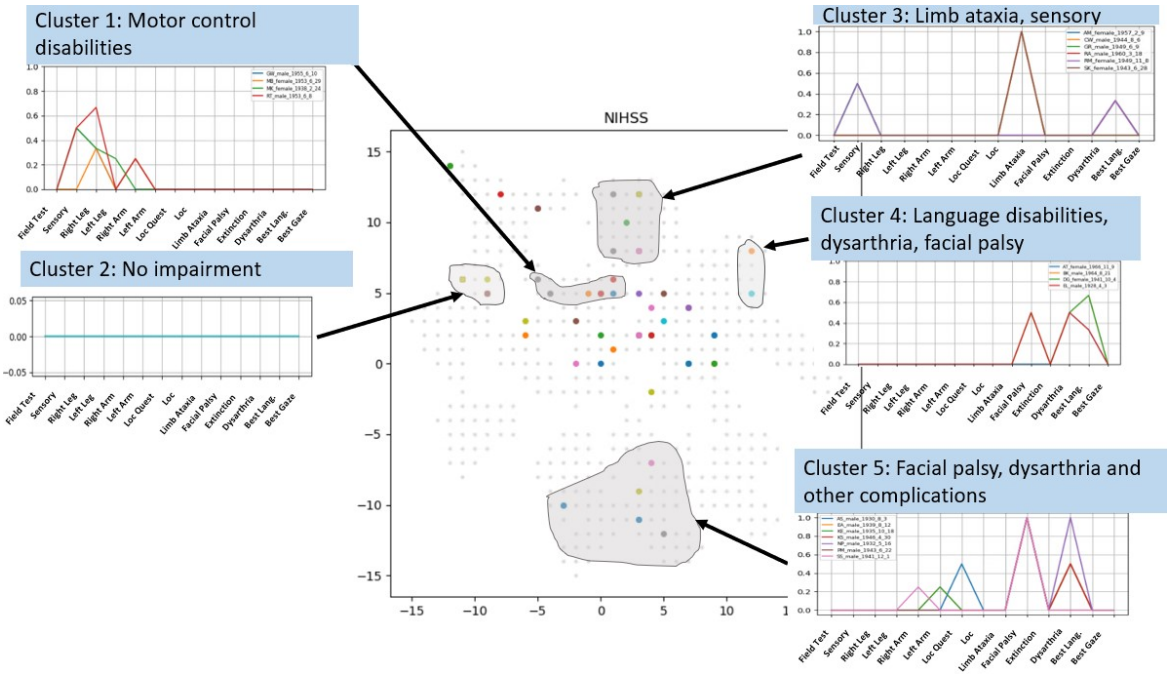


Fig. 6.7 Distinct clusters of mild stroke patients based on NIHSS.

(annotated regions) which can be derived from NIHSS itself. Only the significant clusters are illustrated for clarity. Participants in non-annotated regions do not exhibit significant patterns that can be differentiated. Fig. 6.7 shows the different clusters found in mild stroke survivors based on the aspects of NIHSS.

It was found out that, although all the participants are categorized as mild stroke, there are different variations of impairments present among the cohort. The latent representation showed a clear separation between participants having motor control disabilities, language disabilities, facial palsy, and other complications. Participants in cluster 1 showcased motor control impairment mostly in the right leg and right arm as well as in sensory functions. Participants in cluster 3 showed more than average of the impairment index for sensory function and limb ataxia. Participants in cluster 4 have reported impairment mostly related to language and speech as they have shown higher impairment index for dysarthria, extinction, moderate aphasia. They have also reported a moderate level of facial palsy. Similar characteristics can be observed in cluster 5 as well, however, all participants in cluster 5 have reported severe facial palsy while the majority have reported impairment for dysarthria. It can be asserted that these survivors suffer from language disabilities due to complications associated with facial palsy. Compared to cluster 4, cluster 5 group also show mild disability with both left and right arms and moderate impairment in consciousness questions. Among the participants in the cohort, cluster 5 stroke survivors showed a majority of complications while cluster

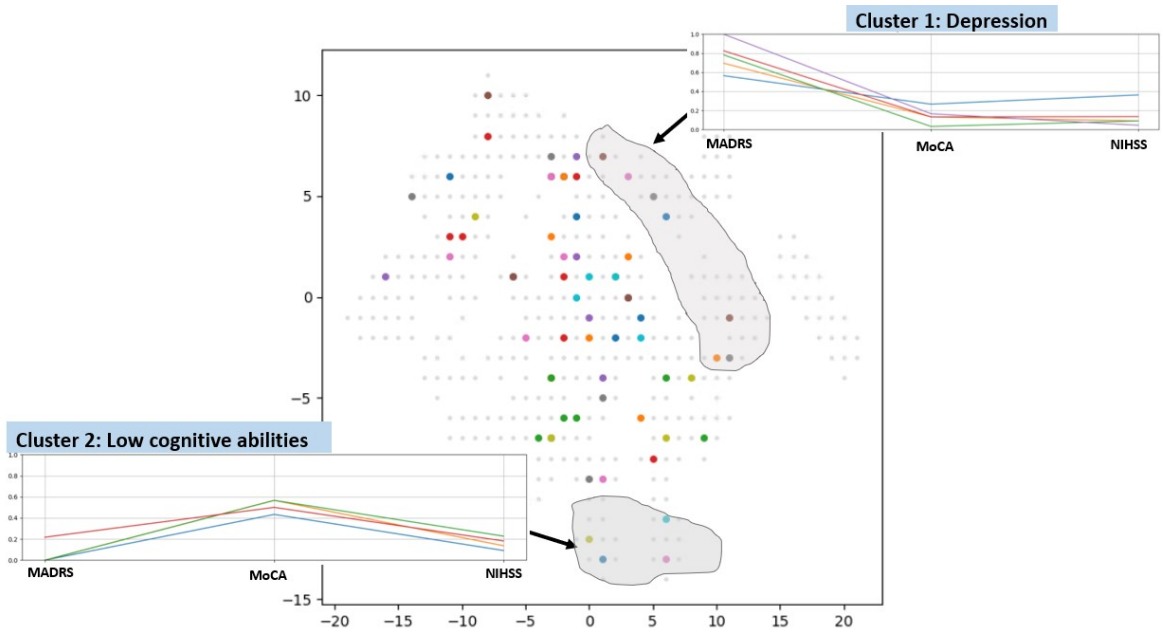


Fig. 6.8 Distinct clusters of mild stroke patients at day 7 post-stroke.

2 survivors showed no impairment following stroke. These insights emphasized that, even though the entire study cohort is classified merely as mild stroke survivors, variations of impairments occur among survivors classified as mild.

Different profiles of mild stroke patients at different time points of their patient trajectories

The START study provides participants with stroke assessment measures at 3-7 days, 90 days, 365 days post-stroke. The measures are NIHSS, MOCA, MADRS, RAPA, WSAS, MRS and SIS which assess the stroke impairment in different domains. The GSOM algorithm was applied to participant data at these three time points separately to infer different groupings of mild stroke survivors across various time-points.

Variants of impairment at 3-7 days post-stroke

At 3-7 days post-stroke, only MADRS and MoCA assessment outcomes were reported. Based on these data, the AI module generated three significant clusters indicating different characteristics of mild stroke patients. Fig. 6.8 illustrates the clusters identified at day 7.

Using the AI module, it was possible to differentiate stroke survivors with severe depression and cognitive impairments separately. It can be seen that the participants grouped in cluster 1 suffer from severe depression, however, their cognitive abilities are not severely impaired. On the contrary, cluster 2 participants have indicated high impairment in cognitive

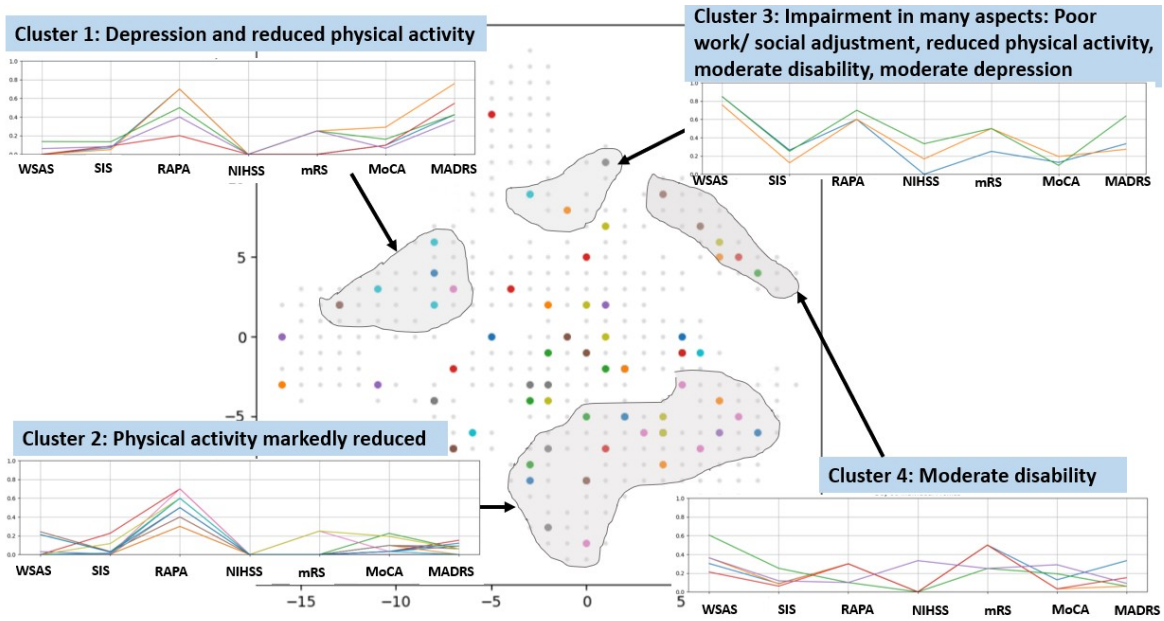


Fig. 6.9 Distinct clusters of mild stroke patients at day 90 post-stroke.

abilities while their depression level is not significant. However, based on the NIHSS, all participants are classified only as mild stroke survivors. We position the use of automated profiling of participants adjunct to basic NIHSS screening so that the stroke survivors can be provided customized care, specific to their complexities.

Profiling at 90 days post-stroke

At 90 days post-stroke, scores related to RAPA, MADRS, mRS, SIS, WSAS and MoCA were reported. Based on these data, the AI module generated three significant clusters indicating different characteristics of mild stroke patients. Fig. 6.9 illustrates the clusters identified at day 90.

It was observed that participants in cluster 1 exhibited reduced physical activity which was varying from low to moderate in the impairment index. As a notable exception, this group showed increased depression comparative to other participants. Cluster 2 participants also showed markedly reduced physical activity, however, they did not show signs of depression which differentiated them from cluster 1. The NIHSS is low for these two groups which align with the less impairment shown.

Cluster 3 participants also showed indications of depression, however, they also showed impairment in many domains such as poor work/ social adjustment, reduced physical activity and moderate disability. It can be noted that this group had the most severe complications out of other participants. Cluster 4 participants showed marginal similarity with cluster 3,

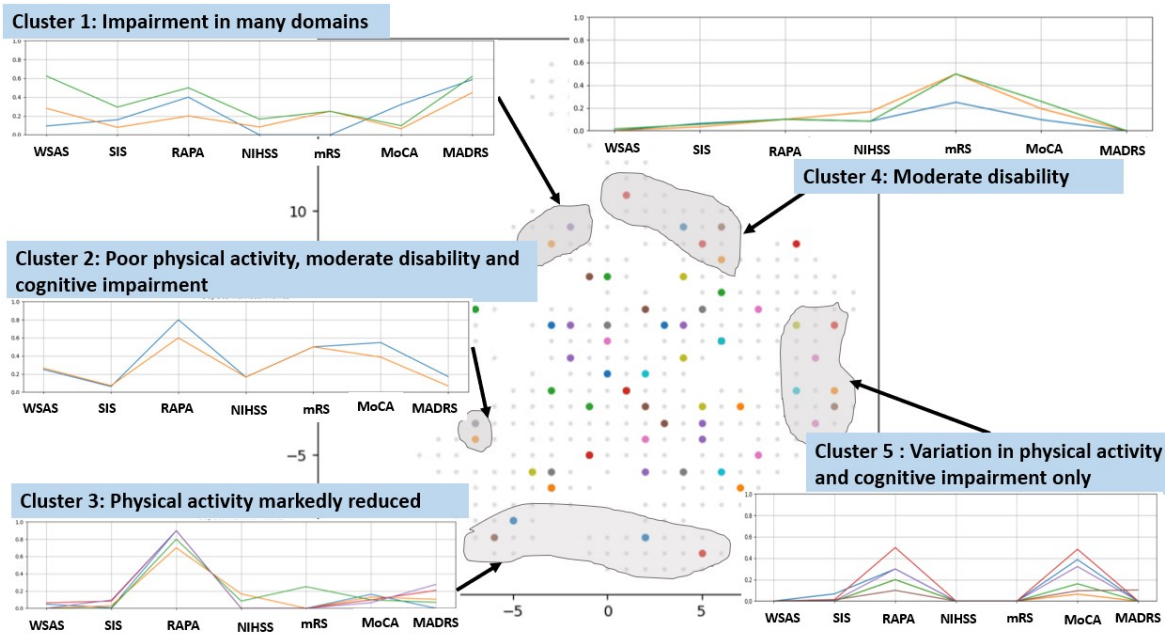


Fig. 6.10 Distinct clusters of mild stroke patients at day 365 post-stroke.

however, the impairment severity was low in this group. It can be comprehended that the NIHSS score also varies in clusters 3 and 4, indicating an increased disability.

The automated separation of participants based on their severity and variants of impairment act as evidence to support that, although participants are classified as mild, there are different groupings of complexities that occur among stroke survivors. The granular level of analysis at day 90 post-stroke enabled to detect several groupings of mild stroke survivors; group with higher levels of depression and reduced physical activity (cluster 1), group with reduced physical activity only (cluster 2), group with increased impairment in multiple domains (cluster 3) and group with low to moderate disability in many aspects (cluster 4).

Profiling at 365 days post-stroke

At 365 days post-stroke, scores related to RAPA, MADRS, mRS, SIS, WSAS and MoCA were reported. Fig. 6.10 illustrates the clusters generated by the AI module on day 365. It was noted that, participants in cluster 1 exhibited impairment in multiple domains such as impairment associated with work/ social adjustment, physical activity. As a notable exception, this group showed a higher level of depression which was not significant in other groups. Similarly, cluster 2 showed increased disability in different domains. However, in contrast to cluster 1, cluster 2 showed increased impairment in cognitive disabilities. From this, it can be deduced that participants in cluster 1 and 2 suffer from complexities in many domains and need to be closely monitored in rehabilitation.

Survivors in cluster 3 can be differentiated from others due to markedly reduced physical activity. They did not show notable impairments in other domains. This group should be targeted more from care related to uplifting motor functionalities. Cluster 4 showed increased disability as they have indicated a higher score in mRS which is used to rank the severity of the disability. Cluster 5 indicated a combination of moderate impairment in cognitive abilities and reduced physical activity. The constellation of impairments in each group demonstrated variants of impairment stroke survivors at 365 days post-stroke.

6.2.5 Discussion

Given the fact that life following a stroke affects both the physical and mental functionality of a person, it is imperative to comprehend the complexities associated (King, 1996). It has been established that vague measures at determining quality of life (QoL) following stroke impede clinician-decision making (Bosworth, 2001). Therefore, a multitude of assessments must be analyzed together in order to provide a more comprehensive view of stroke survivorship. The classification of stroke severity is often done using NIHSS screening which is a widely accepted measure in clinical settings. However, despite the classification of mild by NIHSS, it has been reported that 'mild' stroke survivors undergo an array of complexities associated with quality of life (Edwards *et al.*, 2006; Carlsson *et al.*, 2009). Therefore, traditional assessments and merely NIHSS screening are not capable of capturing the underlying reality of stroke survivorship.

To investigate the variations of mild stroke survivorship, we proposed an self-structuring based AI framework using latent representations to aggregate outcomes from multiple assessments such as cognitive function, physical activity, depression and social adjustment. The fusion of data from multiple assessments enables generating an overview of each individual, which is otherwise challenging to assess by conventional means. This new approach permitted to illustrate different profiles of stroke survivors despite the 'mild' classification by NIHSS. Therefore, it can be inferred that NIHSS alone is no longer content to classify a stroke survivor as mild. This serves as evidence to facilitate the multi-dimensional granular level of analysis apart from traditional or conventional methods of stroke screening.

We assume several potential implications for clinicians arising from this study. First, we provide evidence to showcase different profiles of impairment that exist among mild stroke survivors. The identification of different groupings among mild stroke survivors shows the inadequacy of using only NIHSS to classify severity. The evidence presented in this study relating to various groupings of mild stroke survivors confirms that the stroke severity classification should not merely rely on neurological functions, but should also incorporate cognitive, physical and social activity as measures. This enables widening the scope of

pre-screening following stroke as well as demonstrates the need for incorporating multiple aspects in the severity classification process.

Second, the distinct profiles enable providing targeted care and rehabilitation to stroke survivors focusing on the domain of impairment. The identification of cognitive aspects and depression in mild stroke patients can initiate treatments related to mental health and QoL. This also promotes early intervention therapies as the detection of participants with similar levels of depression could be used to facilitate counseling and necessary care at the early stages. This could also be extended to support group therapies where patients with similar symptoms can participate in order to share their experiences. Similarly, stroke survivors undergoing different levels of impairment could be recommended for necessary treatments. Such need-based care and precautions would enrich the post-stroke quality of life in mild stroke patients. We assert that such profiling and targeted treatments would uplift the current rehabilitation and facilitate patient-centered care.

Third, we propose the use of AI in patient data analysis as it would uncover latent patterns and groupings, which are otherwise challenging to assess. We suggest the plausibility of integrating AI-enabled insights for decision-making and designing strategies for rehabilitation that are associated with uplifting quality of life in patients. Using the framework presented in this study, clinicians can feed in all the data related to different stroke assessments and visualize distinct subgroupings of stroke survivors. The framework is scalable to accommodate data from a large number of patients, thus relevant to apply and understand latent representations of patient information at a larger scale. While this serves as a cost-effective decision-making platform it also categorizes patients based on the similarity of their impairment permitting clinicians and therapists to strategically design treatment and rehabilitation programs for patients who have similar disabilities.

As limitations of this study, we acknowledge that the portion of participants with missing data could be improved. As future enhancements, the cluster analysis could be improved to identify and track each individual's recovery information over time. This will enable to observe the survivorship patterns of each individual over time.

Altogether, while presenting the prospect of using AI in clinical settings, we believe that conventional and traditional means of stroke severity screening should be revisited to incorporate impairment from different domains. Mild stroke survivorship should be investigated at a granular level to explore and investigate if the categorization of mild 'is really mild?'. The discovery of different patient profiles will systematically empower patient-centered rehabilitation based on individual treatment needs thereby advancing the due supportive care with stroke survivorship.

In conclusion, healthcare sector is rapidly moving towards a digital transformation through disruptive technologies and cultural change. From wearable devices to ingestible sensors, from mobile health apps to AI, from robotic carers to electronic records, all constituents of the domain of digital health enables a seamless representation of a digital ecosystem, leading to countless opportunities to uplift the healthcare of the general public. In this case-study, we analyzed clinical outcomes of a stroke survivor patient cohort during three timepoints: 7-days, 3-months and 12-months post-stroke, with the aim to identify distinct clinical profiles and recovery trajectories, in which we utilized the new self-structuring based AI techniques developed in this thesis.

We used the fundamentals of MSKRF conceptual framework and its materialization, RTGSOM algorithm, for the analysis. For this particular case-study, five biological bases we derived to develop the MSKRF conceptual framework were involved; (i) invariant representation of memory, (ii) transience of memory, (iii) sequential information storage, (iv) auto-associative recall of memory, and (v) hierarchical abstraction in memory, which have been introduced in Section 2.3. These new capabilities brought in to solve this particular problem was significant in uplifting the understanding of stroke patient survival patterns and their recovery trajectories, leading to accurate and informed decision-making.

6.3 Summary and Research Questions Revisited

This chapter demonstrated the capabilities of the novel computation models developed in this thesis using two key application domains: Smart City and Digital Health. The first case-study focused on Smart Cities that endeavour to deliver safe, sustainable, effective asset utilization and service provision, amidst rapid urbanization. Video surveillance being a prominent application area under Smart Cities, the case-study aimed to focus on a number of unaddressed issues in video surveillance, such as sub-optimality, latency, predictive accuracy, and most importantly the contextualization of all detected salient objects for further decision-making. In Section 6.1, we proposed a new generative self-structuring and deep learning based approach to overcome the challenges posed in smart-city environments, where we demonstrated an adaptation for a License Plate Detection (LPD) use-case. An evaluation of the proposed GenLS approach was conducted using a state-of-the-art benchmark dataset captured using a single low-cost camera under different weather and recording conditions, in a realistic setting.

This case-study presented a successful implementation of the proposed MSKRF concept framework for a real-world situation, when an environment (smart city) is being represented in a digital form by using advanced data sensing technologies, generate large, multi-modal,

multi-source, dense, high frequent and non-stationary datasets. The new self-structuring AI capabilities brought in to solve this particular problem was significant in uplifting the predictive modelling capabilities of the deep learning techniques, as well provided a contextualization of the digital environment leading to further decision-making, thereby making the new self-structuring AI the missing piece of a successful intelligent surveillance system. The case-study was submitted as a journal article entitled *A Generative Latent Space Approach for Real-time Smart City Surveillance* (Nawaratne *et al.*, 2020b).

Section 6.2 presented a case-study in Digital Health focused on neuroscience and mental health. We analyzed clinical outcomes of a stroke survivor patient cohort during three time-points: 7-days, 3-months and 12-months post-stroke, and identified distinct clinical profiles and recovery trajectories using the new self-structuring based AI techniques developed in this thesis. The findings supported to create a comprehensive view of stroke experience thereby contributing towards uplifting the quality of life and rehabilitation of stroke survivors. This work was submitted as a journal article entitled *"Is mild really mild?": Patient profiling using Artificial Intelligence* (Nawaratne *et al.*, N.D.).

In doing so, we successfully addressed the third research question (RQ3): **"How can such a self-structuring AI based continuous lifelong learning architecture with memory be developed in to technology platforms to advance AI systems in perpetual data intensive environments, such as national security, smart cities, and digital health?"**.

Chapter 7

Conclusion

This chapter concludes the thesis with a summary of research contributions, limitations of the current research and directions for future research. A conclusive summary of the research is presented in Section 7.1 detailing the key contributions made to expand the body of knowledge related to self-structuring AI, continuous lifelong learning and unsupervised machine learning. Section 7.2 describes how those contributions made possible to address the key research questions. The chapter is concluded in Section 7.3 with a discussion of the advances in the field of continual lifelong learning during the past few years while shedding light on the impact and limitations of the methods proposed in this thesis. We conclude this thesis by pointing out several promising research directions towards realizing continual lifelong learning.

7.1 Summary of Contributions

The advancement of IoT and Big Data have resulted in events and situations being more holistically represented digitally and changes in such situations are captured over time. This has enabled a new digital world that provides a closer representation of the natural world than ever. The data being generated from multiple data capturing technologies are vast in volume, formed in both structured and unstructured forms, constitute of different modalities and varieties, and are being generated in high velocity, imposing significant challenges to current AI systems. This demands an advancement of AI technologies to autonomously adapt and update its knowledge of the environment over time.

In the context of AI, the problem of learning from such continuous streams of data, adapting to the external environment and associate with different tasks with the goal of augmenting the acquired knowledge for problem solving and future learning is termed *Continuous Lifelong learning* (CLL). CLL is a learning paradigm that focuses on a higher

and realistic time-scale where data and tasks become available real-time and the access to past data is limited. The stability-plasticity dilemma poses a major hindrance for the effective performance of CLL systems, which is a problem that has been studied for decades. As such when AI systems are exposed to continuous streams of non-stationary data, the new knowledge severely disrupts past knowledge, forgetting what it has learnt previously. This is well-known as catastrophic forgetting, which leads to an abrupt performance decrease or, in the worst case, to the past knowledge being completely overwritten by the new. In order to overcome these challenges, memory requires a degree of plasticity for the integration of new knowledge, while being stable in order to prevent the disruption of existing knowledge.

In contrast to AI systems, the evolution of the human over 6 million years reveals a remarkable phenomenon in terms of human learning and the ability of the human to adapt to continuously changing. The human brain is capable to constantly adapting and exploiting new information in this complex, volatile and evolving physical world (Shatz, 1992). Humans accumulate knowledge gained over a lifetime and use this accumulated knowledge to assist future learning and decision making with possible adaptations. The primary bodily component of humans to achieve such a remarkable ability is the biological brain. Chapter 2 studied the structural, functional and behavioural facets of the biological brain, primarily focusing on visual perception system and memory system, to understand the ability of the human to demonstrate complex behaviours, skills whilst having a memory formulation that can continuously learn and adapt.

Drawing from the understanding obtained on the structural and functional formation of the biological brain, Section 2.3 proposed seven key constituents in human neuronal system that has the potential to aid its artificial counterpart to advance capabilities in perceiving and representing natural environment in order to process Big Data, and derive insights that can be transformed to actions and recommendations. Founded upon the neurophysiological inspiration and the features of the big data and digital environment, Section 2.4 presented the landscape of AI systems positioned within natural world, providing a high-level overview and scope for the thesis. Drawing on the basis of the landscape, Section 2.5 proposed an overarching conceptual framework, *Multi-layered Self-structuring Knowledge Representation Framework (MSKRF)*, in which the conceptualisation, design and development of new AI system could be materialized. MSKRF consists of four layers: (i) Sensory Inputs, (ii) Latent Representation (LR), (iii) Cognitive Representation (CR), and (iv) Continual Knowledge Acquisition. The intermediate representation layers were designed to resemble the biological brain, in which a representation of the natural world is developed for its interaction with it. Thus, the intermediate representation layers require to cater to the 'unknown' latent and cognitive derivations which will occur during actual situations/events. Thereby, both

representation layers should be able to adapt its representation along with continual changes in data distribution over time, to resemble evolving external environment. In Section 2.5, catering to the need to adapt to the unknown structure of the natural representation, we proposed Self-Structuring Artificial Intelligence (SSAI) with an unsupervised learning paradigm as a solution to design these intermediate representation layers.

Chapter 3 brought to fruition the proposed MSKRF framework from initial conceptualization and design to a practical and usable AI technology for real applications. Section 3.1 investigated the feasibility of modelling the indeterministic natural environments using machine learning paradigms in order to facilitate representation learning. The investigation narrowed down to unsupervised self-organization as a viable prospect for representation learning, in which we examined both natural and biological prospects that use self-organization for multitude of representation mechanism in Section 3.2. We identified experience-driven self-organization as a key constituent in biological brains that enable humans to continuously acquire knowledge in their lifetime, i.e., lifelong learning of humans. Section 3.3 explored the foundation of SSAI from biological and statistical perspective to understand how self-organization could be facilitated, both biologically and in nature.

This investigation resulted in evidence that support the argument that self-structuring can enable self-organization. An in-depth survey on existing SSAI techniques, their limitations and prospects resulted in selecting Growing Self-Organizing Maps (GSOM) as the viable candidate algorithm to base SSAI for the proposed MSKRF. Section 3.5 demonstrated the capabilities of the GSOM algorithm through a practical exploration in a Smart City platform forcing on intelligent video surveillance. The conceptual formulation of SSAI and the smart city based practical exploration was presented in the journal article entitled *self-building artificial intelligence and machine learning to empower big data analytics in smart cities* (Alahakoon *et al.*, 2020). An extended study to utilize SSAI in IoT data interoperability environments was presented in the journal article entitled *Self-evolving intelligent algorithms for facilitating data interoperability in IoT environments* (Nawaratne *et al.*, 2018).

Chapter 4 focused on the development of a computational model for representation learning to materialize the intermediate learning layers, latent and cognitive representation layers, of the MSKRF. Section 4.1 studied the limitations posed in current machine learning algorithms due to their focus purely being on persistence, thereby proposed to integrated transience as a solution. This lead to the implementation of a strategic forgetting mechanism for GSOM followed by an experimental evaluation of topology preservation for the new GSOM with transience (TGSOM) algorithm. The novel algorithm was presented in the conference article entitled *HT-GSOM: dynamic self-organizing map with transience for human activity recognition* (Nawaratne *et al.*, 2019b).

Section 4.2 explored non-stationary and sequential nature of environmental stimuli, in which sequences of data are generated by a series of temporally and spatially connected observations. It was identified that sequential data plays a vital role in biological perception system since all sensorimotor data are given as sequences of time-varying stimuli, which makes it pertinent to be incorporated in the AI counterparts. As a solution, TGSOM algorithm was extended by implementing a recurrent information processing mechanism to capture sequential information from input stimuli. This recurrent TGSOM algorithm (RT-GSOM) was presented in the journal article entitled *Recurrent Self-Structuring Machine Learning for Video Processing using Multi-Stream Hierarchical Growing Self-Organizing Maps* (Nawaratne *et al.*, 2020a).

The natural environment is highly complex; thus, human nervous system is equipped to sense this complex environment with multiple modalities. When an event occurs, more than one sensor detects the events, generating redundant neural signals underpinned by the multisensory processing of biological brain. To capture such environments in AI systems, Section 4.3 laid the foundation to process multiple information streams by implementing a multi-stream self-organization architecture. The proposed multi-stream self-organization architecture is experimented using dense activity recognition video data sets in Section 4.3.1 in order to confirm its validity and usability in real-world settings. The algorithm formulation and experiments were presented in the journal article entitled *Hierarchical Two-Stream Growing Self-Organizing Maps with Transience for Human Activity Recognition* (Nawaratne *et al.*, 2019a).

This research defined continual lifelong learning as the ultimate goal of the proposed conceptual AI framework. Chapter 3 and Chapter 4 were directed towards the development of the algorithmic foundation, while Chapter 5 focused on development of variations of continual lifelong learning AI systems using the developments of this thesis so far. Section 5.1 introduced the challenges associated with CLL that needs to be addressed in order to achieve CLL in computational models. These challenges are two-fold where the first relates to the evolving nature of input stimuli (data) while the second relates to the evolving nature of tasks computational models target to achieve. Section 5.2 addressed the first challenge associated with evolving nature of data, followed by proposing a new unsupervised deep learning based active learning approach and its experimental evaluation in the context of intelligent video surveillance focused on anomaly detection. The algorithms and experiments were presented in the journal article entitled *Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance* (Nawaratne *et al.*, 2019c).

Section 5.3 addressed the second challenge associated with the evolving nature of tasks. Followed by an in-depth review on existing Complementary Learning Systems (CLS) based

computational models, this thesis designed and developed a new self-organization based CLL approach, named LifeNet, by incorporating constituents of CLS theory. LifeNet was designed using an architecture of RTGSOM adopted from the developments in Chapter 4. The LifeNet completed the materialization of the proposed MSKRF, which was evaluated using a series of benchmark datasets on object recognition and human activity recognition. The preliminary analysis towards the development of LifeNet was presented in the conference article entitled *Incremental knowledge acquisition and self-learning for autonomous video surveillance* (Nawaratne *et al.*, 2017).

Chapter 6 demonstrated the capabilities of novel algorithmic developments presented in this thesis using two case-studies from Smart City domain and Digital Health domain. The first case-study focused on Smart Cities domain aimed to detect and localize salient objects in surveillance video streams. We addressed a number of unaddressed issues such as sub-optimality, latency, predictive accuracy, and most importantly the contextualization of all detected salient objects for further decision-making. Section 6.1 proposed a Generative Latent Space (GenLS) approach that was developed using the proposed RTGSOM algorithm, to overcome aforementioned challenges, where we demonstrated an adaptation for a License Plate Detection (LPD) use-case. An evaluation of the proposed GenLS approach is conducted using a state-of-the-art benchmark dataset captured using a single low-cost camera under different weather and recording conditions, in a realistic setting. The case-study was submitted as a journal article entitled *A Generative Latent Space Approach for Real-time Smart City Surveillance* (Nawaratne *et al.*, 2020b).

Section 6.2 presented the second case-study in Digital Health domain focused on neuroscience and mental health. In this case-study, we systematically investigated different profiles of survivors classified as 'mild' stroke severity based on tests measuring their stroke impairment and impact using self-structuring latent representations to uncover latent patterns from multiple measures, at different times in the recovery trajectory; with ongoing impairments and impact even 12-months post-stroke. The temporal resolution patterns from the clinical data were successfully captured using the latent representation algorithm developed using our proposed RTGSOM algorithm. The findings supported to create a comprehensive view of mild stroke experience thereby contributing towards uplifting the quality of life and rehabilitation of mild stroke survivors. This work was submitted as a journal article entitled *"Is mild really mild?": Patient profiling using Artificial Intelligence* (Nawaratne *et al.*, N.D.).

In summary, we argued for the necessity for a new type of learning capability for AI to confront Big Data challenges. This thesis is inspired by the remarkable advancement of humans, provided with unprecedented capabilities to adapt to diverse environmental conditions by the evolution for over 6 million years through simple lifestyles to complex,

under unpredictable environmental conditions and circumstances. This thesis is inspired by the remarkable capability of continuous lifelong learning in humans, which can assist AI systems to further advance its capability to perform in complex, volatile and evolving digital environments. Moreover, this thesis acknowledged the vital importance of unsupervised self-learning and concepts of self-structuring in AI for the development of a representation of the natural environment. Founding upon the neurophysiological structural and functional facets of the human brain, this thesis attempt to develop AI systems capable of continuously acquire knowledge from the natural environment and thereby able to be used to derive insights that can be transformed into useful actions and recommendations. This work can be considered as a stepping stone for the development of more autonomous AI required for the future.

7.2 Addressing the research questions

This section describes how the above contributions have addressed the research questions delineated in Chapter 1.

1. **How can computational continual lifelong learning enable the natural world to be represented digitally, making use of continuous streams of data from a variety of digital sensors? What aspects of the structural and functional facets of neurophysiological studies can be used as a foundation premise to develop techniques for computational continual lifelong learning in data intensive digital environments?**

   The first research question investigates the structural and functional facets of neurophysiological studies that can be used as a foundation premise to develop computational continual lifelong learning. The research question is decomposed into four sub-questions as follows.

   (a) **How has the new digital world, made up of Big Data and IoT, transformed AI systems in perceiving natural environments, acquiring and updating knowledge for past and future tasks?**

   This sub-question is addressed in Section 2.1 exploring advancements in digital ecosystems with respect to how AI systems perceive natural environments and identified key limitations in current state-of-the-art AI systems, suggesting the need for a new thinking about conceptualizing AI systems.

(b) **What are the core structural components and functional mechanisms in the human neurophysiological system that support the continual lifelong learning in humans?**

This sub-question is addressed in Section 2.2 by reviewing the key functions and structural formulations of the biological brain. In light of the comprehensive introduction to the biological brain, the thesis presents seven biological bases that have been essential for the survival and continuance of humans.

(c) **How can these neurophysiological facets of humans be used to inspire artificial representation of natural world in data intensive digital environments?**

Section 2.2.2 addresses this sub-question by providing an in-depth discussion on the identified neurophysiological facets, in terms of the seven biological bases with respect to their role and responsibility in biological system, what limitations of AI systems can be addressed through each of these bases and in terms of their importance for the advancement of AI systems.

(d) **How can such neurophysiological inspiration be used to combine the features of big data and digital environment to form an overarching conceptual framework?**

This thesis combine the identified biological bases and a set of cognitive theories to develop a conceptual framework, Multi-layered Self-structuring Knowledge Representation Framework (MSKRF), to model the landscape for AI agents in representing the natural world. The landscape is presented in Section 2.4 and the conceptual framework is proposed in Section 2.5.

2. **What are the computational and machine learning constituents of continuous lifelong learning for materializing the proposed conceptual framework?**

This research question investigates the algorithmic and technical perspectives to design and develop the proposed MSKRF conceptual framework. The research question is decomposed into five sub-questions.

(a) **What are the computational and machine learning foundations for representation learning in the digital world that have been proven through both neurophysiological and ecological studies?**

This sub-question is addressed in Section 3.1 through Section 3.3 by reviewing key functions and natural formulations of self-organization and the need for self-structuring foundation in order to represent the natural world. Drawing on

the inspiration from nature and its natural phenomenon, this thesis introduced computational models that are capable to self-structure.

(b) **What are the structural and algorithmic limitations in current AI for achieving continuous lifelong learning and what fundamental architectural changes will address these limitations?**

This sub-question is addressed in Section 3.4 with a comprehensive exploration of existing self-structuring computational models to identify structural and algorithmic limitations in current AI that limit the achievement of continuous lifelong learning. The research determined that the pre-defined structure of existing computation methods limit their representation capability, thus allowing self-growth capability will enable a better representation. Thereby, we narrowed down the base algorithm to Growing Self Organization Map (GSOM) which captures the essence of an effective self-organization based representation mechanism. Further we demonstrated the validity and robustness of GSOM algorithm through a practical exploration under Smart City context in Section 3.5.

(c) **How can the knowledge embedded in computational models preserve stability and plasticity when introduced to continuous data streams?**

Chapter 4 addressed this sub-question by advancing the GSOM algorithm by combining the underlying concepts of self-structuring, transience and recurrent learning. To address this sub-question, we first discussed the importance of transience as an essential capability in a successful memory system to retain memory plasticity and the importance of capturing non-stationary temporal information from data streams. Section 4.1 provided a detailed review on how the existing computational models attempted to achieve plasticity without compromising the stability of the memory system, and Section 4.2 provided a review of existing methods that incorporate temporal information processing. These led to the development of novel RTGSOM algorithm that incorporated recurrent and transience capabilities which helps to preserve stability and plasticity of the knowledge model.

Followed by the algorithmic extension, we evaluated the performance of the algorithms using a suite of synthetic datasets that consist of a number of datasets with varying complexities and properties resembling problems in the real world. The experimental evaluation demonstrated that the RTGSOM has the capability to encapsulate plasticity in the neuronal latent representation without the loss of stability. Further, the experiment demonstrated how the self-organization with

transience will discard the outdated information and overfitting knowledge in its knowledge acquisition, without the loss of stability.

(d) **With multiple facets and characteristics of data being captured to represent actions, events and situations, how can a comprehensive representation be developed in digital environments?**

This sub-question was addressed in Section 4.3, with the design and development of a hierarchical multi-stream architecture that is able to capture multiple feature streams from videos and produce a unified stream of insights. The proposed architecture is based on RTGSOM that introduced transience and recurrent learning. We demonstrated the proposed model using three benchmark video datasets and the results confirm its validity and usability for human activity recognition.

(e) **What neurophysiological theories enable the development of a computationally plausible memory formulation to achieve continuous lifelong learning?**

This sub-question is addressed in Chapter 5. We based the memory formulations upon the two-fold computation challenges in continuous lifelong learning, that are: i) the evolving nature of input stimuli (data), and ii) the evolving nature of tasks computational models target to achieve. We presented two memory formulation techniques to address this research question.

Section 5.2 addressed the first challenge associated with evolving nature of data followed by proposing a new unsupervised deep learning based active learning approach in Section 5.2.1 and its experimental evaluation in the context of intelligent video surveillance focused on anomaly detection in Section 5.2.2. The second challenge in CLL, the evolving nature of tasks computational models target to achieve is addressed in Section 5.3.

3. **How can such a self-structuring AI based continuous lifelong learning architecture with memory be developed in to technology platforms to advance AI systems in perpetual data intensive environments, such as national security, smart cities, and digital health?**

The final research question was addressed in Chapter 6 by demonstrating the capabilities of the novel computation models developed in this thesis using two key application domains: Smart City and Digital Health. The first case-study was a Generative Latent Space (GenLS) approach developed using the proposed RTGSOM algorithm to address challenges in License Plate Detection in a smart city surveillance context. The second case-study was a patient profiling and trajectory analysis technique for stroke survivors/patients, which supported to create a comprehensive view of stroke

experience thereby contributing towards uplifting the quality of life and rehabilitation of stroke survivors.

7.3 Future Research Directions

The main contribution of the thesis is the conceptualization, design and development of the Multi-layered Self-structuring Knowledge Representation Framework (MSKRF). The fruition of the MSKRF was experimented and demonstrated using a number of benchmark and real-world datasets. The hierarchical models of the MSKRF are designed for learning batches of training data, thus assumes that a portion of the training data is available or stored for model development. This is an improvement from traditional AI techniques that assume the entire training data is available at the beginning. However, in more natural scenarios AI systems should incrementally acquire and process knowledge from the perceptual cues as they become available over time in an online manner.

The concept of predictive coding (Rao and Ballard, 1999) has generated significant interest in the recent past as a mechanism for human brain inspired predictive machines which can continuously develop a mental map of the environment and as such enable the machine to understand and adapt its actions accordingly. In the development of AI, rather than attempting to develop a pre-coded and pre-trained knowledge model, the ideal solution could be a knowledge model that adapt based on natural cues that become available over time. As argued by Wang and Fan (2018), even state-of-the-art deep learning algorithms fail in the type of strategic reasoning needed to predict and understand someone else's incentives and goals whether it is an AI or human. The missing piece to this development towards an advanced AI is the type of deep communication and understanding skills that stems from a critical human cognitive ability: Theory of Mind (TOM) (Rabinowitz *et al.*, 2018).

In addition, Jung *et al.* (2015) investigated the interactions between top-down predictions and bottom-up regression in hierarchical self-organization architectures for spatio-temporal in dynamic data intensive environments. The recurrent learning capability of the proposed RTGSOM algorithm has potential to embody the feedforward and feedback connectivity of the self-organized network, which in turn will aid the development of such a top-down and bottom-up mental map that will enable to correctly predict from incoming streams of data.

As such, we propose future research on continuous lifelong learning should focus towards a hierarchical memory architecture of TOM enhanced bi-directional predictive processing to build AI with improved ability of autonomous and ongoing prediction.

Appendix A

Supplementary Material

A.1 Topography Evaluation of TGSOM with FCPS Data Suite

This section provides the detailed representations of the TGSOM evaluation using FCPS Data Suite. Here we detail the TGSOM latent representations for all the parameter configurations.

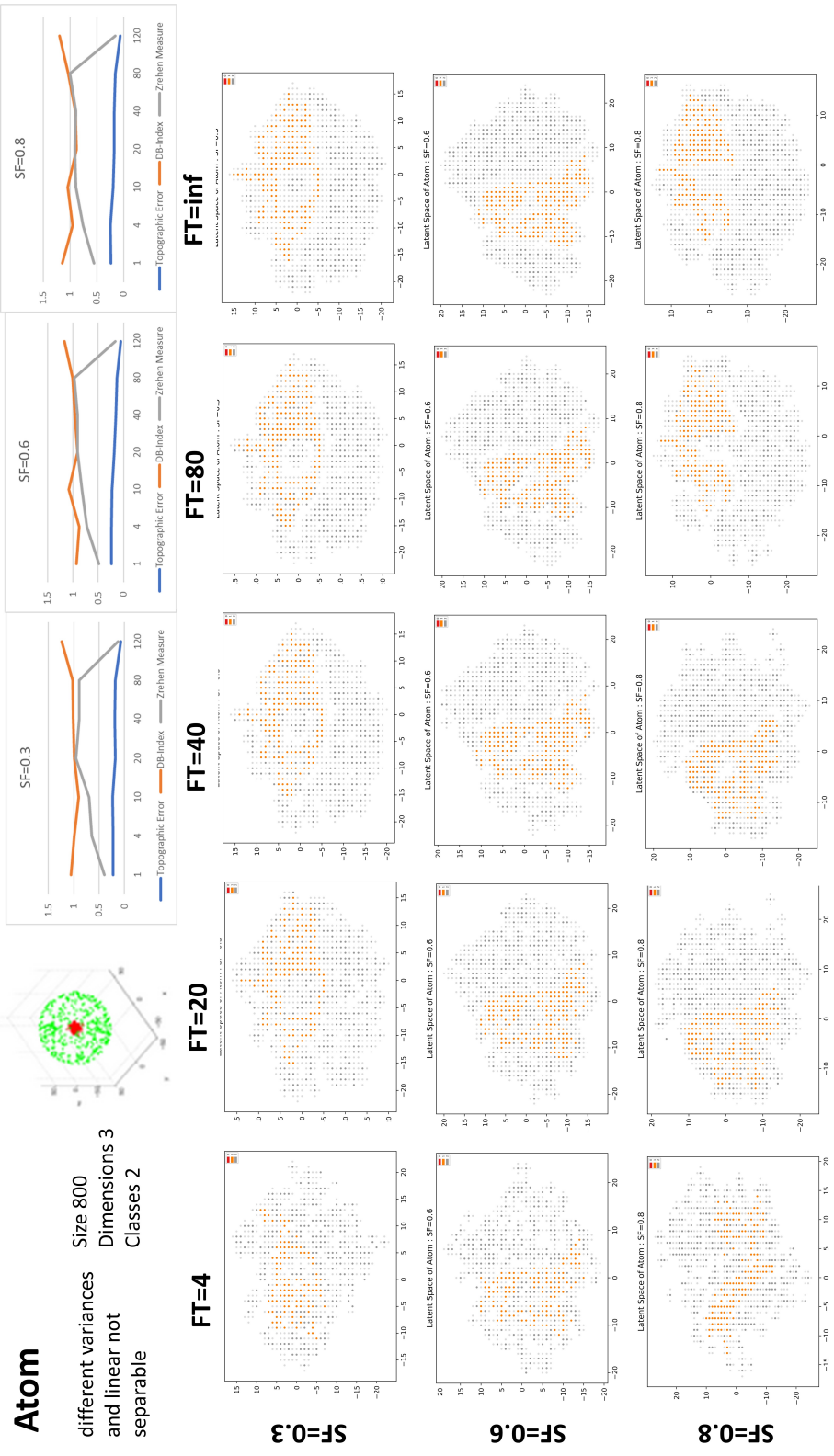


Fig. A.1 Topography Evaluation of ATOM dataset

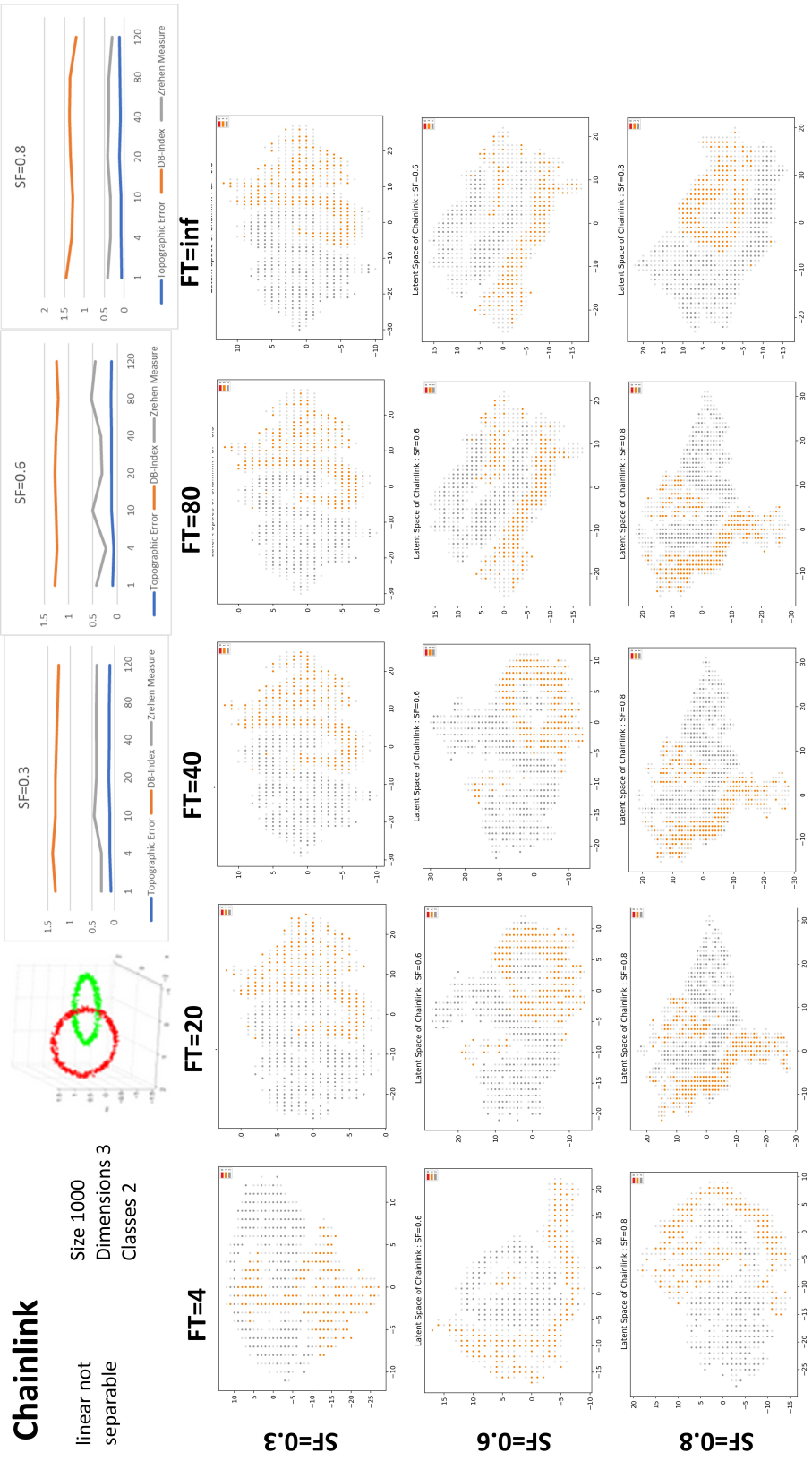


Fig. A.2 Topography Evaluation of ChainLink dataset

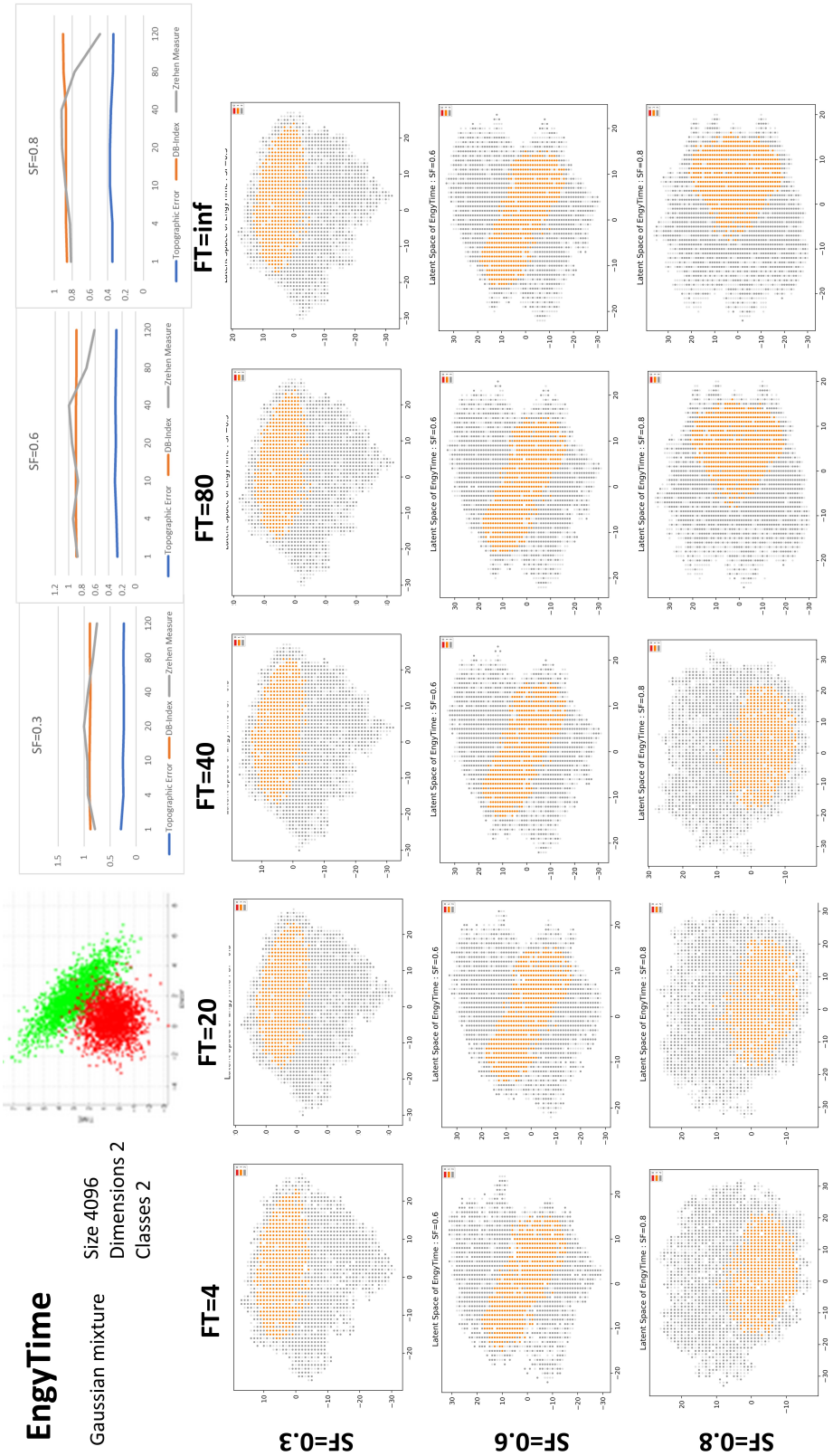


Fig. A.3 Topography Evaluation of ENGYTIME dataset

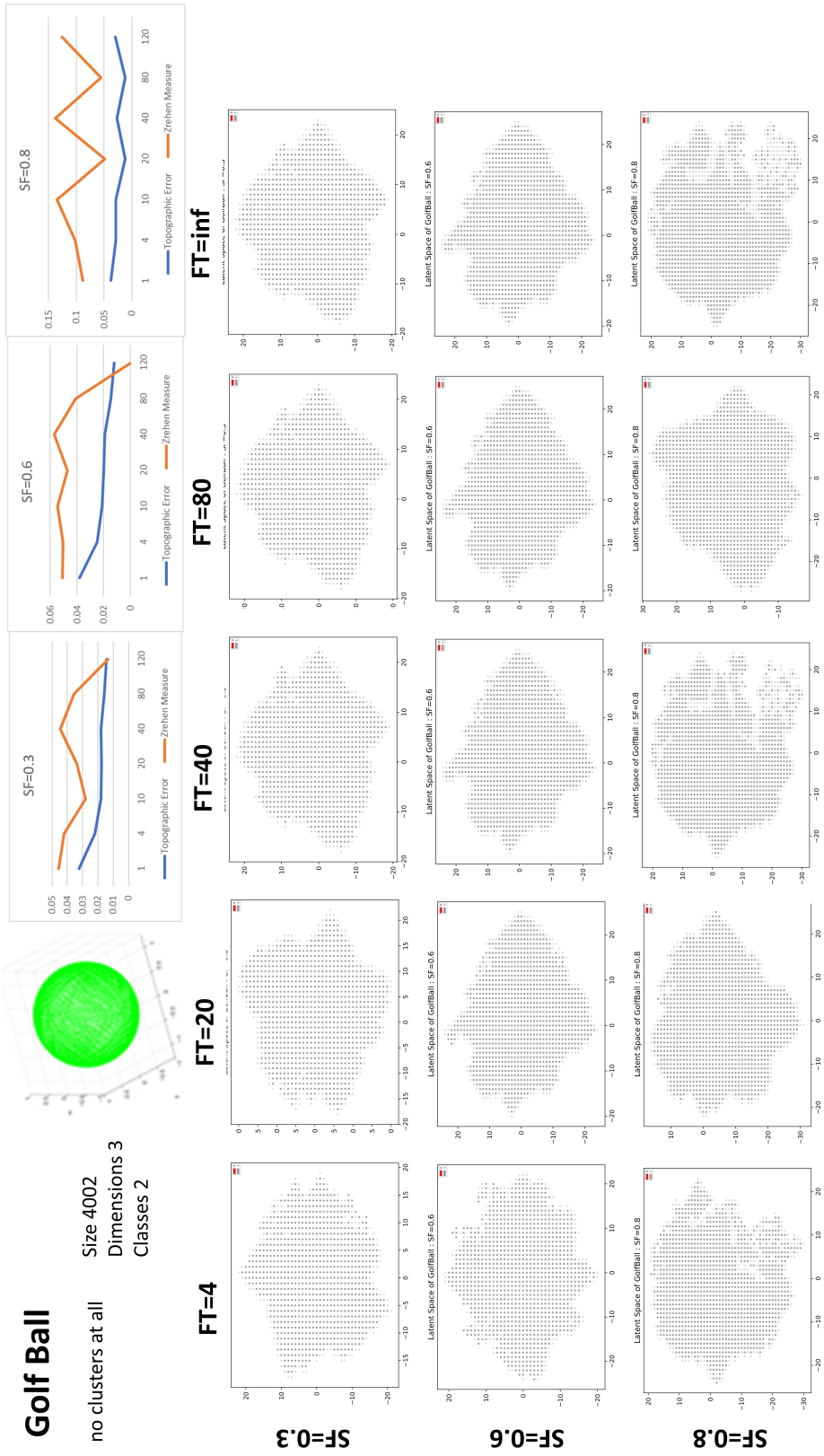


Fig. A.4 Topography Evaluation of Golf Ball dataset

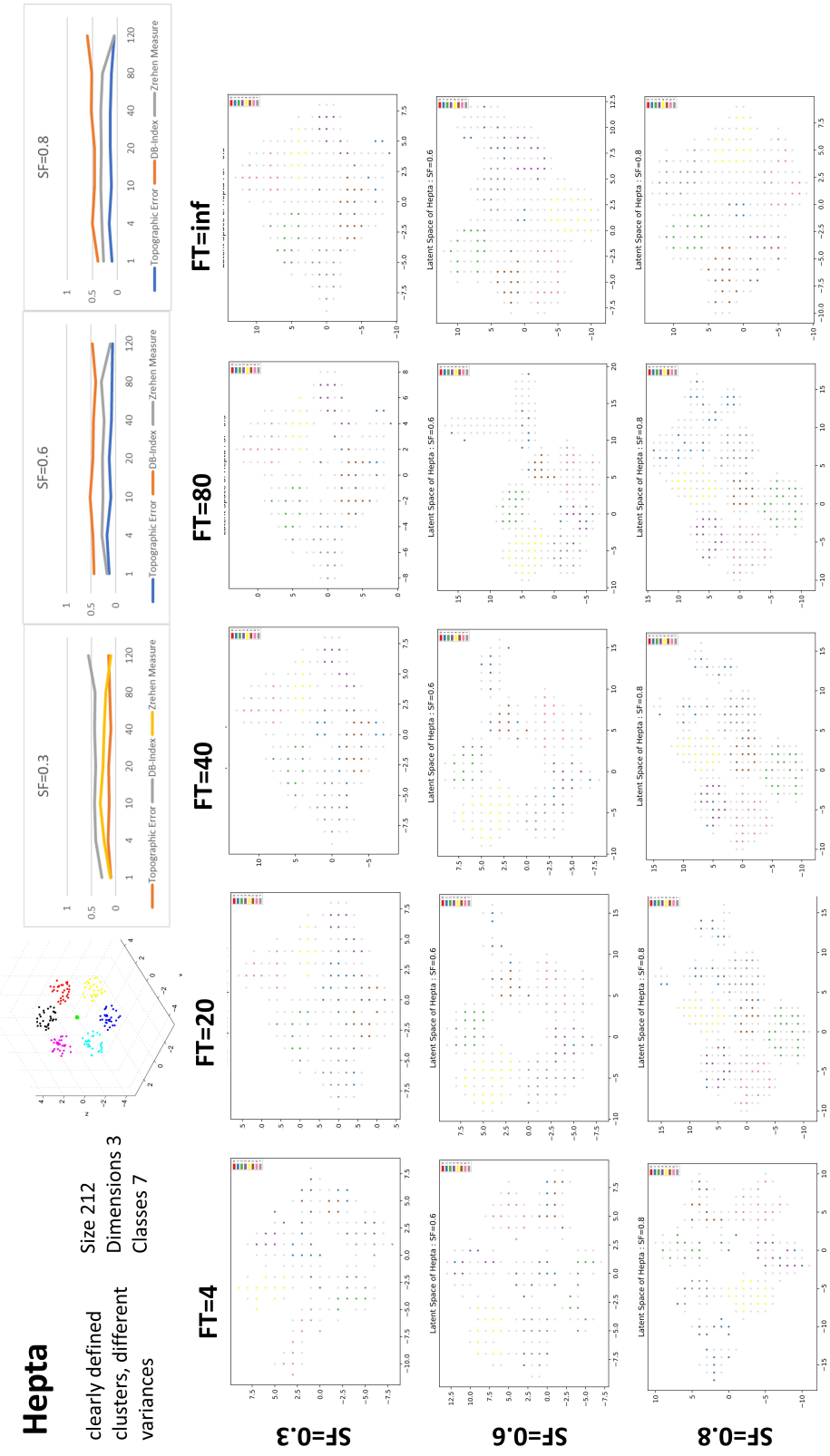


Fig. A.5 Topography Evaluation of HEPTA dataset

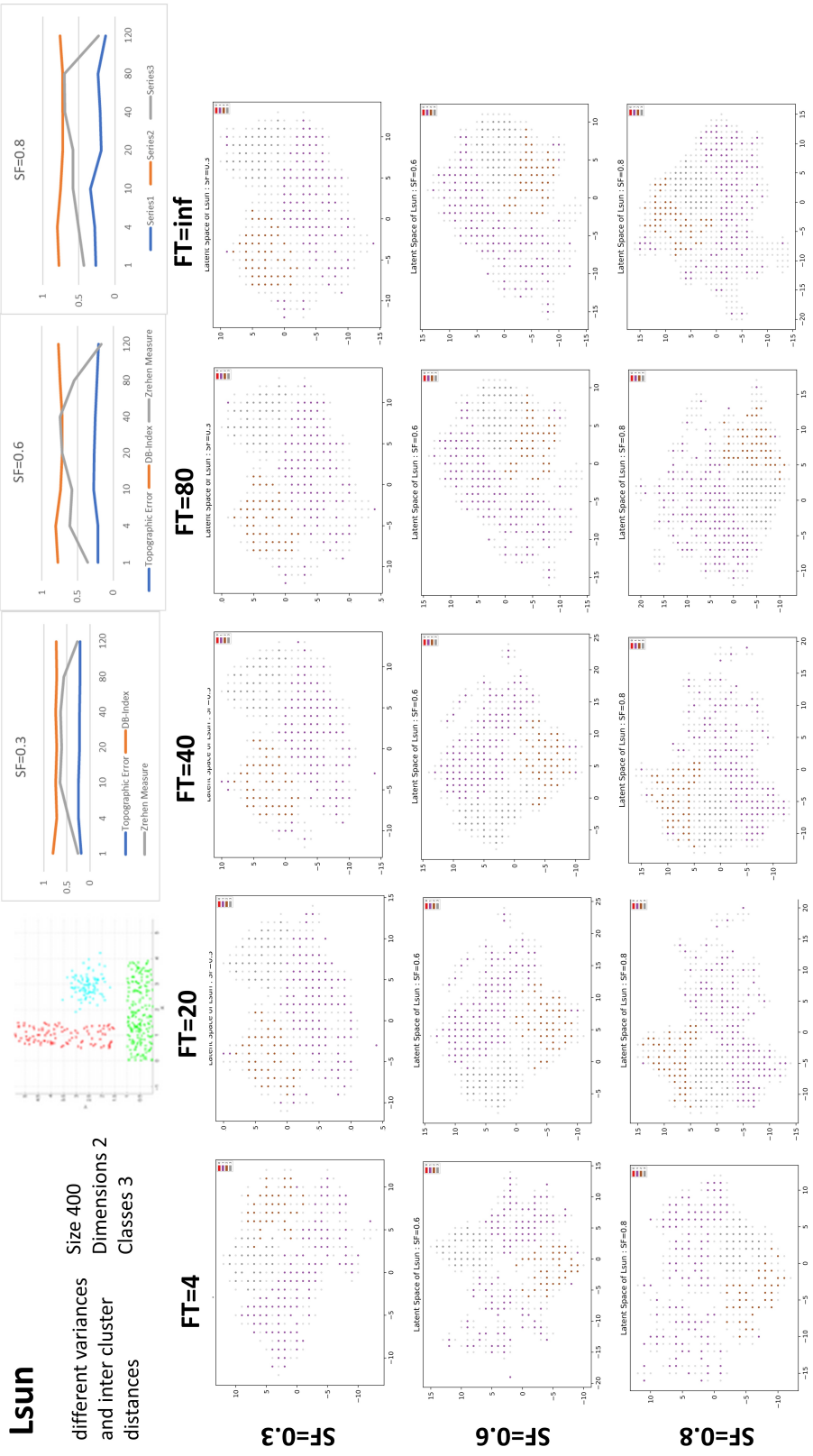


Fig. A.6 Topography Evaluation of LSUN dataset

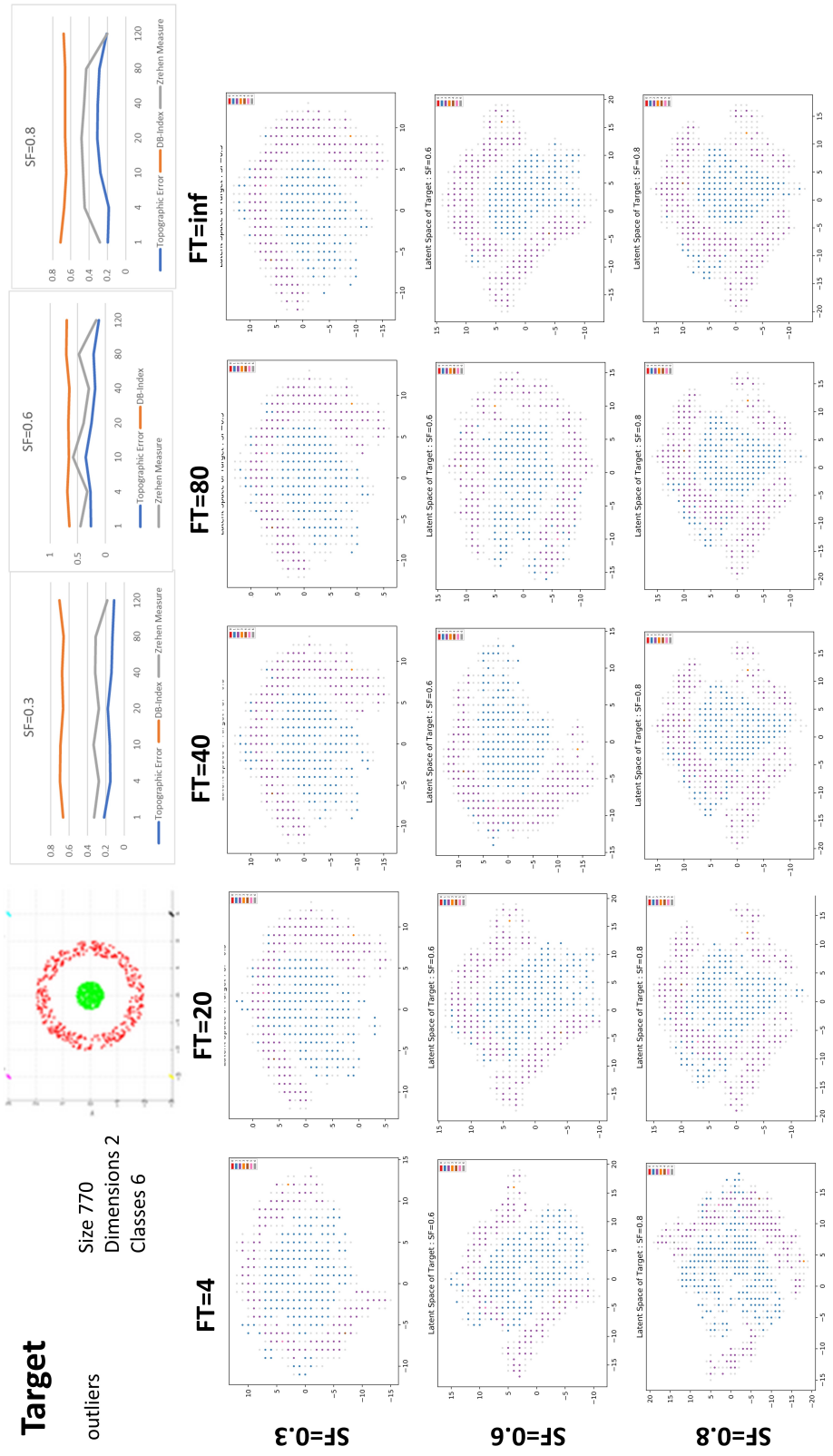


Fig. A.7 Topography Evaluation of TARGET dataset

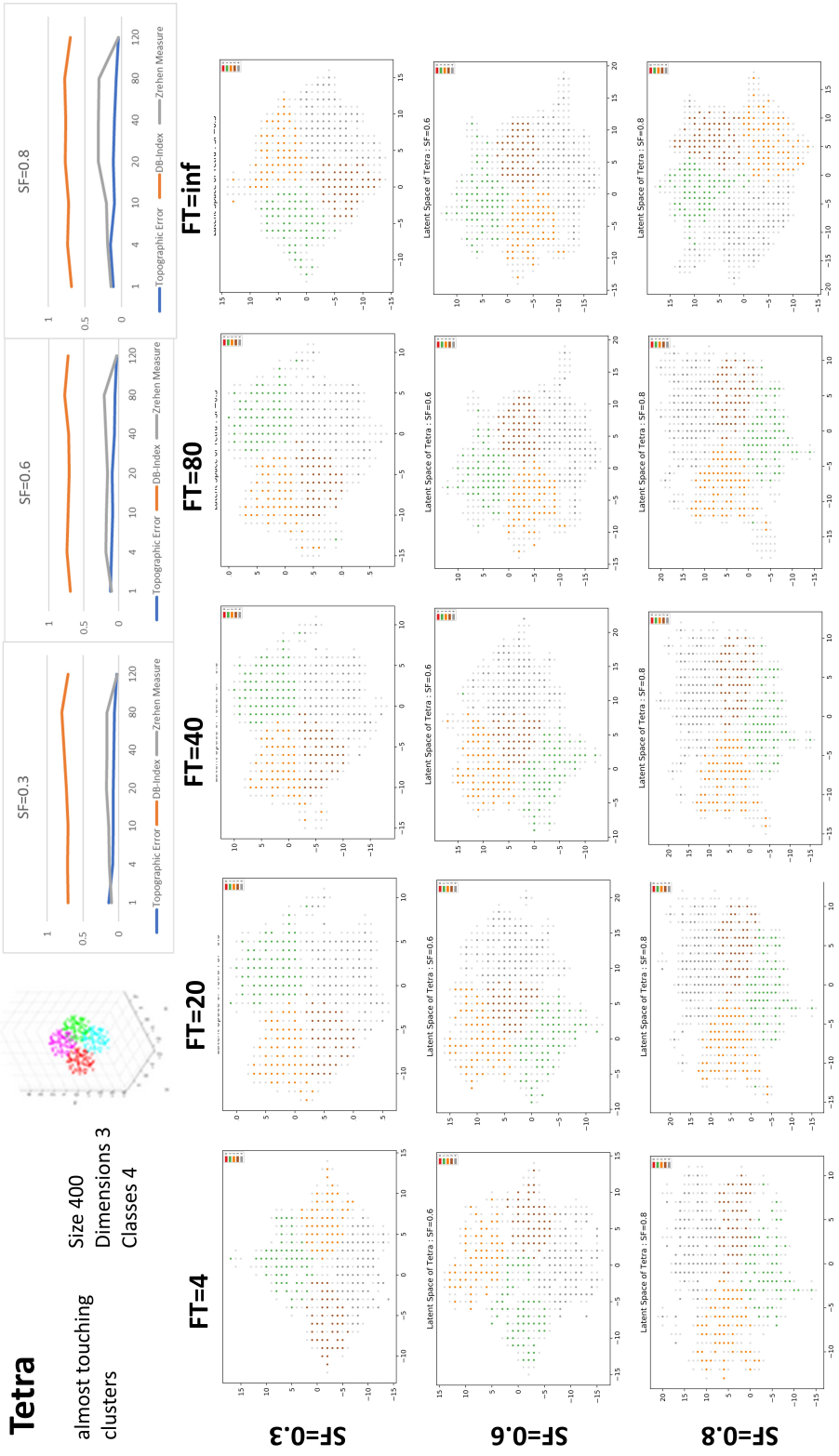


Fig. A.8 Topography Evaluation of TETRA dataset

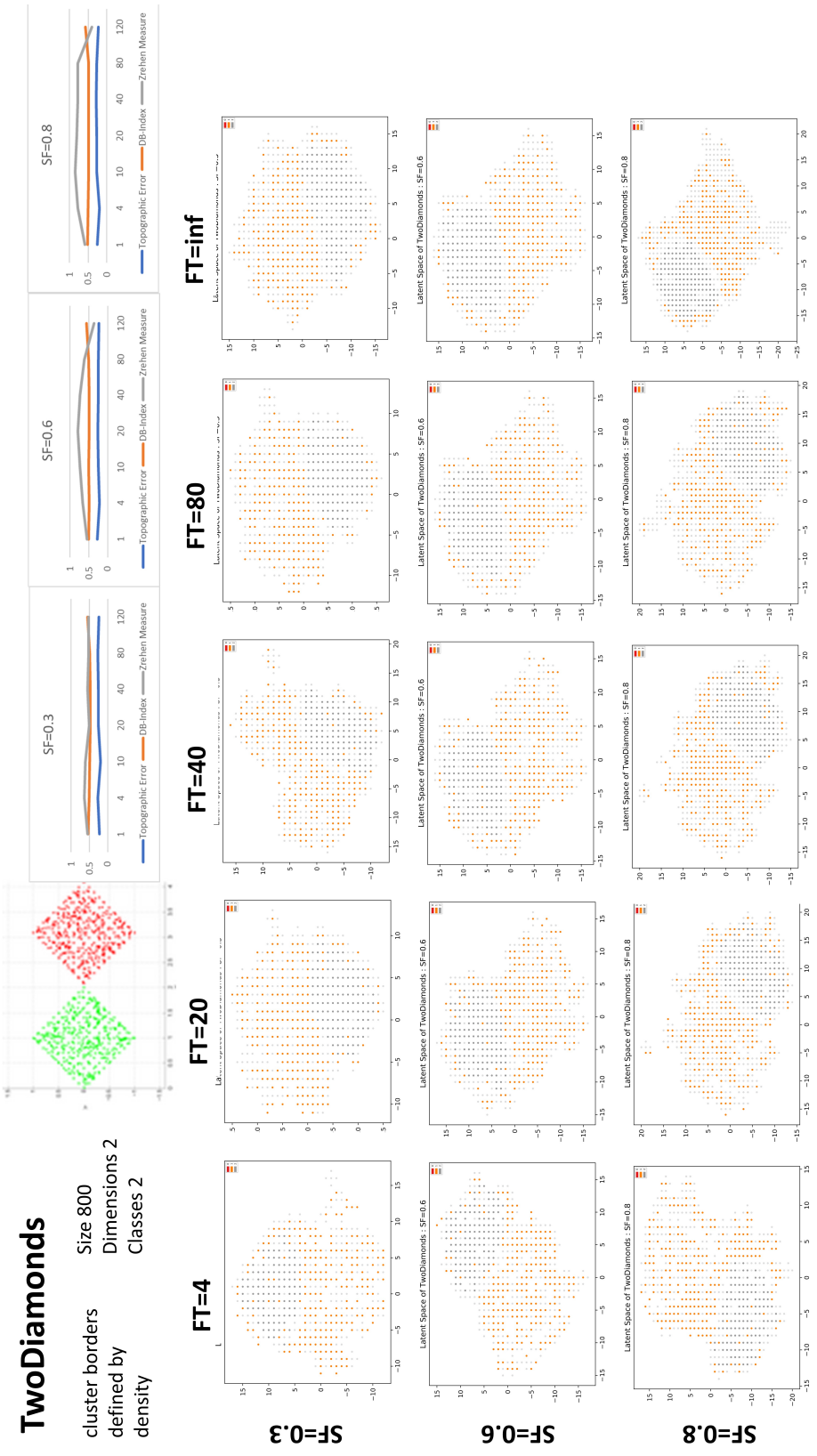


Fig. A.9 Topography Evaluation of TWO DIAMONDS dataset

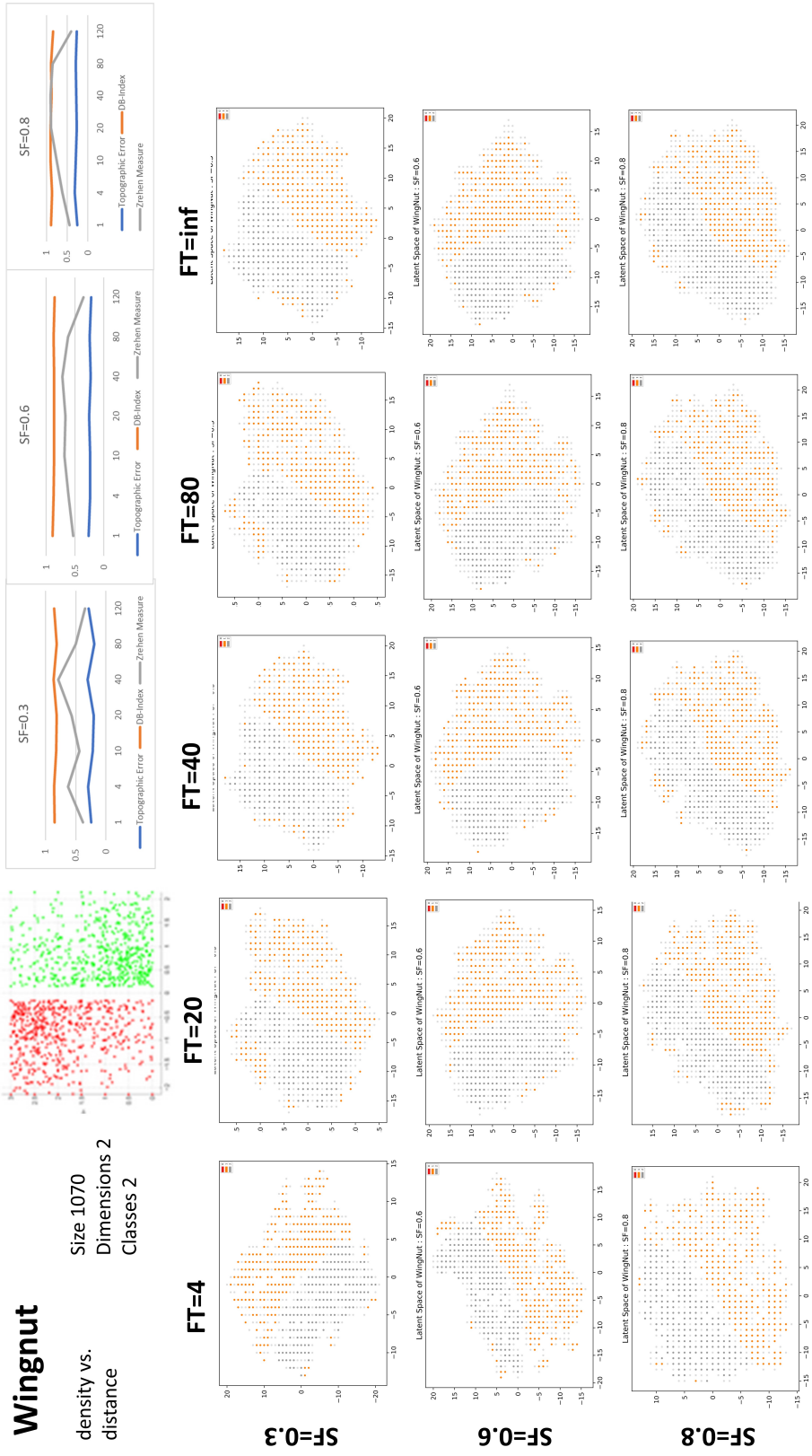


Fig. A.10 Topography Evaluation of WINGNUT dataset

A.2 Stroke Patient Profiling Insights Module

This section provides sample screenshots from the visualization tool developed for the Stroke Patient Profiling case-study presented in Section 6.2. The outcomes of the study were integrated into an interactive visualization platform which enables investigating data in a systematic way. Fig. A.11 (A) shows the outcomes of the GSOM based on the START data, whereas Fig. A.11 (B) shows the aggregated demographic details.

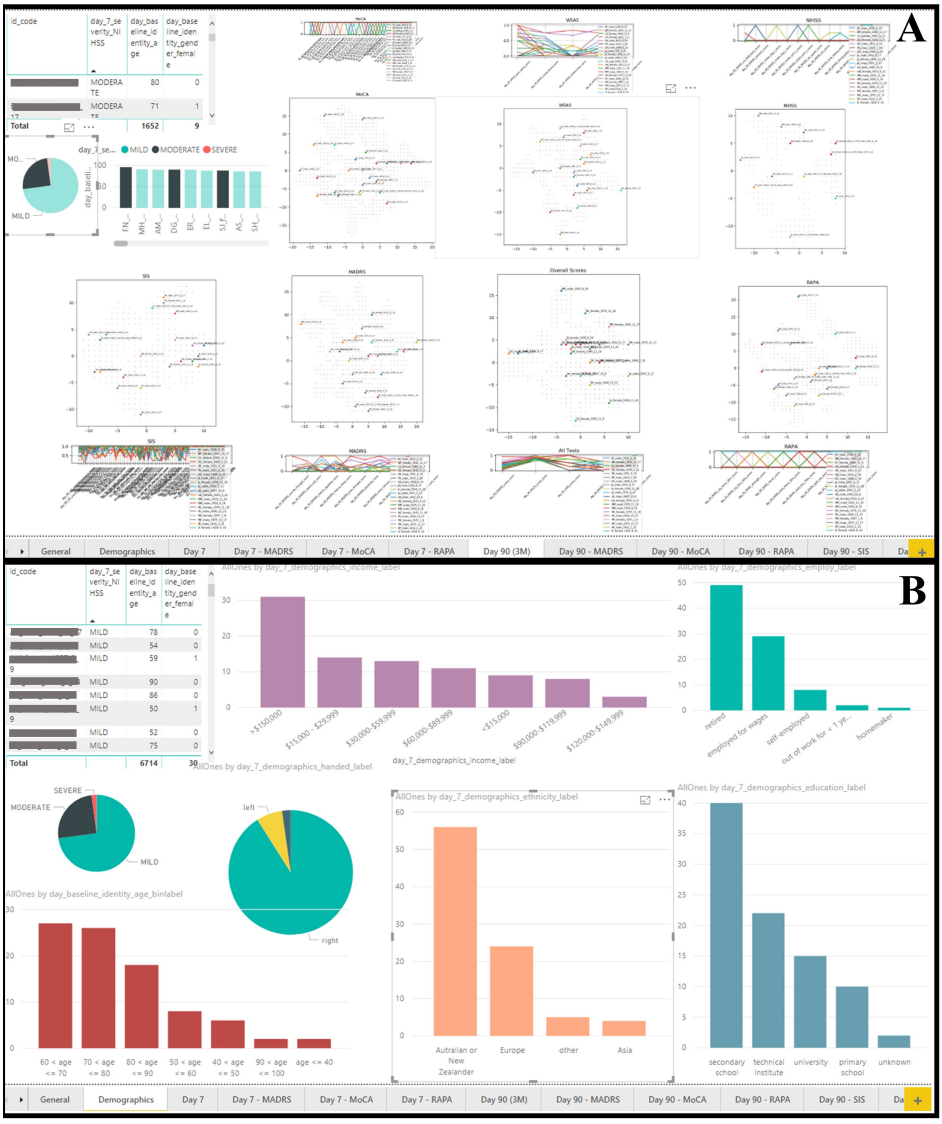


Fig. A.11 Screenshots of Stroke Patient Profiling Insights Module

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. *et al.* (2016) Tensorflow: A system for large-scale machine learning in: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* pp. 265–283

Ahmed, E., Yaqoob, I., Hashem, I.A.T., Khan, I., Ahmed, A.I.A., Imran, M. and Vasilakos, A.V. (2017) The role of big data analytics in internet of things *Computer Networks* **129**, pp. 459–471

Alahakoon, D., Halgamuge, S.K. and Srinivasan, B. (2000) Dynamic self-organizing maps with controlled growth for knowledge discovery *IEEE Transactions on neural networks* **11**(3), pp. 601–614

Alahakoon, D., Nawaratne, R., Xu, Y., De Silva, D., Sivarajah, U. and Gupta, B. (2020) Self-building artificial intelligence and machine learning to empower big data analytics in smart cities *Information Systems Frontiers* pp. 1–20

Aljundi, R. (2019) Continual learning in neural networks *arXiv preprint arXiv:1910.02718*

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B.C., Awwal, A.A.S. and Asari, V.K. (2018) The history began from alexnet: A comprehensive survey on deep learning approaches *arXiv preprint arXiv:1803.01164*

Amarasiri, R., Alahakoon, D., Premaratne, M. and Smith, K. (2005) Enhancing clustering performance of feature maps using randomness

Andreakis, A., Hoyningen-Huene, N.v. and Beetz, M. (2009) Incremental unsupervised time series analysis using merge growing neural gas in: *International Workshop on Self-Organizing Maps* pp. 10–18 Springer

Anwander, A., Tittgemeyer, M., von Cramon, D.Y., Friederici, A.D. and Knösche, T.R. (2007) Connectivity-based parcellation of broca's area *Cerebral cortex* **17**(4), pp. 816–825

Anyoha, R. (2017) The history of artificial intelligence

Anzola, D., Barbrook-Johnson, P. and Cano, J.I. (2017) Self-organization and social science *Computational and Mathematical Organization Theory* **23**(2), pp. 221–257

Appleyard, B. (2011) *The brain is wider than the sky: why simple solutions don't work in a complex world* Hachette UK

Azevedo, F.A., Carvalho, L.R., Grinberg, L.T., Farfel, J.M., Ferretti, R.E., Leite, R.E., Filho, W.J., Lent, R. and Herculano-Houzel, S. (2009) Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain *Journal of Comparative Neurology* **513**(5), pp. 532–541

Baldi, P. (2012) Autoencoders, unsupervised learning, and deep architectures in: *Proceedings of ICML workshop on unsupervised and transfer learning* pp. 37–49

Baoyun, W. (2009) Review on internet of things [j] *Journal of electronic measurement and instrument* **12**, pp. 1–7

Barghout-Stein, L. (1999) *On differences between peripheral and foveal pattern masking* University of California, Berkeley

Barlow, H. (1979) Three theories of cortical function in: *Developmental Neurobiology of Vision* pp. 1–16 Springer

Barlow, H. (2001) The exploitation of regularities in the environment by the brain *Behavioral and Brain Sciences* **24**(4), pp. 602–607

Barlow, H.B. (1961) The coding of sensory messages *Current problems in animal behavior*

Barlow, H.B. (1983) Understanding natural vision in: *Physical and biological processing of images* pp. 2–14 Springer

Barros, P. (2017) Modeling affection mechanisms using deep and self-organizing neural networks Ph.D. Thesis University of Hamburg, Germany Hamburg, DE

Basavaraj, G. and Kusagur, A. (2017) Vision based surveillance system for detection of human fall in: *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* pp. 1516–1520 IEEE

Bassett, D.S. and Gazzaniga, M.S. (2011) Understanding complexity in the human brain *Trends in cognitive sciences* **15**(5), pp. 200–209

Beauchamp, M.S. (2015) The social mysteries of the superior temporal sulcus *Trends in cognitive sciences* **19**(9), pp. 489–490

Behrooz, K.P. and Behzad, K.P. (1993) Evaluation of quantization error in computer vision *Physics-Based Vision: Principles and Practice: Radiometry* **1**, p. 292

Bengio, Y., Courville, A. and Vincent, P. (2013) Representation learning: A review and new perspectives *IEEE transactions on pattern analysis and machine intelligence* **35**(8), pp. 1798–1828

Bergstra, J. and Bengio, Y. (2012) Random search for hyper-parameter optimization *Journal of machine learning research* **13**(Feb), pp. 281–305

Best, J. (2019) Everything you need to know about the future of healthcare

Betsch, B.Y., Einhäuser, W., Körding, K.P. and König, P. (2004) The world from a cat's perspective–statistics of natural videos *Biological cybernetics* **90**(1), pp. 41–50

Blank, M., Gorelick, L., Shechtman, E., Irani, M. and Basri, R. (2005) Actions as space-time shapes in: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* vol. 2 pp. 1395–1402 IEEE

Blog, G.D. (2017) How machine learning with tensorflow enabled mobile proof-of-purchase at coca-cola

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M.K. (1990) Occam's razor *Readings in machine learning* pp. 201–204

Boccaletti, S., Grebogi, C., Lai, Y.C., Mancini, H. and Maza, D. (2000) The control of chaos: theory and applications *Physics reports* **329**(3), pp. 103–197

Bond, A.H. (2004) An information-processing analysis of the functional architecture of the primate neocortex *Journal of Theoretical Biology* **227**(1), pp. 51–79

Born, R.T. and Bradley, D.C. (2005) Structure and function of visual area mt *Annu. Rev. Neurosci.* **28**, pp. 157–189

Bosworth, H. (2001) Health-related quality of life after stroke: A comprehensive review *Stroke: Journal of the American Heart Association* **32**(4)

Braddick, O.J., O'Brien, J.M., Wattam-Bell, J., Atkinson, J., Hartley, T. and Turner, R. (2001) Brain areas sensitive to coherent visual motion *Perception* **30**(1), pp. 61–72

Brains (2019) Brains

Bremner, A.J., Lewkowicz, D.J. and Spence, C. (2012) *Multisensory development* Oxford University Press

Brott, T., Adams Jr, H.P., Olinger, C.P., Marler, J.R., Barsan, W.G., Biller, J., Spilker, J., Holleran, R., Eberle, R. and Hertzberg, V. (1989) Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* **20**(7), pp. 864–870

Budd, S., Robinson, E.C. and Kainz, B. (2019) A survey on active learning and human-in-the-loop deep learning for medical image analysis *arXiv preprint arXiv:1910.02923*

Butt, A. (2019) Ai is the fourth industrial revolution technology

Bzdok, D. (2017) Classical statistics and statistical learning in imaging neuroscience *Frontiers in neuroscience* **11**, p. 543

Bzdok, D., Altman, N. and Krzywinski, M. (2018) Points of significance: statistics versus machine learning

Bzdok, D., Krzywinski, M. and Altman, N. (2017) Points of significance: machine learning: a primer

Camazine, S., Deneubourg, J.L., Franks, N.R., Sneyd, J., Bonabeau, E. and Theraula, G. (2003) *Self-organization in biological systems* Princeton university press

Capra, F. (1996) *The web of life: A new scientific understanding of living systems* Anchor

Cardullo, F., Sweet, B., Hosman, R. and Coon, C. (2011) The human visual system and its role in motion perception in: *AIAA Modeling and Simulation Technologies Conference* p. 6422

Carey, L.M., Crewther, S., Salvado, O., Lindén, T., Connelly, A., Wilson, W., Howells, D.W., Churilov, L., Ma, H., Tse, T. *et al.* (2015) Stroke imaging prevention and treatment (start): a longitudinal stroke cohort study: clinical trials protocol *International Journal of Stroke* **10**(4), pp. 636–644

Carlsson, G., Möller, A. and Blomstrand, C. (2004) A qualitative study of the consequences of'hidden dysfunctions' one year after a mild stroke in persons<? 75 years *Disability and rehabilitation* **26**(23), pp. 1373–1380

Carlsson, G.E., Möller, A. and Blomstrand, C. (2003) Consequences of mild stroke in persons *Cerebrovascular Diseases* **16**(4), pp. 383–388

Carlsson, G.E., Möller, A. and Blomstrand, C. (2009) Managing an everyday life of uncertainty–a qualitative study of coping in persons with mild stroke *Disability and rehabilitation* **31**(10), pp. 773–782

Carpenter, G.A. and Grossberg, S. (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine *Computer vision, graphics, and image processing* **37**(1), pp. 54–115

Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly detection: A survey *ACM computing surveys (CSUR)* **41**(3), pp. 1–58

Chappell, G.J. and Taylor, J.G. (1993) The temporal kohønen map *Neural networks* **6**(3), pp. 441–445

Chatfield, K., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) Return of the devil in the details: Delving deep into convolutional nets *arXiv preprint arXiv:1405.3531*

Chaudhry, R., Ravichandran, A., Hager, G. and Vidal, R. (2009) Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions in: *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp. 1932–1939 IEEE

Chen, Z. and Liu, B. (2016) Lifelong machine learning *Synthesis Lectures on Artificial Intelligence and Machine Learning* **10**(3), pp. 1–145

Chen, Z. and Liu, B. (2018) Lifelong machine learning *Synthesis Lectures on Artificial Intelligence and Machine Learning* **12**(3), pp. 1–207

Chew, L.P. (1989) Constrained delaunay triangulations *Algorithmica* **4**(1-4), pp. 97–108

Choi, E., Lee, K. and Choi, K. (2019) Autoencoder-based incremental class learning without retraining on old data *arXiv preprint arXiv:1907.07872*

Cohen, B. and Muñoz, P. (2016) Sharing cities and sustainable consumption and production: towards an integrated framework *Journal of cleaner production* **134**, pp. 87–97

Commons, W. (2015) A simplified schema of the human visual pathway

Craik, K.J.W. (1952) *The nature of explanation* vol. 445 CUP Archive

Creutzfeldt, O.D. (1977) Generality of the functional structure of the neocortex *Naturwissenschaften* **64**(10), pp. 507–517

Cziko, G.A. (1989) Unpredictability and indeterminism in human behavior: Arguments and implications for educational research *Educational researcher* **18**(3), pp. 17–25

Dalal, N. and Triggs, B. (2005a) Histograms of oriented gradients for human detection in: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* vol. 1 pp. 886–893 vol. 1

Dalal, N. and Triggs, B. (2005b) Histograms of oriented gradients for human detection in: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* vol. 1 pp. 886–893 IEEE

Danielsson, P.E. (1980) Euclidean distance mapping *Computer Graphics and image processing* **14**(3), pp. 227–248

De Silva, D. and Alahakoon, D. (2010) Incremental knowledge acquisition and self learning from text in: *The 2010 International Joint Conference on Neural Networks (IJCNN)* pp. 1–8 IEEE

De Silva, D., Yu, X., Alahakoon, D. and Holmes, G. (2011) Incremental pattern characterization learning and forecasting for electricity consumption using smart meters in: *2011 IEEE International Symposium on Industrial Electronics* pp. 807–812 IEEE

DeGraba, T.J., Hallenbeck, J.M., Pettigrew, K.D., Dutka, A.J. and Kelly, B.J. (1999) Progression in acute stroke: value of the initial nih stroke scale score on patient stratification in future trials *Stroke* **30**(6), pp. 1208–1212

Dimitrov, V. (2003) *A new kind of social science: Study of self-organization of human dynamics* Lulu. com

Doyle, P.J. (2002) Measuring health outcomes in stroke survivors *Archives of physical medicine and rehabilitation* **83**, pp. S39–S43

Draelos, T.J., Miner, N.E., Lamb, C.C., Cox, J.A., Vineyard, C.M., Carlson, K.D., Severa, W.M., James, C.D. and Aimone, J.B. (2017) Neurogenesis deep learning: Extending deep networks to accommodate new classes in: *2017 International Joint Conference on Neural Networks (IJCNN)* pp. 526–533 IEEE

Dua, D. and Graff, C. (2017) UCI machine learning repository

Duncan, P.W., Bode, R.K., Lai, S.M., Perera, S., in Neuroprotection Americas Investigators, G.A. *et al.* (2003) Rasch analysis of a new stroke-specific outcome scale: the stroke impact scale *Archives of physical medicine and rehabilitation* **84**(7), pp. 950–963

Duncan, P.W., Samsa, G.P., Weinberger, M., Goldstein, L.B., Bonito, A., Witter, D.M., Enarson, C. and Matchar, D. (1997) Health status of individuals with mild stroke *Stroke* **28**(4), pp. 740–745

Duncan, P.W., Wallace, D., Lai, S.M., Johnson, D., Embretson, S. and Laster, L.J. (1999) The stroke impact scale version 2.0: evaluation of reliability, validity, and sensitivity to change *Stroke* **30**(10), pp. 2131–2140

Education, I.C. (2020) What is machine learning?

Edwards, D.F., Hahn, M., Baum, C. and Dromerick, A.W. (2006) The impact of mild stroke on meaningful activity and life satisfaction *Journal of stroke and cerebrovascular diseases* **15**(4), pp. 151–157

Elman, J.L. (1990) Finding structure in time *Cognitive science* **14**(2), pp. 179–211

Epshtein, B., Ofek, E. and Wexler, Y. (2010) Detecting text in natural scenes with stroke width transform in: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp. 2963–2970 IEEE

Felleman, D.J. and Van, D.E. (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)* **1**(1), pp. 1–47

Field, D.J. (1987) Relations between the statistics of natural images and the response properties of cortical cells *Josa a* **4**(12), pp. 2379–2394

Foote, K.D. (2016) A brief history of the internet of things

Fourtané, S. (2018) Connected vehicles in smart cities: The future of transportation

Frankland, P.W. and Bontempi, B. (2005) The organization of recent and remote memories *Nature Reviews Neuroscience* **6**(2), pp. 119–130

Fraser, C., Power, M., Hamdy, S., Rothwell, J., Hobday, D., Hollander, I., Tyrell, P., Hobson, A., Williams, S. and Thompson, D. (2002) Driving plasticity in human adult motor cortex is associated with improved motor function after brain injury *Neuron* **34**(5), pp. 831–840

Fratto, N. (2018) Machine un-learning: Why forgetting might be the key to ai

French, R.M. (1999) Catastrophic forgetting in connectionist networks *Trends in cognitive sciences* **3**(4), pp. 128–135

Fritzke, B. (1991) Let it grow-self-organizing feature maps with problem dependent cell structure in: *Artificial neural networks* Citeseer

Fritzke, B. (1994) Growing cell structures—a self-organizing network for unsupervised and supervised learning *Neural networks* **7**(9), pp. 1441–1460

Fritzke, B. (1995) A growing neural gas network learns topologies in: *Advances in neural information processing systems* pp. 625–632

García, J., Gardel, A., Bravo, I. and Lázaro, J.L. (2014) Multiple view oriented matching algorithm for people reidentification *IEEE Transactions on Industrial Informatics* **10**(3), pp. 1841–1851

Gartner (2017) Gartner says 8.4 billion connected "things" will be in use in 2017, up 31 percent from 2016

Gepperth, A. and Karaoguz, C. (2016) A bio-inspired incremental learning architecture for applied perceptual problems *Cognitive Computation* **8**(5), pp. 924–934

Goodale, M.A., Milner, A.D. *et al.* (1992) Separate visual pathways for perception and action *Trends in Neurosciences*

Granter, S.R., Beck, A.H. and Papke Jr, D.J. (2017) Alphago, deep learning, and the future of the human microscopist *Archives of pathology & laboratory medicine* **141**(5), pp. 619–621

Green, D.G., Sadedin, S. and Leishman, T.G. (2008) Self-organization in: *Encyclopedia of Ecology* pp. 3195–3203 Elsevier

Green, T.L. and King, K.M. (2010) Functional and psychosocial outcomes 1 year after mild stroke *Journal of Stroke and Cerebrovascular Diseases* **19**(1), pp. 10–16

Grossman, E. and Blake, R. (2001) Brain activity evoked by inverted and imagined biological motion *Vision research* **41**(10-11), pp. 1475–1482

Groves, P.M. and Thompson, R.F. (1970) Habituation: a dual-process theory. *Psychological review* **77**(5), p. 419

Gubbi, J., Buyya, R., Marusic, S. and Palaniswami, M. (2013) Internet of things (iot): A vision, architectural elements, and future directions *Future generation computer systems* **29**(7), pp. 1645–1660

Guelzim, T. and Obaidat, M.S. (2016) Cloud computing systems for smart cities and homes in: *Smart Cities and Homes* pp. 241–260 Morgan Kaufmann, Boston

Hadsell, R., Rao, D., Rusu, A.A. and Pascanu, R. (2020) Embracing change: Continual learning in deep neural networks *Trends in Cognitive Sciences*

Hamker, F.H. (2001) Life-long learning cell structures—continuously learning without catastrophic interference *Neural Networks* **14**(4-5), pp. 551–573

Hand, B., Page, S.J. and White, S. (2014) Stroke survivors scoring zero on the nih stroke scale score still exhibit significant motor impairment and functional limitation *Stroke research and treatment* **2014**

Hart, J.K. and Martinez, K. (2015) Toward an environmental internet of things *Earth and Space Science* **2**(5), pp. 194–200

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K. and Davis, L.S. (2016) Learning temporal regularity in video sequences in: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 733–742

Hasan, M. and Roy-Chowdhury, A.K. (2014) Incremental activity modeling and recognition in streaming videos in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 796–803

Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E. and Chiroma, H. (2016) The role of big data in smart city *International Journal of Information Management* **36**(5), pp. 748–758

Hawkins, D.M. (2004) The problem of overfitting *Journal of chemical information and computer sciences* **44**(1), pp. 1–12

Hawkins, J., Ahmad, S. and Cui, Y. (2017) A theory of how columns in the neocortex enable learning the structure of the world *Frontiers in neural circuits* **11**, p. 81

Hawkins, J. and Blakeslee, S. (2007) *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines* Macmillan

Hawkins, J., George, D. and Niemasik, J. (2009) Sequence memory for prediction, inference and behaviour *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**(1521), pp. 1203–1209

Hawkins, J., Lewis, M., Klukas, M., Purdy, S. and Ahmad, S. (2019) A framework for intelligence and cortical function based on grid cells in the neocortex *Frontiers in neural circuits* **12**, p. 121

Haxby, J.V., Hoffman, E.A. and Gobbini, M.I. (2000) The distributed human neural system for face perception *Trends in cognitive sciences* **4**(6), pp. 223–233

Heath, N. (2018) Tesla's autopilot: Cheat sheet

Hebb, D.O. (1949) *The organization of behavior: a neuropsychological theory* J. Wiley; Chapman & Hall

Hein, G. and Knight, R.T. (2008) Superior temporal sulcus—it's my area: or is it? *Journal of cognitive neuroscience* **20**(12), pp. 2125–2136

Herculano-Houzel, S. (2009) The human brain in numbers: a linearly scaled-up primate brain *Frontiers in human neuroscience* **3**, p. 31

Herzog, M.H. and Clarke, A.M. (2014) Why vision is not both hierarchical and feedforward *Frontiers in computational neuroscience* **8**, p. 135

Hetherington, P. (1989) Is there'catastrophic interference'in connectionist networks? in: *Proceedings of the 11th annual conference of the cognitive science society* pp. 26–33 LEA

Hinkle, J.L. (2014) Reliability and validity of the national institutes of health stroke scale for neuroscience nurses *Stroke* **45**(3), pp. e32–e34

Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the knowledge in a neural network *arXiv preprint arXiv:1503.02531*

Hinton, G.E. and Van Camp, D. (1993) Keeping the neural networks simple by minimizing the description length of the weights in: *Proceedings of the sixth annual conference on Computational learning theory* pp. 5–13

Hirsch, H.V. and Spinelli, D. (1970) Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats *Science* **168**(3933), pp. 869–871

Ho, W.T., Lim, H.W. and Tay, Y.H. (2009) Two-stage license plate detection using gentle adaboost and sift-svm in: *2009 First Asian Conference on Intelligent Information and Database Systems* pp. 109–114 IEEE

Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory *Neural computation* **9**(8), pp. 1735–1780

Hofman, M.A. (2014) Evolution of the human brain: when bigger is better *Frontiers in neuroanatomy* **8**, p. 15

Hong, X., Guan, S.U., Man, K.L. and Wong, P.W. (2020) Lifelong machine learning architecture for classification *Symmetry* **12**(5), p. 852

Hubel, D.H. and Wiesel, T.N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex *The Journal of physiology* **160**(1), pp. 106–154

Ikezu, T. and Gendelman, H.E. (2016) *Neuroimmune pharmacology* Springer

James, W. (2007) *The principles of psychology* vol. 1 Cosimo, Inc.

Jayaratne, M., Alahakoon, D., De Silva, D. and Yu, X. (2018) Bio-inspired multisensory fusion for autonomous robots in: *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society* pp. 3090–3095 IEEE

Jha, P. and Patnaik, K.S. (2020) Self-driving cars: Role of machine learning in: *Handbook of Research on Emerging Trends and Applications of Machine Learning* pp. 490–507 IGI Global

Jung, M., Hwang, J. and Tani, J. (2015) Self-organization of spatio-temporal hierarchy via learning of dynamic visual image patterns on action sequences *PloS one* **10**(7)

Kahn, D. (2013) Brain basis of self: self-organization and lessons from dreaming *Frontiers in psychology* **4**, p. 408

Karlsson, M.P. and Frank, L.M. (2009) Awake replay of remote experiences in the hippocampus *Nature neuroscience* **12**(7), p. 913

Kasner, S.E. (2006) Clinical interpretation and use of stroke scales *The Lancet Neurology* **5**(7), pp. 603–612

Keller, G.B., Bonhoeffer, T. and Hübener, M. (2012) Sensorimotor mismatch signals in primary visual cortex of the behaving mouse *Neuron* **74**(5), pp. 809–815

Kemker, R. and Kanan, C. (2017) Fearnet: Brain-inspired model for incremental learning *arXiv preprint arXiv:1711.10563*

Khadartsev, A.A. and Eskov, V.M. (2014) Chaos theory and self-organization systems in recovery medicine: A scientific review *Integrative Medicine International* **1**(4), pp. 226–233

Khalilia, M. and Popescu, M. (2014) Topology preservation in fuzzy self-organizing maps in: *Advance Trends in Soft Computing* pp. 105–114 Springer

Kim, D., Lee, M. and Kwak, N. (2017) Matching video net: Memory-based embedding for video action recognition in: *2017 International Joint Conference on Neural Networks (IJCNN)* pp. 432–438 IEEE

Kim, J. and Grauman, K. (2009) Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates in: *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp. 2921–2928 IEEE

King, R.B. (1996) Quality of life after stroke *Stroke* **27**(9), pp. 1467–1472

Kiran, B.R., Thomas, D.M. and Parakkal, R. (2018) An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos *Journal of Imaging* **4**(2), p. 36

Kiritsis, D. (2011) Closed-loop plm for intelligent products in the era of the internet of things *Computer-Aided Design* **43**(5), pp. 479–501

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. *et al.* (2017) Overcoming catastrophic forgetting in neural networks *Proceedings of the national academy of sciences* **114**(13), pp. 3521–3526

Kiviluoto, K. (1996) Topology preservation in self-organizing maps in: *Proceedings of International Conference on Neural Networks (ICNN'96)* vol. 1 pp. 294–299 vol.1

Kohonen, T. (1982) Self-organized formation of topologically correct feature maps *Biological cybernetics* **43**(1), pp. 59–69

Kohonen, T. (1990) The self-organizing map *Proceedings of the IEEE* **78**(9), pp. 1464–1480

Konkel, L. (2018) The brain before birth: Using fmri to explore the secrets of fetal neurodevelopment

Konorski, J. (1948) Conditioned reflexes and neuron organization. *The American Journal of Psychotherapy*

Kosmopoulos, D.I., Voulodimos, A.S. and Doulamis, A.D. (2012) A system for multicamera task recognition and summarization for structured environments *IEEE Transactions on Industrial Informatics* **9**(1), pp. 161–171

Krizhevsky, A., Nair, V. and Hinton, G. (2014) The cifar-10 dataset

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet classification with deep convolutional neural networks in: *Advances in neural information processing systems* pp. 1097–1105

Krugman, P.R. and Krugman, P. (1996) *The self-organizing economy* vol. 122 Blackwell Oxford

Kumaran, D., Hassabis, D. and McClelland, J.L. (2016) What learning systems do intelligent agents need? complementary learning systems theory updated *Trends in cognitive sciences* **20**(7), pp. 512–534

Kurpiel, F.D., Minetto, R. and Nassu, B.T. (2017) Convolutional neural networks for license plate detection in images in: *2017 IEEE International Conference on Image Processing (ICIP)* pp. 3395–3399 IEEE

Lang, K.J., Waibel, A.H. and Hinton, G.E. (1990) A time-delay neural network architecture for isolated word recognition *Neural networks* **3**(1), pp. 23–43

LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning *nature* **521**(7553), pp. 436–444

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition *Proceedings of the IEEE* **86**(11), pp. 2278–2324

Lewis, J.W., Beauchamp, M.S. and DeYoe, E.A. (2000) A comparison of visual and auditory motion processing in human cerebral cortex *Cerebral Cortex* **10**(9), pp. 873–888

Lewkowicz, D.J. (2014) Early experience and multisensory perceptual narrowing *Developmental psychobiology* **56**(2), pp. 292–315

Li, Z. (2002) A saliency map in primary visual cortex *Trends in cognitive sciences* **6**(1), pp. 9–16

Liu, A.A., Su, Y.T., Nie, W.Z. and Kankanhalli, M. (2016) Hierarchical clustering multi-task learning for joint human action grouping and recognition *IEEE transactions on pattern analysis and machine intelligence* **39**(1), pp. 102–114

Liu, B. (2018) Natural intelligence the human factor in ai

Liu, H., Chen, S. and Kubota, N. (2013) Intelligent video systems and analytics: A survey *IEEE Transactions on Industrial Informatics* **9**(3), pp. 1222–1233

Liu, J., Luo, J. and Shah, M. (2009) Recognizing realistic actions from videos "in the wild" in: *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp. 1996–2003 IEEE

Lowel, S. and Singer, W. (1992) Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity *Science* **255**(5041), pp. 209–212

Lu, C., Shi, J. and Jia, J. (2013a) Abnormal Event Detection at 150 FPS in MATLAB in: *2013 IEEE International Conference on Computer Vision* pp. 2720–2727

Lu, C., Shi, J. and Jia, J. (2013b) Abnormal event detection at 150 fps in matlab in: *Proceedings of the IEEE international conference on computer vision* pp. 2720–2727

Lungarella, M. and Sporns, O. (2005) Information self-structuring: Key principle for learning and development in: *Proceedings. The 4th International Conference on Development and Learning, 2005* pp. 25–30 IEEE

Luo, W., Liu, W. and Gao, S. (2017) Remembering history with convolutional lstm for anomaly detection in: *2017 IEEE International Conference on Multimedia and Expo (ICME)* pp. 439–444 IEEE

Luria, A.R. and Solotaroff, L.T. (1987) *The mind of a mnemonist: A little book about a vast memory.* Harvard University Press

Luvizon, D.C., Nassu, B.T. and Minetto, R. (2016) A video-based system for vehicle speed measurement in urban roadways *IEEE Transactions on Intelligent Transportation Systems* **18**(6), pp. 1393–1404

Maaten, L.v.d. and Hinton, G. (2008) Visualizing data using t-sne *Journal of machine learning research* **9**(Nov), pp. 2579–2605

MacKay, D.J. and Mac Kay, D.J. (2003) *Information theory, inference and learning algorithms* Cambridge university press

Madjiheurem, S. and Toni, L. (2019) Representation learning on graphs: A reinforcement learning application in: *The 22nd International Conference on Artificial Intelligence and Statistics* pp. 3391–3399 PMLR

Mahadevan, V., Li, W., Bhalodia, V. and Vasconcelos, N. (2010a) Anomaly detection in crowded scenes in: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp. 1975–1981

Mahadevan, V., Li, W., Bhalodia, V. and Vasconcelos, N. (2010b) Anomaly detection in crowded scenes in: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp. 1975–1981 IEEE

Marsland, S., Shapiro, J. and Nehmzow, U. (2002) A self-organising network that grows when required *Neural networks* **15**(8-9), pp. 1041–1058

Martin-Schild, S., Albright, K.C., Tanksley, J., Pandav, V., Jones, E.B., Grotta, J.C. and Savitz, S.I. (2011) Zero on the nihss does not equal the absence of stroke *Annals of emergency medicine* **57**(1), pp. 42–45

Martinetz, T., Schulten, K. *et al.* (1991) A" neural-gas" network learns topologies *Artificial Neural Networks*

Matsugu, M., Mori, K., Mitari, Y. and Kaneda, Y. (2003) Subject independent facial expression recognition with robust face detection using a convolutional neural network *Neural Networks* **16**(5-6), pp. 555–559

McClelland, J.L., McNaughton, B.L. and O'Reilly, R.C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* **102**(3), p. 419

McCloskey, M. and Cohen, N.J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem in: *Psychology of learning and motivation* vol. 24 pp. 109–165 Elsevier

McLaren, D. and Agyeman, J. (2015) *Sharing cities: a case for truly smart and sustainable cities* MIT press

Mehran, R., Oyama, A. and Shah, M. (2009) Abnormal crowd behavior detection using social force model in: *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp. 935–942 IEEE

Mermillod, M., Bugaiska, A. and Bonin, P. (2013) The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects *Frontiers in psychology* **4**, p. 504

Mici, L., Parisi, G.I. and Wermter, S. (2018) A self-organizing neural network architecture for learning human-object interactions *Neurocomputing* **307**, pp. 14–24

Miller, G.A. (1995) Wordnet: a lexical database for english *Communications of the ACM* **38**(11), pp. 39–41

Mind, S.A. (2012) Edges of perception *SCIENTIFIC AMERICAN MIND* p. 46–53

Minetto, R., Thome, N., Cord, M., Leite, N.J. and Stolfi, J. (2014) Snoopertext: A text detection system for automatic indexing of urban scenes *Computer Vision and Image Understanding* **122**, pp. 92–104

Molnár, Z. and Pollen, A. (2014) How unique is the human neocortex? *Development* **141**(1), pp. 11–16

Montazzolli Silva, S. and Rosito Jung, C. (2018) License plate detection and recognition in unconstrained scenarios in: *Proceedings of the European Conference on Computer Vision (ECCV)* pp. 580–596

Montgomery, S.A. and Åsberg, M. (1979) A new depression scale designed to be sensitive to change *The British journal of psychiatry* **134**(4), pp. 382–389

Mountcastle, V. (1978) An organizing principle for cerebral function: the unit module and the distributed system *The mindful brain*

Mountcastle, V.B. (1957) Modality and topographic properties of single neurons of cat's somatic sensory cortex *Journal of neurophysiology* **20**(4), pp. 408–434

Mundt, J.C., Marks, I.M., Shear, M.K. and Greist, J.M. (2002) The work and social adjustment scale: a simple measure of impairment in functioning *The British Journal of Psychiatry* **180**(5), pp. 461–464

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E. (2015) Deep learning applications and challenges in big data analytics *Journal of Big Data* **2**(1), p. 1

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L. and Chertkow, H. (2005) The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment *Journal of the American Geriatrics Society* **53**(4), pp. 695–699

Nawaratne, R., Adikari, A., Alahakoon, D. and Carey, L. (N.D.) "is mild really mild?": Patient profiling using artificial intelligence unpublished

Nawaratne, R., Adikari, A., Alahakoon, D., De Silva, D. and Chilamkurti, N. (2020a) Recurrent self-structuring machine learning for video processing using multi-stream hierarchical growing self-organizing maps *Multimedia Tools and Applications*

Nawaratne, R., Alahakoon, D., De Silva, D., Chhetri, P. and Chilamkurti, N. (2018) Self-evolving intelligent algorithms for facilitating data interoperability in iot environments *Future Generation Computer Systems* **86**, pp. 421–432

Nawaratne, R., Alahakoon, D., De Silva, D., Kumara, H. and Yu, X. (2019a) Hierarchical two-stream growing self-organizing maps with transience for human activity recognition *IEEE Transactions on Industrial Informatics*

Nawaratne, R., Alahakoon, D., De Silva, D. and Yu, X. (2019b) Ht-gsom: dynamic self-organizing map with transience for human activity recognition in: *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)* vol. 1 pp. 270–273 IEEE

Nawaratne, R., Alahakoon, D., De Silva, D. and Yu, X. (2019c) Spatiotemporal anomaly detection using deep learning for real-time video surveillance *IEEE Transactions on Industrial Informatics* **16**(1), pp. 393–402

Nawaratne, R., Bandaragoda, T., Adikari, A., Alahakoon, D., De Silva, D. and Yu, X. (2017) Incremental knowledge acquisition and self-learning for autonomous video surveillance in: *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society* pp. 4790–4795 IEEE

Nawaratne, R., De Silva, D. and Nguyen, S. (2020b) A generative latent space approach for real-time smart city surveillance

Neftci, E.O. and Averbeck, B.B. (2019) Reinforcement learning in artificial and biological systems *Nature Machine Intelligence* **1**(3), pp. 133–143

Nelson, C.A. (2000) Neural plasticity and human development: The role of early experience in sculpting memory systems *Developmental Science* **3**(2), pp. 115–136

Nitta, A., Hayashi, K., Hasegawa, T. and Nabeshima, T. (1993) Development of plasticity of brain function with repeated trainings and passage of time after basal forebrain lesions in rats *Journal of Neural Transmission/General Section JNT* **93**(1), pp. 37–46

Oh, J., Guo, X., Lee, H., Lewis, R.L. and Singh, S. (2015) Action-conditional video prediction using deep networks in atari games in: *Advances in neural information processing systems* pp. 2863–2871

Olshausen, B.A. and Field, D.J. (1996) Natural image statistics and efficient coding *Network: computation in neural systems* **7**(2), pp. 333–339

Ordóñez, F.J. and Roggen, D. (2016) Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition *Sensors* **16**(1), p. 115

Ortiz, G.A. and L. Sacco, R. (2014) National institutes of health stroke scale (nihss) *Wiley StatsRef: Statistics Reference Online*

Osuwa, A.A., Ekhoragbon, E.B. and Fat, L.T. (2017) Application of artificial intelligence in internet of things in: *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)* pp. 169–173 IEEE

O'Neill, J., Pleydell-Bouverie, B., Dupret, D. and Csicsvari, J. (2010) Play it again: reactivation of waking experience and memory *Trends in neurosciences* **33**(5), pp. 220–229

O'Reilly, R.C., Bhattacharyya, R., Howard, M.D. and Ketz, N. (2014) Complementary learning systems *Cognitive science* **38**(6), pp. 1229–1248

Pan, S.J. and Yang, Q. (2009) A survey on transfer learning *IEEE Transactions on knowledge and data engineering* **22**(10), pp. 1345–1359

Parisi, G.I. (2017) Multimodal learning of actions with deep neural network self-organization Ph.D. Thesis University of Hamburg, Germany Hamburg, DE

Parisi, G.I., Kemker, R., Part, J.L., Kanan, C. and Wermter, S. (2019) Continual lifelong learning with neural networks: A review *Neural Networks*

Parisi, G.I., Magg, S. and Wermter, S. (2016) Human motion assessment in real time using recurrent self-organization in: *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* pp. 71–76 IEEE

Parisi, G.I., Tani, J., Weber, C. and Wermter, S. (2018) Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization *Frontiers in neurorobotics* **12**, p. 78

Parisi, G.I., Weber, C. and Wermter, S. (2015) Self-organizing neural integration of pose-motion features for human action recognition *Frontiers in neurorobotics* **9**, p. 3

Peng, B., Lei, J., Fu, H., Zhang, C., Chua, T.S. and Li, X. (2018) Unsupervised video action clustering via motion-scene interaction constraint *IEEE Transactions on Circuits and Systems for Video Technology*

Polishetty, R., Roopaei, M. and Rad, P. (2016) A next-generation secure cloud-based deep learning license plate recognition for smart cities in: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* pp. 286–293 IEEE

Qiu, F.T. and Von Der Heydt, R. (2005) Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules *Neuron* **47**(1), pp. 155–166

Quadrato, G., Elnaggar, M.Y. and Di Giovanni, S. (2014) Adult neurogenesis in brain repair: cellular plasticity vs. cellular replacement *Frontiers in neuroscience* **8**, p. 17

Quiroga, R.Q. (2017) *The Forgetting Machine: Memory, Perception, and the" Jennifer Aniston Neuron"* Benbella Books

Rabinowitz, N.C., Perbet, F., Song, H.F., Zhang, C., Eslami, S. and Botvinick, M. (2018) Machine theory of mind *arXiv preprint arXiv:1802.07740*

Rao, R.P. and Ballard, D.H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects *Nature neuroscience* **2**(1), pp. 79–87

Rebuffi, S.A., Kolesnikov, A., Sperl, G. and Lampert, C.H. (2017) icarl: Incremental classifier and representation learning in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* pp. 2001–2010

Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster r-cnn: Towards real-time object detection with region proposal networks in: *Advances in neural information processing systems* pp. 91–99

Richards, B.A. and Frankland, P.W. (2017) The persistence and transience of memory *Neuron* **94**(6), pp. 1071–1084

Riemer, M., Klinger, T., Bouneffouf, D. and Franceschini, M. (2019) Scalable recollections for continual lifelong learning in: *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 33 pp. 1352–1359

Robins, A. (1993) Catastrophic forgetting in neural networks: the role of rehearsal mechanisms in: *Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems* pp. 65–68 IEEE

Robins, A. (1995) Catastrophic forgetting, rehearsal and pseudorehearsal *Connection Science* **7**(2), pp. 123–146

Rodriguez, A., Whitson, J. and Granger, R. (2004) Derivation and analysis of basic computational operations of thalamocortical circuits *Journal of cognitive neuroscience* **16**(5), pp. 856–877

Rolls, E.T. (2008) Memory, attention, and decision-making *Chapter 2. OUP*

Ruder, S. (2016) An overview of gradient descent optimization algorithms *arXiv preprint arXiv:1609.04747*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. *et al.* (2015) Imagenet large scale visual recognition challenge *International journal of computer vision* **115**(3), pp. 211–252

Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R. and Hadsell, R. (2016) Progressive neural networks *arXiv preprint arXiv:1606.04671*

Said, O. and Masud, M. (2013) Towards internet of things: Survey and future vision *International Journal of Computer Networks* **5**(1), pp. 1–17

Sargano, A.B., Angelov, P. and Habib, Z. (2017) A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition *applied sciences* **7**(1), p. 110

Schacter, D.L., Addis, D.R. and Buckner, R.L. (2007) Remembering the past to imagine the future: the prospective brain *Nature reviews neuroscience* **8**(9), pp. 657–661

Schmid, M.C., Schmiedt, J.T., Peters, A.J., Saunders, R.C., Maier, A. and Leopold, D.A. (2013) Motion-sensitive responses in visual area v4 in the absence of primary visual cortex *Journal of Neuroscience* **33**(48), pp. 18740–18745

Schoenemann, P.T. (2006) Evolution of the size and functional areas of the human brain *Annu. Rev. Anthropol.* **35**, pp. 379–406

Schuldt, C., Laptev, I. and Caputo, B. (2004) Recognizing human actions: a local svm approach in: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* vol. 3 pp. 32–36 IEEE

Schweitzer, F. (1997) *Self-organization of complex structures: From individual to collective dynamics* CRC Press

Sekeres, M.J., Bonasia, K., St-Laurent, M., Pishdadian, S., Winocur, G., Grady, C. and Moscovitch, M. (2016) Recovering and preventing loss of detailed memory: differential rates of forgetting for detail types in episodic memory *Learning & Memory* **23**(2), pp. 72–82

Selmi, Z., Halima, M.B. and Alimi, A.M. (2017) Deep learning system for automatic license plate detection and recognition in: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* vol. 1 pp. 1132–1138 IEEE

Seltzer, B. and Pandya, D.N. (1980) Converging visual and somatic sensory cortical input to the intraparietal sulcus of the rhesus monkey *Brain research* **192**(2), pp. 339–351

Shalev-Shwartz, S. *et al.* (2011) Online learning and online convex optimization *Foundations and trends in Machine Learning* **4**(2), pp. 107–194

Shatz, C.J. (1992) The developing brain *Scientific American* **267**(3), pp. 60–67

Shepherd, G.M. (1975) Axons, dendrites and synapses in: *Membranes, Ions, and Impulses* pp. 165–170 Springer

Shreyas, V., Bharadwaj, S.N., Srinidhi, S., Ankith, K. and Rajendra, A. (2020) Self-driving cars: An overview of various autonomous driving systems in: *Advances in Data and Information Sciences* pp. 361–371 Springer

Simonyan, K. and Zisserman, A. (2014a) Two-stream convolutional networks for action recognition in videos in: *Advances in neural information processing systems* pp. 568–576

Simonyan, K. and Zisserman, A. (2014b) Very deep convolutional networks for large-scale image recognition *arXiv preprint arXiv:1409.1556*

Singer, W. (1986) The brain as a self-organizing system *European archives of psychiatry and neurological sciences* **236**(1), pp. 4–9

Skår, J. (2003) Introduction: Self–organization as an actual theme *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **361**(1807), pp. 1049–1056

Solutions, L.S.H. (2019) The state of the smart home market

Sorabji, R. (1974) Body and soul in aristotle *Philosophy* **49**(187), pp. 63–89

Sours, C., Raghavan, P., Foxworthy, W.A., Meredith, M.A., El Metwally, D., Zhuo, J., Gilmore, J.H., Medina, A.E. and Gullapalli, R.P. (2017) Cortical multisensory connectivity is present near birth in humans *Brain imaging and behavior* **11**(4), pp. 1207–1213

Spilker, J., Kongable, G., Barch, C., Braimah, J., Bratina, P., Daley, S., Donnarumma, R., Rapp, K. and Sailor, S. (1997) Using the nih stroke scale to assess stroke patients *Journal of Neuroscience Nursing* **29**(6), pp. 384–393

Spoms, O. and Pegors, T. (2003) Generating structure in sensory data through coordinated motor activity in: *Proceedings of the International Joint Conference on Neural Networks, 2003.* vol. 4 pp. 2796–vol IEEE

Sporns, O. and Pegors, T.K. (2004) Information-theoretical aspects of embodied artificial intelligence in: *Embodied artificial intelligence* pp. 74–85 Springer

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting *The journal of machine learning research* **15**(1), pp. 1929–1958

Stein, B.E. and Meredith, M.A. (1993) *The merging of the senses.* The MIT Press

Stein, B.E., Stanford, T.R. and Rowland, B.A. (2014) Development of multisensory integration from the perspective of the individual neuron *Nature Reviews Neuroscience* **15**(8), pp. 520–535

Stiehler, A. (2018) The fourth industrial revolution

Stooke, A., Lee, K., Abbeel, P. and Laskin, M. (2020) Decoupling representation learning from reinforcement learning *arXiv preprint arXiv:2009.08319*

Strickert, M. and Hammer, B. (2005) Merge som for temporal data *Neurocomputing* **64**, pp. 39–71

Su, J., Vargas, D.V. and Sakurai, K. (2019) One pixel attack for fooling deep neural networks *IEEE Transactions on Evolutionary Computation* **23**(5), pp. 828–841

Sur, M. and Leamey, C.A. (2001) Development and plasticity of cortical areas and networks *Nature Reviews Neuroscience* **2**(4), pp. 251–262

Tan Min, L. (2016) Artificial intelligence: The next frontier

Tang, C., Bian, M., Liu, X., Li, M., Zhou, H., Wang, P. and Yin, H. (2019) Unsupervised feature selection via latent representation learning and manifold regularization *Neural Networks* **117**, pp. 163–178

Tarapore, D., Lungarella, M. and Gómez, G. (2006) Quantifying patterns of agent–environment interaction *Robotics and Autonomous Systems* **54**(2), pp. 150–158

Taupin, P. and Gage, F.H. (2002) Adult neurogenesis and neural stem cells of the central nervous system in mammals *Journal of neuroscience research* **69**(6), pp. 745–749

Tishby, N. (2017) New theory cracks open the black box of deep learning *Quantum Magazine*

Tole, A.A. *et al.* (2013) Big data challenges *Database systems journal* **4**(3), pp. 31–40

Tononi, G. and Cirelli, C. (2014) Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration *Neuron* **81**(1), pp. 12–34

Topolski, T.D., LoGerfo, J., Patrick, D.L., Williams, B., Walwick, J. and Patrick, M.M.B. (2006) The rapid assessment of physical activity (rapa) among older adults *Preventing chronic disease* **3**(4)

Touvron, H., Vedaldi, A., Douze, M. and Jégou, H. (2020) Fixing the train-test resolution discrepancy: Fixefficientnet *arXiv preprint arXiv:2003.08237*

Tudor Ionescu, R., Smeureanu, S., Alexe, B. and Popescu, M. (2017) Unmasking the abnormal events in video in: *Proceedings of the IEEE International Conference on Computer Vision* pp. 2895–2903

Turing, A.M. (1936) On computable numbers, with an application to the entscheidungsproblem *J. of Math* **58**(345-363), p. 5

Turing, A.M. (2004) Computing machinery and intelligence (1950) *The Essential Turing: The Ideas that Gave Birth to the Computer Age. Ed. B. Jack Copeland. Oxford: Oxford UP* pp. 433–64

Ultsch, A. (2005) Clustering with som fundamental Clustering Problems Suite, https://www.uni-marburg.de/fb12/arbeitsgruppen/datenbionik/data?language_sync=1

Umakanthan, S. (2016) Human action recognition from video sequences Ph.D. Thesis Queensland University of Technology

Uylings, H.B. (2006) Development of the human cortex and the concept of "critical" or "sensitive" periods *Language Learning* **56**, pp. 59–90

van de Ven, G.M., Siegelmann, H.T. and Tolias, A.S. (2020) Brain-inspired replay for continual learning with artificial neural networks *Nature communications* **11**(1), pp. 1–14

Voegtlin, T. (2002) Recursive self-organizing maps *Neural networks* **15**(8-9), pp. 979–991

Vu, H. (2017) Deep abnormality detection in video data. in: *IJCAI* pp. 5217–5218

Wang, J. and Fan, S. (2018) Thinking like a human: What it means to give ai a theory of mind

Wang, L. and Sng, D. (2015) Deep learning algorithms with applications to video analytics for a smart city: A survey *arXiv preprint arXiv:1512.03131*

Wang, T. and Snoussi, H. (2012) Histograms of Optical Flow Orientation for Visual Abnormal Events Detection in: *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance* pp. 13–18

Webber, C.S. (1991) Competitive learning, natural images and cortical cells *Network: Computation in Neural Systems* **2**(2), pp. 169–187

Wikenheiser, A.M. and Redish, A.D. (2015) Decoding the cognitive map: ensemble hippocampal sequences and decision making *Current opinion in neurobiology* **32**, pp. 8–15

Williams, J.B. and Kobak, K.A. (2008) Development and reliability of a structured interview guide for the montgomery-åsberg depression rating scale (sigma) *The British Journal of Psychiatry* **192**(1), pp. 52–58

Williams, L.S., Weinberger, M., Harris, L.E., Clark, D.O. and Biller, J. (1999) Development of a stroke-specific quality of life scale *Stroke* **30**(7), pp. 1362–1369

Wilson, J.L., Hareendran, A., Grant, M., Baird, T., Schulz, U.G., Muir, K.W. and Bone, I. (2002) Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified rankin scale *Stroke* **33**(9), pp. 2243–2246

Winocur, G., Moscovitch, M. and Bontempi, B. (2010) Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions *Neuropsychologia* **48**(8), pp. 2339–2356

Wongthongtham, P., Kaur, J., Potdar, V. and Das, A. (2017) Big data challenges for the internet of things (iot) paradigm in: *Connected Environments for the Internet of Things* pp. 41–62 Springer

Wu, S., Moore, B.E. and Shah, M. (2010) Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes in: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp. 2054–2060 IEEE

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Zhang, Z. and Fu, Y. (2018) Incremental classifier learning with generative adversarial networks *arXiv preprint arXiv:1802.00853*

Xie, L., Ahmad, T., Jin, L., Liu, Y. and Zhang, S. (2018) A new cnn-based method for multi-directional car license plate detection *IEEE Transactions on Intelligent Transportation Systems* **19**(2), pp. 507–517

Xie, S. and Guan, Y. (2016) Motion instability based unsupervised online abnormal behaviors detection *Multimedia Tools and Applications* **75**(12), pp. 7423–7444

Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K. and Woo, W.c. (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting in: *Advances in neural information processing systems* pp. 802–810

Xu, D., Yan, Y., Ricci, E. and Sebe, N. (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion *Computer Vision and Image Understanding* **156**, pp. 117–127

Yang, Y., Saleemi, I. and Shah, M. (2012) Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions *IEEE transactions on pattern analysis and machine intelligence* **35**(7), pp. 1635–1648

Yarrow, S., Razak, K.A., Seitz, A.R. and Seriès, P. (2014) Detecting and Quantifying Topography in Neural Maps *PLOS ONE* **9**(2), p. e87178

Zaharescu, A. and Wildes, R. (2010) Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing in: *European Conference on Computer Vision* pp. 563–576 Springer

Zahra, D., Qureshi, A., Henley, W., Taylor, R., Quinn, C., Pooler, J., Hardy, G., Newbold, A. and Byng, R. (2014) The work and social adjustment scale: reliability, sensitivity and value *International Journal of Psychiatry in Clinical Practice* **18**(2), pp. 131–138

Zha, S., Luisier, F., Andrews, W., Srivastava, N. and Salakhutdinov, R. (2015) Exploiting image-trained cnn architectures for unconstrained video classification *arXiv preprint arXiv:1503.04144*

Zhang, W., Yang, D. and Wang, H. (2019) Data-driven methods for predictive maintenance of industrial equipment: A survey *IEEE Systems Journal* **13**(3), pp. 2213–2227

Zhang, Y. and Yang, Q. (2017) A survey on multi-task learning *arXiv preprint arXiv:1707.08114*

Zhao, B., Fei-Fei, L. and Xing, E.P. (2011) Online detection of unusual events in videos via dynamic sparse coding in: *CVPR 2011* pp. 3313–3320 IEEE

Zhu, H. (2020) Generalized representation learning methods for deep reinforcement learning pp. 5216–5217

Zou, Y., Shi, Y., Wang, Y., Shu, Y., Yuan, Q. and Tian, Y. (2018) Hierarchical temporal memory enhanced one-shot distance learning for action recognition in: *2018 IEEE International Conference on Multimedia and Expo (ICME)* pp. 1–6 IEEE

Zrehen, S. (1993) Analyzing kohonen maps with geometry in: *International Conference on Artificial Neural Networks* pp. 609–612 Springer