Human behavior understanding from motion and bodily cues using deep neural networks

HUMAN BEHAVIOR UNDERSTANDING FROM MOTION AND BODILY CUES USING DEEP NEURAL NETWORKS

HUMAN BEHAVIOR UNDERSTANDING FROM MOTION AND BODILY CUES USING DEEP NEURAL NETWORKS

Dissertation

to obtain the degree of doctor at Maastricht University,
on the authority of the Rector Magnificus Prof. Dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public on
Tuesday, June 15, 2021, at 10:00 hours

by

Dario DOTTI

Faculty of Science and Engineering
Department of Data Science and Knowledge Engineering

This dissertation has been approved by the

Promotor:	Dr. S. Asteriadis
Promotor:	Prof. Dr. G. Weiss
Co-promotor:	Dr. M. Popa

Composition of the doctoral committee:

Prof. Dr. Ir. R.L.M. Peeters (chair), Maastricht University
Prof. Dr. M. Valstar, University of Nottingham
Prof. Dr. B. Jansma, Maastricht University
Dr. H.S. Hung, TU Delft
Dr. J. Niehues, Maastricht University

This PhD has been funded by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement *N*° 690090 (ICT4Life project).

SIKS Dissertation Series No. 2021-17

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover inspired by: A. Gormley, FEELING MATERIAL sculptures
Designed by: D.Dotti

ISBN 978-94-6423-300-1

Science never solves a problem without creating ten more.

George Bernard Shaw

CONTENTS

Summary **xi**

1 Introduction **1**
1.1 Context and Motivation . . . 1
1.1.1 Ambient Assisted Living Applications . . . 3
1.1.2 Smart Surveillance Applications . . . 6
1.1.3 Affective computing Applications . . . 7
1.2 Research Questions . . . 9
1.3 Thesis Overview . . . 11
References . . . 13

2 Introduction to state-of-the-art automatic human behavior analysis **17**
2.1 Introduction to the Theoretical Framework . . . 17
2.2 Artificial Neural Network . . . 18
2.2.1 Loss function . . . 20
2.2.2 Backpropagation . . . 21
2.3 Autoencoder Neural Network . . . 21
2.4 Long Short-Term Memory Network . . . 23
2.5 Convolutional Neural Network . . . 24
2.6 Literature review . . . 26
2.7 Ambient Assisted Living . . . 26
2.8 Vision-based smart surveillance . . . 28
2.8.1 Trajectory based analysis . . . 29
2.8.2 Surveillance datasets used in this dissertation . . . 32
2.9 Human motion analysis . . . 33
2.9.1 Skeleton-motion Features . . . 33
2.9.2 Social and nonsocial interaction . . . 36
2.10 Personality Computing . . . 37
2.10.1 Personality recognition . . . 38
2.10.2 Personality datasets used in this dissertation . . . 40
References . . . 42

3 Motion pattern discovery and path prediction **51**
3.1 Introduction . . . 51
3.2 Feature Extraction . . . 53
3.2.1 Occupancy Histogram (OH) . . . 53
3.2.2 Adapted Histogram of Oriented Tracklets (AHOT) . . . 53
3.2.3 Motion Descriptor SPEED and CAHOT . . . 54
3.2.4 Sparse Autoencoders (SAE) . . . 55
3.2.5 Unsupervised Learning . . . 56

3.3 Experiments . . . 56
3.3.1 Pedestrians vs. Auto-vehicles labeling in the LOST Dataset . . . 56
3.3.2 Labeling of Normal and Abnormal events . . . 58
3.4 Experimental results . . . 59
3.4.1 Abnormal Behavior prediction performance . . . 59
3.4.2 Qualitative Results . . . 59
3.5 Abnormal Behavior in Healthcare: The ICT4Life platform . . . 61
3.5.1 Motivation . . . 61
3.5.2 Introduction . . . 61
3.5.3 Abnormal behavior detection in healthcare . . . 63
3.5.4 Multimodal dataset for Abnormal Behavior Detection . . . 65
3.6 Conclusions . . . 71
References . . . 73

4 A Hierarchical Autoencoder Learning Model for Path Prediction and Abnormality Detection 75
4.1 Introduction . . . 75
4.2 Model Description . . . 78
4.2.1 Greedy Hierarchical Learning . . . 80
4.2.2 Bottom Layer Feature Extraction . . . 80
4.2.3 Bottom Layer Learning . . . 80
4.2.4 Second Layer Feature Extraction . . . 81
4.2.5 Second Layer Learning . . . 82
4.3 High Layer Inference . . . 83
4.3.1 Overview . . . 83
4.3.2 Motion Dictionary Construction . . . 83
4.3.3 Path Prediction using Bayesian Networks . . . 84
4.3.4 Long-term path modeling for Destination Prediction . . . 85
4.3.5 Abnormality Detection modeling . . . 86
4.4 Analysis of the Hierarchical Learning Model . . . 86
4.5 Experiments . . . 88
4.5.1 Experiment 1: Path Prediction . . . 88
4.5.2 Path Prediction: Qualitative Results . . . 89
4.5.3 Experiment 2: Destination Prediction on the GC dataset . . . 90
4.5.4 Destination Prediction: Qualitative Results . . . 92
4.5.5 Experiment 3: Abnormality Detection . . . 93
4.5.6 Experiment 4: Path prediction of different objects on the VIRAT dataset . . . 93
4.6 Conclusions . . . 96
References . . . 97

5 Behavior and Personality Analysis in a nonsocial context Dataset 99
5.1 From motion to behaviors using personality based model . . . 99
5.1.1 Personality recognition using trajectory patterns . . . 100

5.2 Introduction to Personality recognition using behavioral cues 103
5.3 Behavior and Personality in a nonsocial context Dataset 104
5.3.1 The Nonosocial Dataset: an extension of the Indoor motion dataset . 104
5.4 An AE-LSTM framework for personality recognition 106
5.4.1 Spatio-Temporal features (Heatmaps) 106
5.4.2 Spatio-Temporal clusters 107
5.4.3 Unsupervised Posture Representation 107
5.4.4 Posture dynamic modeling and personality recognition using LSTM 108
5.5 Experiment and Results 109
5.5.1 Personality Data analysis. 110
5.5.2 Spatio-Temporal Clustering of nonverbal behaviors 111
5.5.3 Personality Recognition 113
5.5.4 Personality visualization 115
5.6 Conclusions. 116
References 118

6 *Being the center of attention*: A Person-Context CNN framework for Personality Recognition 121
6.1 Person-Context Framework 124
6.1.1 Architecture 125
6.1.2 Person motion 125
6.1.3 Social Group Motion 126
6.1.4 Context proxemics 128
6.2 Person-Context Interaction Learning using a CNN architecture. 130
6.2.1 Feature Learning 131
6.3 Experiments 133
6.3.1 Experimental Setup 134
6.3.2 Personality traits configurations: Personality types 134
6.3.3 Ablation study 135
6.3.4 Salsa Dataset separate sessions 135
6.3.5 Personality types recognition 136
6.3.6 Personality traits recognition. 137
6.4 Qualitative Results 138
6.5 Discovered Personality Patterns. 140
6.6 Conclusions. 142
References 143

7 Temporal Triplet Mining for Personality Recognition 147
7.1 Introduction 147
7.2 The Proposed Framework. 150
7.2.1 Motion Features 151
7.2.2 Proxemics Distances 151

7.3 Temporal Identification Similarity Metric Learning (TISML) . . . 153
7.3.1 Definitions . . . 153
7.3.2 Formulation . . . 153
7.3.3 Temporal Triplet Mining (TTM) . . . 154
7.4 Implementation Details . . . 155
7.5 Experiments . . . 156
7.5.1 Datasets and Labels . . . 156
7.5.2 Evaluation protocol . . . 156
7.5.3 TTM versus Random Triplets Mining (RTM) . . . 157
7.5.4 Impact of Time-window Selection . . . 158
7.5.5 Ablation study . . . 160
7.5.6 Comparison with baseline techniques . . . 160
7.5.7 Personality Traits Recognition . . . 162
7.5.8 KNN classifier investigation . . . 163
7.6 Conclusions . . . 163
References . . . 165

8 Conclusions and Future Research **169**
8.1 Answers to the research questions . . . 169
8.1.1 Research question 1: How can motion trajectories be leveraged for the discovery of normal as well as abnormal behavioral patterns? . . . 170
8.1.2 Research question 2: How can motion trajectories be leveraged for real-time surveillance applications? . . . 171
8.1.3 Research question 3: Posture sequence modelling and affective computing: what can we automatically learn about personality using body postures? . . . 171
8.1.4 Research question 4: Are contextual cues informative predictors in addition to posture for personality recognition? . . . 172
8.1.5 Research question 5: Does modelling the temporal nature of human behaviors improve latent representations and consequently personality recognition? . . . 174
8.2 Future Research . . . 175
References . . . 177

Impact paragraph **179**
References . . . 181

Acknowledgements **183**

Curriculum Vitæ **185**

List of Publications **187**
References . . . 187

SIKS Dissertation Series **189**

SUMMARY

Automatic human behavior understanding is considered a core technology that can facilitate a variety of applications. Nevertheless, defining, detecting, and recognizing human behavior is still a big challenge requiring research efforts from the computer vision and machine learning communities.

Technological advancements in the field of Artificial Intelligence (AI) have opened the path to systems capable of learning and sensing the environment in a way that imitates human perception. Machines are very powerful when it comes to learning regular and tangible patterns. However, there is still big room for improvement in the fields concerning the automatic understanding of behaviors and how humans use them to communicate as well as to express their feelings.

In this dissertation, we present novel work in the field of computer vision and behavior understanding using image and video data. In particular, we will mainly focus on the rich information that the human body generates during daily activities. Nonverbal signals refer to the types of humans' daily communication that are nonverbal. From our gestures to our body postures and movements, nonverbal communication conveys large volumes of information that humans read and interpret every day. Therefore, in this dissertation, we pose the critical research question of how to build computational models that can enhance machines' understanding of human intentions, behaviors, personality traits, and activities, by learning meaningful patterns from human motion and bodily cues.

As human behavior understanding is a research topic that can be potentially used to support several fields of our society, in this dissertation, we focus specifically on three research fields: Ambient Assisted Living (AAL), Video Surveillance (VS), and Affective Computing (AC). In Chapter 1, we introduce the main challenges and outcomes from each of them. AAL concerns the use of ambient intelligence techniques, processes, and technologies to enable aging individuals to live independently for as long as possible. Smart AAL applications made with low-cost sensors monitoring and detecting dangerous events in elderly homes can reduce the healthcare economic burden while improving the living conditions of the senior citizens. VS concerns surveillance systems based on a set of cameras that monitor public or private areas. Smart VS applications that automatically understand the filmed events can increase the efficiency of the surveillance staff while reducing the systems' cost. Finally, AC concerns the understanding of human affective cues such as emotions and personality attributes. AC applications that automatically recognize and interpret personality attributes can have great impact on understanding why individuals make certain choices in fields like marketing or human resources.

Human behavior analysis from video data is one of the most complex challenges in the computer vision community as movements are difficult to define and lack clear semantic structures. Moreover, the categorization of movements is a non-trivial problem

for several reasons. Movements associated with the same activity can vary in duration or expressivity. For instance, walking behaviors can depend on the individuals making the actions, e.g., elderly usually walk with a slower pace than young individuals, or depending on the context, e.g., in crowded spaces we may walk in a more zigzag pattern compared to when we walk in free spaces. In Chapter 2, we introduce the theoretical frameworks used in this dissertation as well as we present the state-of-the-art methods in the fields of Ambient Assisted Living (AAL), Video Surveillance (VS), and Affective Computing (AC).

In this thesis, movements are extracted as chronological sequences of multi-dimensional locations called trajectories. Trajectory information provides meaningful insights about motion towards a destination and its related motion patterns. As trajectories are simply multi-dimensional location information, they are very easy to store and privacy compliant, hence, they are commonly used for surveillance applications. Specifically, in Chapter 3, we use trajectory based methods for the detection of abnormal events in outdoor public spaces as well as private homes. Additionally, in Chapter 4, we continue to investigate abnormal behavior detection applications proposing a real-time framework based on trajectory data.

Although trajectory data is important for general surveillance applications, it does not provide rich information about the articulated motion of the human body. Therefore, in order to obtain more fine-grained insights about human body motion and behaviors, in Chapter 5, we introduce a framework that encodes skeleton joints information to learn spatio-temporal sequences related to human body postures.

In computer vision, contextual information has been shown to improve several challenging tasks such as action recognition and scene understanding. Building on these findings, in Chapter 6, we aim to extend our research by understanding the mutual relation between behaviors that come intrinsically from individuals (e.g. motion) and information that comes from the context (e.g. social/nonsocial situations). Additionally, in Chapter 7, we deepen our investigation on the temporal evolution of human behaviors and their similarity using Deep Metric Learning techniques. Finally, in Chapter 8, we address the research questions that guided our research throughout this PhD and draw conclusions and recommendations for future works.

In this dissertation, five research questions are addressed. The first question consists of *How can motion trajectories be leveraged for the discovery of normal as well as abnormal behavioral patterns?* In Chapter 3, we explore various temporal features in combination with spatial information to encode objects' motions using trajectory data. Using an unsupervised approach, we investigate the detection of normal as well as abnormal motion events in indoor as well as outdoor scenarios.

The second question consists of *How can motion trajectories be leveraged for real-time surveillance applications?* In Chapter 4, we tackle this question by proposing a hierarchical framework, based on Autoencoders, for modeling motion trajectories in real-time. The hierarchical architecture is designed to capture short, noisy spatio-temporal trajectories in the lower levels while learning meaningful motion transitions in the higher levels. Finally, we model temporal motion transitions to infer the future trajectory step in real-time.

The third question consists of *Posture sequence modelling and affective computing: what can we automatically learn about personality using body postures?* In Chapter 5, we introduce a novel approach to learn upper body posture representations and their evolution in time using an Autoencoder in combination with a Long Short-Term Memory Network. In this chapter, we study body posture sequences and their link to personality attributes. To do so, we propose a novel dataset where forty-six participants were recorded performing six tasks in unconstrained indoor environments and their personality scores were reported using a self-assessment questionnaire on the big-5 personality traits.

The fourth question focuses on *Are contextual cues informative predictors in addition to posture for personality recognition?* In Chapter 6, our goal is to map the mutual relation between individual behaviors and contextual information to personality attributes. To do so, we introduce a novel Convolutional Neural Network framework that analyzes the behavioral events in the scene at multiple levels of granularity. Firstly, we encode individual movements from every person in the scene. Secondly, we explore the interaction between individuals in small social groups, by studying how communication dynamics are affected by the personality of single individuals involved in the group. Thirdly, we explore how individuals use their personal space in different situations such as in social as well as nonsocial scenarios.

Finally, the fifth question consists of *Does modelling the temporal nature of human behaviors improve their latent representations and consequently personality recognition?* In Chapter 7, continuing the research line of the previous chapter, we aim at expanding the use of body motion as well as context information to learn their interaction dynamics in time. We propose a novel model that encodes temporally adjacent motion and context descriptors as they are likely to belong to the same semantic behavior. The learning process is carried out using a Deep Metric Learning strategy with the goal of finding meaningful movements that enhance the discovery of discriminative personality patterns.

Overall, as nonverbal communication (e.g. body movements, body postures, and expressions) convey rich information about behaviors, in this thesis, we proposed several novel methods to extract, learn, and visualize meaningful patterns of human behaviors. Our findings show that by examining the interaction between movements and contextual cues, we can enhance machines' understanding of how humans behave in different environments.

1

INTRODUCTION

1.1. CONTEXT AND MOTIVATION

We are in the midst of a wave of technological innovations that are revolutionizing several sectors of our society. Mass digitalization, sensors that are increasingly more ubiquitous (Internet Of Things (IoT)), and Machine Learning (ML) techniques leveraging the big amount of collected data are the main drivers of this new revolution. This technological growth aims to tackle some of the grand challenges of our society, such as renewable energy, healthcare, surveillance, and automation. In this context, the integration of these technologies in our society remains a fundamental challenge to be tackled. Since the days of early computers, the field of Human Computer Interaction (HCI) has undergone great advancements to bridge the gap between man and machine. Great inventions were produced due to these efforts, including the keyboard, the mouse and the touch screen. In recent years, the technological development in the field of HCI was greatly affected by Artificial Intelligence (AI), opening the path for more natural and intuitive interactions between humans and machines. Researchers have began to integrate human communication modalities such as vision, body gestures and natural language understanding into computers and robots. For example, one of the most successful commercialized product utilizing natural speech interaction is Amazon Alexa [1], where the users can control the device using voice commands. In order to convert speech to text, Alexa is embedded with Natural Language Processing (NLP) techniques. Another example is Google Soli [2], where the users can control the smartphone applications using natural gestures. Google smart sensors are designed to recognize human motion at several scales, making the navigation more genuine.

In this thesis, we will mainly focus on the rich information that the human body generates during daily activities. Despite the great influence of body language in human communication, the topic has become popular only since the 1960s [3], and yet, the majority of people consider speech as the main form of communication among humans [4]. Several scientists tend to see language communication as a very recent evolutionary event [5], whereas, body expressions (e.g. manual gestures) have been shown to be

much more ancient [6]. Nonverbal signals refer to the portion of our communication that is nonverbal which includes body language, facial expressions, paralinguistics etc. These nonverbal signals shape a major part of humans' daily communication. From our facial expressions to our body movements, the nonverbal communication conveys large volumes of information that humans read and interpret everyday. *"Reading body language is more than just a matter of perception. It entails not only recognizing and coding socially relevant visual information, but also ascribing meaning to those representations"* [7]. While we view ascribing meaning to non-verbal communication as the final step towards intelligent social machines, a critical research question addressed in this dissertation is how to build computational models that can enhance machines' understanding of human intentions, behavior, personality traits and activities, by learning meaningful patterns from human motion and bodily cues.

Given the recent technological advances, fully autonomous systems will gain significant relevance in people's life. It is expected that the level of manual input during human-machine interactions will decrease, letting the technology work autonomously. Under these circumstances, researchers coined the term "technologies as perfect butler" [8], where the system aims to fulfill users' preferences without having them stated manually. In this direction, we believe that automatically recognizing users' behaviors through measurements and observations (passive interaction) can yield great improvement into several fields of our society. For example, in healthcare, smart algorithms can support the hospitals and elderly homes in the monitoring and detection of dangerous/abnormal events [9]. Private homes equipped with smart sensors can learn the normal life cycle of the inhabitants and perform actions automatically like switching on/off the heater, switching on/off the lights, and preparing meals [10]. In public environments, such as train stations and public squares, surveillance cameras are already passively monitoring the daily activities. However, the process of understanding and detecting the salient events in the recorded data is carried out manually. Also in the surveillance field, it is expected that the level of manual work will decrease because of smart systems that can understand and detect important events in an automatic way.

Taking the above into account, the final goal of this work is to investigate AI-endorsed systems able to automatically understand human behaviors in social and nonsocial environments using sensory data. Unlike natural language processing, which is based on languages that have well studied semantic and syntactic structures, human motion analysis is lacking a generic underlying architecture [11]. As a matter of fact, human motion is generated in different forms and levels of complexity. Motion can be generated by the full body that moves coherently to make an activity such as running or jumping, it can be generated by only parts of the body like facial expressions or gestures, or, motion can be seen solely as movement through space towards a destination. Therefore, it is extremely complex for systems to perceive, understand and anticipate human motion in different environments. The challenge of accurately modeling human motion derives from the complexity of human expressivity and the variety of its internal and external stimuli [12]. Individuals' motion may be guided by own intents and internal needs (walking fast due to a delay) or may be guided by external factors such as social rules and norms (walking on a pedestrian sidewalk instead of on the road). Some factors influencing human motion are observable, such as movement quality and movement types, yet several

other factors are not directly observable and need to be deduced from perceptual cues or from the contextual information [12]. Hence, in this thesis, we aim to study observable cues as well as external context cues that can reveal internal drivers such as intentions, attitudes, and goals.

In the computer vision field, terms like actions, activities, and behaviors are often used interchangeably by different authors [13]. In this thesis, we define the following hierarchy: "action", "activity", "behavior". The term "action" describes primitive and atomic entities that are made in sequence to generate more complex activities. For example, to kick a ball, action primitives can be "foot planting", "cocking of kicking limb", "swing" etc. We define the term "activities" as more coarse-grained behaviors that have semantic definitions, for example "playing football" or "making tea". Activities have a clear beginning and a clear end in time, and they usually follow standards defined by the society. Finally, we define the term "behaviors" as coarse-grained entities which neither follow any periodicity nor standard and therefore they are difficult to segment in time, and they are challenging to label. As stated by the authors in [14], the term activity defines the concepts of "what" the user is doing in the environment, while, the term behavior defines "how" the activity or action is performed.

Action and activity recognition have received much attention in the computer vision field [15], this advancement was also fueled by the introduction of several labeled datasets which provided human annotations regarding the most meaningful actions or activities in the data. Nevertheless, we argue that in real life scenarios, human movements are not always well-defined, moreover, researchers cannot always rely on data annotations that determine what to focus on and why. Hence, in this thesis, we shift the attention on the definition and interpretation of *coarse behaviors*. We believe that human behavior understanding can enhance the computer vision world in several ways. First of all, as behaviors are difficult to annotate and segment, techniques like unsupervised learning must be adopted. Consequently, researchers are not constrained by the constant need of human annotations and definitions. Secondly, behavior understanding is not a purely mathematical/computational problem, it should be tackled from different scientific points of view, for example by adopting models from other fields such as Psychology [16] and Art [17]. Finally, behavior understanding using computer vision techniques has great potential due to the wide range of promising applications in critical fields like healthcare and surveillance. Specifically, in this thesis, we focus on the impact of automatic human behavior understanding in the research fields of Ambient Assisted Living (AAL), Video Surveillance (VS), and Affective Computing (AC). In the next sections, an in-depth description of the three main research fields is given.

1.1.1. AMBIENT ASSISTED LIVING APPLICATIONS

Ambient Assisted Living (AAL) concerns the use of ambient intelligence techniques, processes, and technologies to enable aging individuals to live independently for as long as possible [20]. As the average population age in Europe, and in the world in general, is steadily increasing, new challenges surrounding the quality of life of older people and their carers are emerging [21]. The need for caregiving, home assistance, rehabilitation and physical support rises the expenses of the healthcare domain in every government in the world. For example, costs of caring for people with Alzheimer's and other dementia

(a) A wearable device measuring vital signs. Image from [18]

(b) A private home equipped with sensors that make smart inferences, in this case, a falling down event is detected. Image from [19]

Figure 1.1: Two examples of AAL smart applications.

in USA were estimated around $203 billion in 2013, and the projections indicate that the costs will be around $1.2 trillion per year by 2050 [22]. The situation in Europe is similar, as the demographic statistics estimate that the population over 60 years old is 24.5% of the total inhabitants [22] yielding significant financial burden upon the socioeconomic systems of all the European countries. In this context, smart healthcare applications made with low-cost sensors could reduce the economic burden while improving the living conditions of the old population [22]. For example, one of the first improvement brought by AAL was the possibility for seniors and impaired persons to remotely interact with relatives, friends, and doctors using video conferencing technologies [23].

Besides already existing technologies, in the last decades, numerous initiatives have been carried out to propose ICT systems based on off-the-shelf active and/or passive sensors that can remotely monitor different aspects of daily life. For example, the research and innovation framework of ICT for Aging Well and the Europe AAL Joint Program have been launched by the European Union for cultivating the development of innovative ICT-based products, services, and systems to support the process of aging well at home, in the community and at work [22]. These initiatives aim to bring together universities and companies to design breakthrough applications to support the elderly in different AAL scenarios.

The scenarios that AAL is addressing are complex. A key source of this complexity is the inherent heterogeneity of the end-user population including individuals suffering from diverse disabilities and illnesses, but also healthy individuals, who are mainly interested in "lifestyle functionalities" to improve their quality of life [24]. Another significant aspect to consider for AAL technologies is the different house arrangements of the target population. As AAL systems will frequently need to be deployed in preexisting houses, general and modular system infrastructures should be adopted to adjust as much as possible to the different houses' layouts. Finally, AAL products and services are not limited to direct users, but address also health providers, caregivers, and family members, providing tools to record, share, and visualize health information about the

users.

Although the AAL products and services are essentially designed to be stand-alone systems, great effort is undergoing to integrate these systems to provide a more effective and efficient support. The final goal is to provide a more complete overview of the patients' clinical pictures by integrating different care systems (i.e. *integrated care*). Integrated care aims to enhance quality of care and quality of life, consumer satisfaction and system efficiency by cutting across multiple services, providers and settings [25]. For example, the patients' clinical picture is composed by medical exams made in the hospital. Nevertheless, the condition of the patients in their everyday life at home is often only shared verbally during the intake procedure. In this context, smart monitoring systems that provide analysis of everyday activities could be used to enhance the patient's clinical picture. In this way, doctors could take into account the insights provided by the monitoring systems to make a diagnosis, administer a therapy, and follow-up. Finally, the diagnosis as well as the therapy could be fed back into the ICT systems to update and regulate the monitoring metrics.

Among the AAL applications, daily life activity (DLA) monitoring is an essential component for assisted living. Continuous monitoring can play a crucial role in detecting abnormalities or emergencies during the intervals between the frequent or infrequent visits to the doctor. This factor, in combination with more affordable and accessible hardware represents an opportunity to create solutions that can change people's lives now and in the future. Consequently, monitoring technologies are receiving more attention from policy makers, academia, and the industry. These technologies include smart homes, mobile, and wearable sensors among others [26].

Smart homes are regular houses that have been equipped with ambient sensors such as cameras, microphones, and infrared sensors. Sensors acquire daily life information and this knowledge is utilized for automation and provision of comfort as well as for assessing the cognitive and physical health conditions of the users. A question that arises is where in the home should AAL technologies be deployed so as to be harnessed to their maximum benefit [27]. Although accidents can occur anywhere within the home, authors in [27] indicated the kitchen, the bathroom, and the bedroom as those places where accidents are most likely to happen. However, the bathroom and the bedroom locations are more problematic due to the privacy concern of the target population. Limitations of these types of technology are high installation cost, maintenance, in-home coverage, as well as digital know-how required in all the operations.

Mobile and wearable sensors take advantage of the close position to the human body to collect physiological measurements. Most of the smartphones are equipped with various sensors such as accelerometers, gyroscopes, proximity sensors, and global positioning system (GPS), which can be used for detecting user activity and mobility. Wearable sensors such as bands and clothes can measure blood glucose, blood pressure, and cardiac activity. Limitations for these types of sensors include their low accuracy, which is often the case in low-cost systems. Moreover, in the case of senior citizens, wearing bulky or unfamiliar devices may prove detrimental or necessitate a very close supervision.

Studies in automatic monitoring systems in healthcare have highlighted their potential to improve the feeling of security in the aging population living independently [28]. Examples of AAL applications and results are given in Fig. 1.1. Cardiovascular diseases

(a) New York train station movements were recorded in the dataset published by [36]

(b) Data from surveillance cameras include several objects like pedestrians and cars. Image from [37]

Figure 1.2: Two examples of Smart Surveillance applications

are a great burden for the European population, hence, several devices and applications using wearable devices that can measure heart activity have been launched (Fig. 1.1a). Around 36,000 elderly are reported to be fatally injured from falls every year in Europe [29], for this reason, smart devices have been presented for the detection of falling down events (Fig. 1.1b). In the next chapter, a detailed explanation of the algorithms shaping the state-of-the-art for these applications is given.

1.1.2. SMART SURVEILLANCE APPLICATIONS

Visual surveillance systems consist of a set of cameras that monitor in real-time the surrounding environment. The recorded data is controlled by human operators with the goal of detecting possible abnormal events. Nowadays, despite an increasing concern about privacy, a growing number of countries are deploying advanced surveillance systems to monitor, track, and surveil public areas [30]. The process of watching, reviewing, and analysing the data is performed mostly manually [31], and, monitoring long and unconstrained security footage is challenging due to limited human cognitive resources. The ability to hold attention and to react in case of rarely suspicious events is demanding and prone to errors [32]. For example, authors in [33] showed that after 20 minutes of watching surveillance cameras, attention of operators drops due to boredom and the natural hypnotizing effect of monitoring video scenes. Thus, automatically understanding the recorded events and detecting possible abnormal dynamics are fundamental tasks that support as well as enhance the traditional surveillance systems.

Due to the increasing technological power and data availability, researchers have started, not only to be able to understand the events occurring in the scene but more importantly, to predict what can happen next. In this regard, authors in [32] reported a shift in the security paradigm from “investigation of accidents” to “prevention of potential accidents”. This degree of anticipation helps in inferring targets’ routes [34], intents [35], and destinations [36]. For instance, in a road intersection, it is crucial to immediately detect whether a car lost control and is approaching pedestrian areas, or to forecast a pedestrian’s walking path into a potentially restricted zone.

Objects’ navigation (i.e. humans, cars) is affected by multiple factors. External factors, such as context regulations and static obstacles in the environment follow standard

conventions and therefore can be known a priori. On the other hand, internal factors such as goal destinations and attitudes are more challenging to model as they appear unpredictable. Therefore, a major challenge for a smart surveillance system is to embed external and internal factors in order to understand events in the scene, and more importantly, to predict future occurrences.

Navigation in an environment is conveyed by the object position in space. In this thesis, we consider the ground level 2D or 3D trajectory to indicate the object navigation. Object trajectories can be analyzed in a real-time mode as well as in an offline mode.

Real-time monitoring systems track and analyse continuous snapshots of data occurring within short periods of time, producing an output almost simultaneously with the input. These systems have the potential power of detecting abnormal events as soon as they arise. However, real-time systems are also more prone to errors due to the limited reaction time. Anomaly detection systems have success in exposing attacks in real-time, yet have high false positive rates [38]. False positive events occur when an activity is flagged as dangerous but it was determined to be benign upon analysis. False alarms have costs, for example, computational power and valuable resources are spent when irrelevant events are flagged as dangerous.

Offline monitoring systems analyse the data after it has been recorded. Offline processing enables the use of more complex and computationally demanding strategies, therefore provides more accurate results than real-time processing. However, one disadvantage of offline monitoring systems is that they are unable to prevent dangerous events due to their asynchronous nature.

As humans, we possess navigational skills acquired through years of practice [34]. Hence, AI models should be able to imitate navigation strategies that human employ in different scenarios. For example, moving in a crowded public space, such as a train station, requires the ability to avoid other moving objects without losing the destination goal. On the other hand, moving in a less crowded indoor space, such as a house, requires more fragmented and curved movements due to limited space.

Spatio-temporal human trajectories offer valuable information about human motion, therefore, a plethora of applications was proposed such as pedestrian motion analysis in public spaces (Fig. 1.2a) and motion patterns discovery for various vehicles (Fig. 1.2b). In the next chapter, a detailed explanation of the algorithms shaping the state-of-the-art for these applications is given.

1.1.3. AFFECTIVE COMPUTING APPLICATIONS

Affects and emotions are fundamental human experiences that have a great impact on humans' lives, choices, well-being, and so forth. Since antiquity, studies on humors and emotions highlighted that individuals differ in their predisposition to experience certain emotions and feelings [39]. The relatively stable psychological attributes, affective patterns, and desires that distinguish individuals from one another are defined as Personality patterns.

"A helpful analogy is to consider that personality is to emotion as climate is to weather. That is, what one expects is personality, what one observes at any particular moment is emotion" [39]. In general, personality includes patterns of thoughts that shape individual's behaviors, emotions, motivations, and characteristics. Personality psychology is

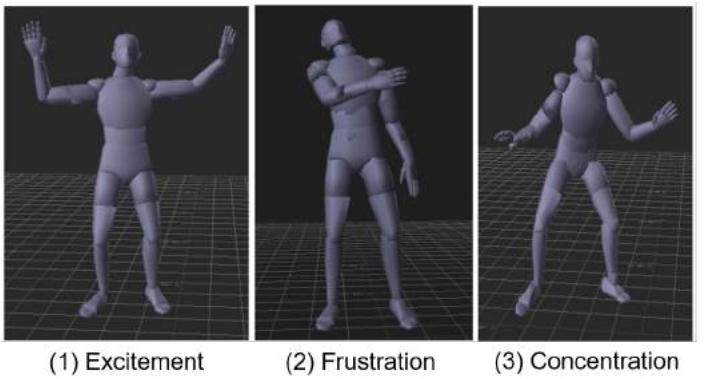


Figure 1.3: Examples of body expressions that reveal affective states of individuals. Pictures taken from [44]

the science that aims at capturing stable individual characteristics that explain and predict observable behavioral differences [40]. Hence, the ability to automatically recognize and interpret personality patterns has a huge impact on several technological applications. Consequently, the field of affective computing has received increasing attention in the last years. In this thesis, we will study how human body postures, expressions, and behaviors can be automatically mapped to personality labels.

Understanding behavioral patterns helps to have a better idea of how people act, think and behave [41], therefore several psychological models have been proposed throughout history. The Big-Five trait model [42] is the most popular personality model, dividing human behavior characteristics in five general dimensions: Extroversion (i.e. talkative, outgoing), Agreeableness (i.e. considerate, forgiving), Conscientiousness (i.e. efficient, motivated), Neuroticism (i.e. tense, full of worries), and Openness to Experience (i.e. inventive, imaginative). However, a long term debate in the personality field showed that considering single traits that vary across individuals fails in describing the configuration of these attributes within each person. In contrast, numerous studies confirmed the theory proposed in [43], in which all the personality traits can be organized into three major types: undercontrolled, resilient, and overcontrolled. Resilient personality types exhibit low levels of neuroticism, and intermediate or above average in the rest of the Big-Five traits. The undercontrolled type usually scores high in neuroticism and extraversion, and finally, the overcontrolled type scores below average on extraversion and above average on neuroticism.

Psychological experiments highlighted the validity of theoretical personality models by predicting accurately participants' life outcomes. For example, authors in [45] detected an association between extrovert, as well as proactive personality to career success indicators. For this reason, the computing community started to adopt and consult personality models for automatic behavior understanding in the field of personality computing [40].

Human behaviors can be verbal, where words expressed by the voice represent the communication channel, and nonverbal, where the body language is the main communication channel. Verbal communication in everyday life is without any doubt impor-

tant, and, since the words are measurable, researchers showed a clear link between the way a certain message is articulated and personality styles [46]. For example, an expressive style is often an outcome of an expressive personality [46], while an aggressive communication style is an outcome of an aggressive personality [47]. However, results showed that verbal communication is easier to train and manipulate to convey biased messages [48]. Words are not the only tool humans use to communicate. The way people stand, the gestures, and the facial expressions are nonverbal ways of communicating feelings that impact individuals' behavioral patterns. Psychological studies showed that the nonverbal channels are much more unconscious than the verbal channel and, therefore, nonverbal signals can be more valuable in revealing the true internal state and the true personality. Hence, in this thesis, we focus on nonverbal behavioral cues such as body posture and expressive movements and their relation with personality attributes.

As human beings, we are able to interpret affective states of other individuals from little information. For instance, in Fig. 1.3, even though only body posture information is shown, we are able to understand the expressed affective states. For these reasons, several studies were carried out to understand how emotions, attitude, and personality are conveyed in dynamic body gestures and postures. In the next chapter, a detailed explanation of the algorithms shaping the state-of-the-art for these applications is given.

1.2. RESEARCH QUESTIONS

Given the fruitful applications that can be derived from human behavior understanding, this thesis focuses on conducting research on several components of human behavior (e.g. motion trajectory, body posture dynamics, personality). Specifically, the following questions have guided our research activities since the beginning of this PhD study:

Research question 1: *How can motion trajectories be leveraged for the discovery of normal as well as abnormal behavioral patterns?*

One of the goals in human behaviour understanding consists in learning an object's regular activity patterns and defining types of deviations which could be considered abnormal. This analysis is useful for modeling normal behaviours in a range of different contexts, such as private houses, work environments, and public spaces.

Abnormality detection yields several challenges: first of all, by definition, abnormal events occur much less often than normal events. Therefore, systems that deal with this problem have to face unbalanced datasets containing a few examples of significant anomalies. One of the most used strategies consists of learning regular activity patterns and categorizing as abnormal everything that significantly deviates from it. However, it is infeasible to collect all possible normal events, hence, systems may suffer from false alarms when recording unseen normal events. Another strategy is to adopt supervised approaches, where normal and abnormal events are defined in the model. However, the problem mentioned before of unbalanced data makes the labeling operation labor-intensive.

Our research aims to tackle both challenges by proposing a semi-supervised approach in which activity patterns are discovered without any human supervision using clustering techniques. Then, we facilitate the integration of experts' opinions for obtaining a semantic interpretation of the scene, making them define what is normal or

abnormal using the obtained clusters. This strategy gives the advantage of labeling sets of similar events instead of labeling the events one by one. Lastly, we use the obtained labels for a supervised abnormality detection task in an indoor as well as an outdoor scenario.

Research question 2: *How can motion trajectories be leveraged for real-time surveillance applications?*

Real-time processing is a key factor for smart surveillance systems. However, real-time behavior understanding is extremely challenging due to the fact that human movements are inherently noisy. Humans, differently from programmed machines, do not always take the shortest or most efficient path towards a goal for many different reasons. Therefore, there is a need for real-time systems that are robust to noise and small deviations to estimate accurately the long-term path progression.

Our research aims to model trajectories from surveillance scenarios using a hierarchical strategy. Low levels of the hierarchy capture local spatio-temporal motion attributes such as spatial orientation and speed, while higher levels contribute to obtaining richer semantic information. The bottom-up approach exploits the inherent statistical correlations between neighboring elements using an increasing spatio-temporal grid. With the proposed strategy, local patterns describing small movements are discovered by the low hierarchical levels, while more discriminative and robust motion patterns are encoded using higher layers. Finally, our research aims to model the dynamics of the learned features for surveillance applications like trajectory path forecasting, direction prediction, and abnormality detection.

Research question 3: *Posture sequence modelling and affective computing: what can we automatically learn about personality using body postures?*

Human behavior consists of multimodal signals such as articulated full body motion, facial expressions, voice and movements in space [12]. A common approach is to consider one or multiple modalities for behavior understanding in predefined scenarios. Every modality has strengths and weaknesses. For example, analysing voice pitch in every day life [49] provides useful information regarding the emotional status of individuals, however, feature extraction may be difficult due to surrounding noise. Nonverbal behavioral cues have been studied extensively in the field of human behavior understanding. Facial expressions have been successfully linked to true human behavior characteristics, however, these features are not robust to camera positions and privacy restrictions. These limitations are critical for applications in smart surveillance, hence, in this thesis, we study body posture sequences and their link to personality attributes. Besides the theoretical motivation explained before, in our opinion, modeling body and posture expressions has several practical advantages: 1) In order to cover wider areas, surveillance cameras are often placed on the ceiling, human postures and upper-body movements are usually clearly visible, making them robust to noise and occlusions. 2) As upper-body parts cover more than half of the human body volume, it is easier to track them in respect to small-volume objects like faces. 3) Body postures can be represented by skeleton joints sketches, and therefore, fully respecting the individuals' privacy.

To deepen the research in the field, we introduced a novel dataset for behavior understanding and personality recognition. Forty-six participants were recorded in an unconstrained indoor space, related to a smart home environment, performing six tasks

resembling Activities of Daily Living (ADL). During the experiment, personality scores were collected using self-assessment questionnaires. Furthermore, our research aims to map body posture sequences to participants' personality attributes with the goal of integrating theories from psychology with computational models.

Research question 4: *Are contextual cues informative predictors in addition to posture for personality recognition?*

In the field of smart monitoring systems, the affective/emotional aspects of human behaviors are often omitted [50], whereas, it has been shown that actions and movements made by the individuals are influenced by their affective state, as well as their personality attributes [51]. For example, authors in [51] show that the pedestrians' walking behaviors in a public environment can be described by affective attributes. Pedestrians are considered aggressive if their walking path is not influenced by the surrounding crowd, whereas pedestrians are considered conservative if their paths are modified to avoid contact with other people.

Building on these findings, our research aims to map the mutual relation between individual behaviors and contextual information to personality attributes with the goal of shedding some light on the influence of surroundings on human behaviors.

Research question 5: *Does modelling the temporal nature of human behaviors improve their latent representations and consequently personality recognition?*

Mapping behavioral patterns to personality labels can allow an intelligent system to be more interactive and adaptive [8], reducing unnecessary manual interaction. For instance, autonomous cars can be customized according to the personality of the driver [52] to behave according to the personal driving inclinations and perceptions with the goal of ensuring a smoother ride. Comparing and categorizing behavioral patterns for personality understanding is a challenging task due to the different lengths that the same behavior can acquire when performed in different scenarios or performed by different individuals. For example, a simple action like walking, can be expressed in several different ways depending on the attitude/personality (i.e. nervous versus calm personality), or depending on the scenario (i.e. crowded versus empty scenario).

Therefore, one of the goals of our research is to expand the use of body information, context learning and their dynamic interaction in time. In particular, we aim to measure and optimize the similarity between temporally related spatio-temporal behavioral samples in order to understand their semantic relation in different scenarios.

1.3. THESIS OVERVIEW

This thesis is divided into eight chapters. Chapter 1 provided a brief introduction to the automatic human behavior understanding research field with applications in various domains such as Ambient Assisted Living, Smart Surveillance, and Affective Computing. Chapter 2 reviews the state-of-the-art algorithms and frameworks regarding the studied domains involving human behavior understanding. In Chapter 3, the first research question is addressed with the use of a semi-supervised framework for the discovery of normal and abnormal behavioral patterns in indoor as well as outdoor scenarios. In Chapter 4, the second research question is studied with the proposition of a hierarchical model for trajectory analysis and real-time prediction of future trajectory paths. In Chapter 5, the third research question is tackled with the proposition of a temporal framework for

the mapping of posture dynamics to personality attributes. In Chapter 6, the fourth research question is addressed with the proposition of a framework that aims to map motion cues generated by the human body and contextual information in the same latent space. In Chapter 7, the fifth research question is investigated utilizing a metric learning strategy that aims to minimize the intra personality class variation and maximize the inter personality class variation in time. Finally, Chapter 8 discusses the outcomes of this dissertation, provides a series of conclusions and presents possible directions for future work in the new emerging fields of Artificial Intelligence, Affective Computing, and the role of Deep Learning architectures.

REFERENCES

[1] V. Perera, T. Chung, T. Kollar, and E. Strubell, "Multi-task learning for parsing the alexa meaning representation language," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[2] Google, "Google soli project." https://atap.google.com/soli/.

[3] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, pp. 1–1, 2018.

[4] B. Pease and A. Pease, *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam, 2008.

[5] B. MacWhinney, "Language evolution and human development," *Origins of the social mind: Evolutionary psychology and child development*, pp. 383–410, 2005.

[6] K. Liebal and J. Call, "The origins of non-human primates' manual gestures," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1585, pp. 118–128, 2012.

[7] C. M. Tipper, G. Signorini, and S. T. Grafton, "Body language in the brain: constructing meaning from expressive movement," *Frontiers in Human Neuroscience*, vol. 9, p. 450, 2015.

[8] H. Achten, "Buildings with an attitude," in *Stouffs, R. andSariyildiz, S.(eds.), ComputationandPerformance–Proceedings of the 31st eCAADe Conference*, vol. 1, pp. 477–485, 2013.

[9] D. Dotti, M. Popa, and S. Asteriadis, "Unsupervised discovery of normal and abnormal activity patterns in indoor and outdoor environments.," in *VISIGRAPP (5: VISAPP)*, pp. 210–217, 2017.

[10] S. Majumder, E. Aghayi, M. Noferesti, H. Memarzadeh-Tehran, T. Mondal, Z. Pang, and M. J. Deen, "Smart homes for elderly healthcare—recent advances and research challenges," *Sensors*, vol. 17, no. 11, p. 2496, 2017.

[11] N. Dael, M. Mortillaro, and K. R. Scherer, "The body action and posture coding system (bap): Development and reliability," *Journal of Nonverbal Behavior*, vol. 36, no. 2, pp. 97–121, 2012.

[12] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[13] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[14] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A survey on ontologies for human behavior recognition," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, pp. 1–33, 2014.

[15] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[16] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[17] A. Hutchinson, "Labanotation," *Journal of Aesthetics and Art Criticism*, vol. 13, no. 2, pp. 276–277, 1954.

[18] H. . E. project, "Ict4life." http://www.ict4life.eu/.

[19] B. J. Kröse, T. van Oosterhout, G. Englebienne, *et al.*, "Video surveillance for behaviour monitoring in home health care," Citeseer, 2014.

[20] F. Cardinaux, D. Bhowmik, C. Abhayaratne, and M. S. Hawley, "Video based technology for ambient assisted living: A review of the literature," *Journal of Ambient Intelligence and Smart Environments*, vol. 3, no. 3, pp. 253–269, 2011.

[21] A. Abtoy, A. Touhafi, A. Tahiri, *et al.*, "Ambient assisted living system's models and architectures: A survey of the state of the art," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 1, pp. 1–10, 2020.

[22] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: a survey," *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015.

[23] M. Amoretti, G. Copelli, M. Muro, M. Picone, and F. Zanichelli, "e-inclusive videoconference services in ambient assisted living environments," in *AmI2009, 3rd European Conference on Ambient Intelligence*, pp. 227–230.

[24] P. Antón, A. Muñoz, A. Maña, and H. Koshutanski, "Security-enhanced ambient assisted living supporting school activities during hospitalisation," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, no. 3, pp. 177–192, 2012.

[25] WHO, "Integrated care models: an overview," 2016.

[26] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 579–590, 2012.

[27] M. J. O'Grady, C. Muldoon, M. Dragone, R. Tynan, and G. M. O'Hare, "Towards evolutionary ambient assisted living systems," *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, no. 1, pp. 15–29, 2010.

[28] D. T. Handler, L. Hauge, A. Spognardi, and N. Dragoni, "Security and privacy issues in healthcare monitoring systems: A case study.," in *HEALTHINF*, pp. 383–388, 2017.

[29] S. Turner, R. Kisser, and W. Rogmans, "Falls among older adults in the eu-28: Key facts from the available statistics," *EuroSafe, Amsterdam*, pp. 1–5, 2015.

[30] S. Feldstein, *The global expansion of AI surveillance*, vol. 17. Carnegie Endowment for International Peace, 2019.

[31] I. Bouchrika, "A survey of using biometrics for smart visual surveillance: Gait recognition," in *Surveillance in Action*, pp. 3–23, Springer, 2018.

[32] A. Hampapur, L. Brown, J. Connell, S. Pankanti, A. Senior, and Y. Tian, "Smart surveillance: applications, technologies and implications," in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, vol. 2, pp. 1133–1138, IEEE, 2003.

[33] M. W. Green, J. Travis, and R. Downs, "The appropriate and effective use of security technologies in us schools," *US Department of Justice, Report NJC178265*, 1999.

[34] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.

[35] A. Alahi, V. Ramanathan, K. Goel, A. Robicquet, A. A. Sadeghian, L. Fei-Fei, and S. Savarese, "Learning to predict human behavior in crowded scenes," in *Group and Crowd Behavior for Computer Vision*, pp. 183–207, Elsevier, 2017.

[36] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3488–3496, 2015.

[37] R. Khirodkar, D. Yoo, and K. Kitani, "Domain randomization for scene-specific car detection and pose estimation," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1932–1940, IEEE, 2019.

[38] K. Goeschel, "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive bayes for off-line analysis," in *SoutheastCon 2016*, pp. 1–6, IEEE, 2016.

[39] W. Revelle and K. R. Scherer, "Personality and emotion," *Oxford companion to emotion and the affective sciences*, vol. 1, pp. 304–306, 2009.

[40] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[41] D. J. Ozer and V. Benet-Martinez, "Personality and the prediction of consequential outcomes," *Annu. Rev. Psychol.*, vol. 57, pp. 401–421, 2006.

[42] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[43] J. H. Block and J. Block, "The role of ego-control and ego-resiliency in the organization of behavior," in *Development of cognition, affect, and social relations: The Minnesota Symposia on child psychology*, vol. 13, pp. 39–101, 1980.

[44] H. Zacharatos, C. Gatzoulis, and Y. L. Chrysanthou, "Automatic emotion recognition based on body movement analysis: a survey," *IEEE computer graphics and applications*, vol. 34, no. 6, pp. 35–45, 2014.

[45] S. E. Seibert, J. M. Crant, and M. L. Kraimer, "Proactive personality and career success.," *Journal of applied psychology*, vol. 84, no. 3, p. 416, 1999.

[46] A. R. Sutin, A. Terracciano, M. H. Kitner-Triolo, M. Uda, D. Schlessinger, and A. B. Zonderman, "Personality traits prospectively predict verbal fluency in a lifespan sample.," *Psychology and aging*, vol. 26, no. 4, p. 994, 2011.

[47] K. E. Dill, C. A. Anderson, K. B. Anderson, and W. E. Deuser, "Effects of aggressive personality on social expectations and social perceptions," *Journal of Research in Personality*, vol. 31, no. 2, pp. 272–292, 1997.

[48] M. Merbaum and H. C. Lukens Jr, "Effects of instructions, elicitations, and reinforcements in the manipulation of affective verbal behavior.," *Journal of abnormal psychology*, vol. 73, no. 4, p. 376, 1968.

[49] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[50] M. Cristani, V. Murino, and A. Vinciarelli, "Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 51–58, IEEE, 2010.

[51] S. Yi, H. Li, and X. Wang, "Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance," *IEEE transactions on image processing*, vol. 25, no. 9, pp. 4354–4368, 2016.

[52] J.-K. Choi, K. Kim, D. Kim, H. Choi, and B. Jang, "Driver-adaptive vehicle interaction system for the advanced digital cockpit," in *2018 20th International Conference on Advanced Communication Technology (ICACT)*, pp. 307–310, IEEE, 2018.

2

INTRODUCTION TO STATE-OF-THE-ART AUTOMATIC HUMAN BEHAVIOR ANALYSIS

2.1. INTRODUCTION TO THE THEORETICAL FRAMEWORK

The performance of machine learning algorithms strongly depends on how the raw data is processed and represented. In order to keep only the most informative parts of the data, a common practice is to provide a collection of characteristics called feature representations. These features are used as examples for the machine to *learn* a specific task. For example, in order to detect objects from RGB images, the representation of objects as a set of features like colors, shapes, and sizes, is far more powerful than feeding the machine raw pixels information.

In conjunction with the feature representation step, machine learning models have to be designed and trained depending on the chosen learning strategy. When the data includes the solutions (labels) of a given task, a supervised learning strategy is applied. A typical supervised learning task is classification, where algorithms are designed to find the best separation between classes of labels. However, when dealing with real-world data, solutions have not always been added side by side with the data; in these situations, an unsupervised learning strategy is applied. A typical unsupervised learning task is clustering, where algorithms aim to find groups of samples with similar feature patterns. In this thesis, we apply the aforementioned strategies depending on the research goals and circumstances. For example, human motion is difficult to label as the specific movements may depend on multiple factors like contexts and external rules. In this case, an unsupervised learning strategy is adopted for the discovery of meaningful behavioral patterns (Chapter 3). Contrarily, applications like personality computing are designed to map personality labels, found with traditional psychological questionnaires, to specific actions and behaviors. In this case, a supervised learning strategy is adopted (Chapter 5).

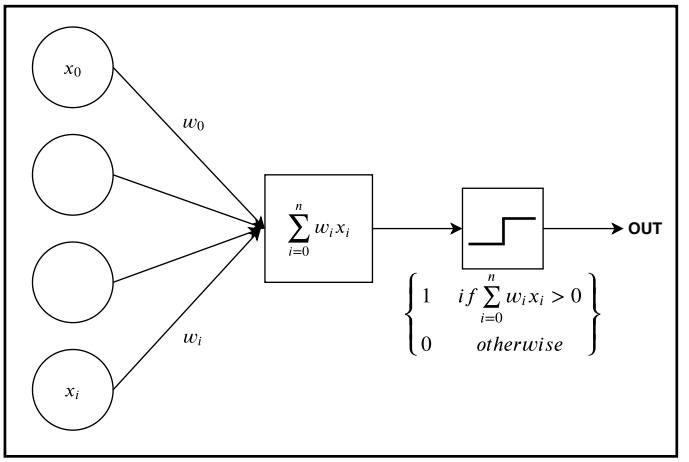


Figure 2.1: First mathematical model of a neuron, called Perceptron. The weighted sum is passed through a threshold function which filters the results to produce a binary output.

In recent years, the *Deep Learning* strategies have improved all the state-of-the-art results in the field of machine learning. Deep learning is a specific type of Machine Learning strategy [1] composed of deep hierarchical structures that are able to learn more meaningful data representations than the standard machine learning algorithms. The idea of deep learning is vaguely inspired by the functioning of the human brain cortex [2], and it consists in learning feature representations in a hierarchical way. Simple and basic features are learned in the early stages, and, the combination of these primary features allows the understanding of more complex and abstract features in the higher hierarchical layers [3]. In terms of computations, the deep learning strategy was motivated by important challenges like high-dimensional data and manifold representation and generalization [1]. As in this manuscript deep learning strategies are extensively used, in the next paragraphs, we will introduce the general deep learning structures as well as the specific techniques used to tackle our research questions. Specifically, we will first introduce the key ideas behind the Artificial Neural Network (Section 2.2), the unsupervised strategy of the Autoencoder Neural network (Section 2.3), the temporal power of the Long-Short Term Memory Network (Section 2.4) and finally, the Convolutional Neural network (Section 2.5).

2.2. ARTIFICIAL NEURAL NETWORK

The history of ANN dates back to 1943 [4] when the first mathematical model of a neuron called perceptron was proposed (Fig. 2.1). This model takes a set of inputs, multiplies each of them by a weight (that vaguely corresponds to the synaptic transmission between nearby neurons), and thresholds the sum of these weighted inputs to an output. The output is a 1 if the sum is above a certain threshold, otherwise, the output is a 0. The perceptron was proposed initially as a single layer, but later, the author in [2] proposed an extension called Multi-layer perceptron (MLP), in which a network of perceptrons

was proposed with the aim of modelling the visual perception phenomena.

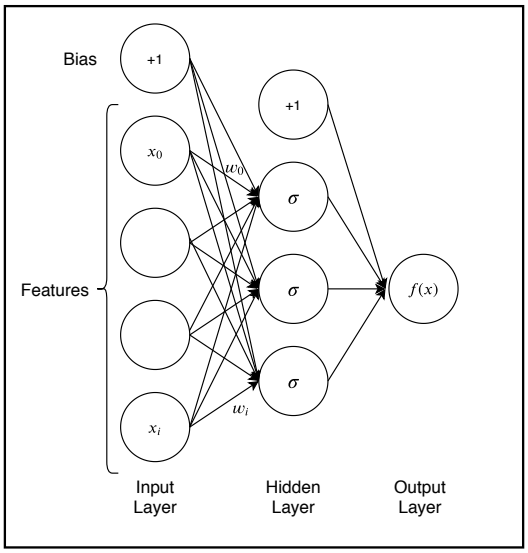


Figure 2.2: Multi-layer perceptrons (MLP) or Neural Network feedforward structure. In the simplest form, MLP is formed by three layers: The input layer, the hidden layer, and the output layer.

In the simplest form, MLP is formed by three layers (Fig. 2.2): the first layer is the input layer, which contains a set of units that correspond to the input data $X = [x_1, \ldots, x_n]$. Bounded to each input value, a set of random scalar values called weights $W = [w_1, \ldots, w_n]$ is initially assigned. The final goal of MLP is to learn these weights in order to minimize the error between the input and the output. A bias input is also present with a constant value of 1 that helps the model to fit the given data. The last layer is the output layer which contains the artificial neurons that are associated with the model outputs (for example, in case of classification, each output corresponds to one target class). As every input x is associated with a label y, the training examples, composed of pairs x, y, specify that the output layer's goal is to produce values as close as possible to y (usually called $\hat{y}$). Between the input and the output layer, there are the hidden layers, where all the core computations happen. Differently from the output layer, the behavior of the hidden layers is not directly specified by the training examples. The learning algorithm, through optimization strategies, chooses how to minimize the difference between the input and the output values. As the hidden layers estimate the output as a function of the input, but the function of the data is unknown, these layers are called "hidden layers" [1]. MLPs are also called neural networks because they are structured as an interconnected chain, and the overall length of this chain defines the depth of the model. These models are called feedforward because information flows from the input layer to the output layer, in other words, there are no feedback connections in which outputs of the model are fed back into itself. When feedforward neural networks are extended to include feedback connections, they are called recurrent neural networks (Section 2.4).

One of the significant differences between the original perceptron [4] and the more

modern artificial neural networks is the addition of an activation function a. As shown in Fig. 2.2, in each step, the weighted sum of the input is calculated, however, the result is not bounded to any limit and, therefore, it can take any value from -inf to inf. Computationally, the main role of the activation function is to restrict the results to a predefined range (e.g. $0,..,1$ or $-1,..,1$). Theoretically, the activation function decides whether the particular neuron is "activated" (high values) or not (low values). As the model calculates sequentially the activation from the input to the output, the activation of the j-th neuron in layer l (x_j^l) is calculated as following:

$$x_j^l = a(\sum_{i=0}^{n^{l-1}} x_i^{l-1} w_i^{(l-1,l)}) \tag{2.1}$$

where $w_i^{(l-1,l)}$ is the weight i that connects the neurons x_i and x_j in the consecutive layer l [5].

There exist many activation functions, the most used are: the sigmoid function (Eq. 2.2), the rectified linear unit (reLU) (Eq. 2.4), and the hyperbolic tangent (tanh) (Eq. 2.3).

$$f(z) = sigmoid(z) = \frac{1}{(1 + exp(-z))} \tag{2.2}$$

$$f(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{2.3}$$

$$f(z) = Relu(z) = max(0, z) \tag{2.4}$$

Another great advantage of the ANN with respect to traditional machine learning algorithms is the introduction of the end-to-end training strategy [6]. This procedure consists of finding an optimal configuration of all its weights W in all the connected layers, with respect to the training data X. The common training strategy for modern neural networks consists of two parts: first, we need a measure of the quality of the model, in other words, we need to know whether our network is delivering an output that resembles the training data. Therefore, we compute the *Cost function* or *Loss function L*, which defines the difference (error) between the input labels y and the predicted output $\hat{y}$. The Loss functions used in this dissertation will be briefly introduced in Section 2.2.1. Second, we need to send this error back to all the nodes, to update all the internal network's parameters minimizing the loss function. This process is called backpropagation and it will be briefly explained in Section 2.2.2.

2.2.1. LOSS FUNCTION

Given a supervised learning strategy, where we have a training set X and labels Y, we aim at fitting all the weights W to minimize the difference between the network's predicted output and the real labels. Specifically, the *Loss function L* is used to define the difference between the network prediction labels $\hat{Y}$ and the true labels Y. In this thesis, the cross-entropy loss function is used as follows:

$$L(\hat{Y}, Y) = -\sum_i y_i \log(\hat{y}_i) \tag{2.5}$$

where $\hat{y}_i$ is the probability of the predicted class i and y_i is the true probability for that class. In other words, the cross entropy loss function measures the distance between two probability distributions, the true class labels distribution and the predicted labels distribution, and, by minimizing, the final goal is to align them.

In the case of an unsupervised learning strategy, or in case the model's output is a continuous value, L aims at fitting all the weights W by minimizing the error between the real output and the predicted output. In these cases, the mean-squared error loss function is used as follows:

$$L(\hat{Y}, Y) = \frac{1}{2}\sum_i (\hat{y}_i - y_i)^2 \tag{2.6}$$

where $\hat{y}_i$ corresponds to the predicted output value and y_i corresponds to the real value.

2.2.2. BACKPROPAGATION

Backpropagation is the process that propagates the error between the predicted network output $\hat{Y}$ and the true labels Y from the last layer back to all the nodes. It has been showed by several types of research [7], that the optimal parameters configuration can be found using Stochastic Gradient Descent (SGD) optimization algorithms. The main idea of this class of algorithms is to iteratively update each particular weight w_{ij}^l to find the minimum of a function using a measure called gradient. It starts at a random location in the parameter space, and then, iteratively reduces the error until it reaches a local minimum. At each step of the iteration, it determines the direction of the steepest descent in the parameter space and takes a step along that direction. The gradient descent is computed as follows [5]:

$$\Delta w_{ij} = -lr \frac{\delta L}{\delta w_{ij}^{l+1,l}} \tag{2.7}$$

The partial derivative $\frac{\delta L}{\delta w_{ij}}$ computes the slope (gradient) of the function between consecutive layers $(l+1, l)$. lr is the learning rate parameter which gives control over the steps' size to take. Selecting the right learning rate is critical. If the learning rate is too large, we will take big steps risking to overstep the local minimum. While, if the learning rate is too high, the function will be more conservative, running more iterations of gradient descent, and therefore, increasing the training time.

2.3. AUTOENCODER NEURAL NETWORK

As introduced in the previous section, the ANNs goal is to find an optimal approximation of the input data X with respect to predefined labels Y. However, in several real-world situations the labels are not known a priori. Therefore, a sub-set of neural networks was dedicated to solving the challenge of unsupervised learning (i.e. learning without the supervision of labels).

In particular, the Autoencoders neural networks aim to learn the underlying structure of the data by learning an approximation to the identity function, in other words,

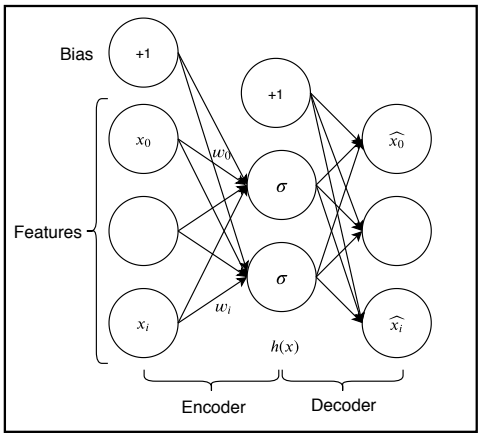


Figure 2.3: Autoencoder Neural Network architecture. The encoder and the decoder functions are designed to reconstruct the input features.

setting $Y = X$. This operation seems trivial, however, by setting constraints (or regularizers) to the hidden layers, a sub-set of meaningful and general features can be found. In the simplest form, for a single hidden layer autoencoder, the encoder f_0 and decoder g_0 functions are designed to reconstruct the input data X, represented as a vectorized set of input features $X = [x_1, \ldots, x_n]$, as good as possible (Figure 2.3). Therefore, given input data X, through the encoding step f_0, and the decoding step g_0, the goal is to obtain the reconstructed data $\hat{X}$ as follows:

$$h(X) = f_0(W_1 X + b) \tag{2.8}$$

where $h(X)$ represents the encoded set of features used to represent the input X. Subsequently, the decoding step is computed by the function g_0 and the reconstruction result is denoted by $\hat{X}$:

$$\hat{X} = g_0(W_2 h(X) + c) \tag{2.9}$$

$\{W_1, W_2\}$ are the weight matrices and $\{b, c\}$ are the encoding and decoding bias parameters. f_0 and g_0 are the non-linear activation functions. The optimization goal is to minimize the error between the input data X and the reconstructed data $\hat{X}$ using one of the loss functions explained previously (Eq. 2.6, Eq. 2.5). As the autoencoders are solely a special case of feedforward ANN, backpropagation strategy (Eq. 2.7) is used to minimise the used loss function.

In this manuscript, two types of Autoencoder structures have been studied: the undercomplete Autoencoder [8], and sparsity Autoencoder [9]. The former technique designs a network with a lower number of hidden units than the input units, while the latter technique forces all the hidden units to have minimal activations by using a sparsity parameter. The simple idea of an undercomplete AE is: when the dimension of h is smaller than the dimension of X, the network is forced to learn a "compressed" representation of the input data, and, therefore, filter out redundant information. On the other hand,

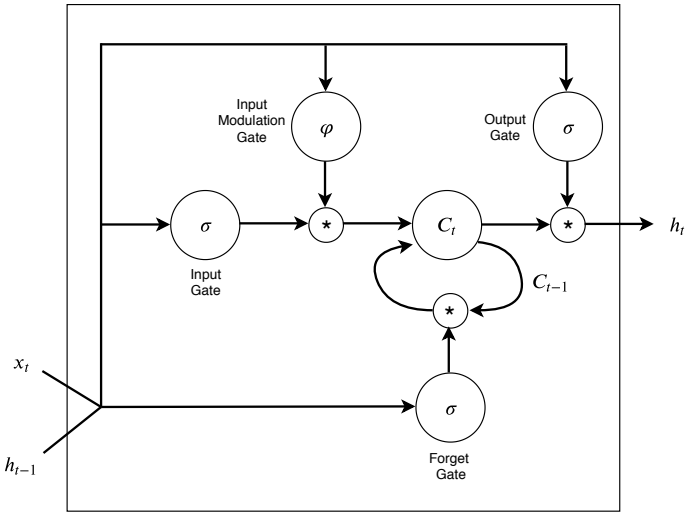


Figure 2.4: Internal activity of an LSTM module. Gated units allow the network to update its states, being able to store information over different time frames.

the sparse AE does not focus on the dimension of h but, rather, it adds a regularizer in the loss function which forces the units to all have a minimal activation. In this way, the hidden units are forced to avoid redundant information letting only meaningful information go through. AEs were shown to yield good performance in several real-world applications [10] due to the ability to learn meaningful representations of the data without labels. In this thesis, real-world data such as objects' motion trajectories and body postures are analyzed using AE networks. As these types of data require excessive labor for labeling frame by frame information [11], we leverage the AE networks for finding the underlying structure of the unlabelled data.

2.4. LONG SHORT-TERM MEMORY NETWORK

Differently from the feedforward neural networks, where the activation's flow goes only in one direction, from the input to the output, the Recurrent Neural Network (RNN) is a special type of method where the activation's flow is recurrent [12]. Given their structure, RNNs are specialized in analyzing temporal sequences. The general idea behind their structure is the parameter sharing across different parts of the networks [1], in this way, they are able to generalize information seen at different time positions in the temporal sequence.

In this thesis, we use a special type of RNN called Long-short memory network (LSTM) [13] that was shown to improve the RNN, being able to learn long term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem using gated units (Fig. 2.4). Gated units allow the network to store information (i.e. pieces of evidence for a particular feature) over a long duration [1]. However, once that information has been used, it might be useful for the network to update the old state. The most important part of the LSTM module is the cell state (C_t in the figure) which contains a

self-loop controlled by the forget gate that decides which information has to be updated and which are still important in the long-term period. In specific, the input information flows into the cell state regulated by the LSTM gates. The first gate is called the "forget gate" being regulated by a sigmoid function σ which sets the weights between 0 and 1 for each number coming from the previous cell C_{t-1}. High values show that the information is important and therefore has to be kept, while low values show that the information can be forgotten.

The input gate is responsible for the addition of new information to the cell state. This process can be divided into three sub-processes: First, similar to the forget state, the input information from h_{t-1} and x_t are regulated by using a sigmoid function 2.2. Second, a new vector is created containing all the possible values that can be added to the cell state (called input modulation gate in the figure) by using a tanh function 2.3. Third, information is added to the cell state C_t by multiplying the value of the regulated input to the created vector.

Finally, there is the output gate which has the role of selecting the output at time t. The functioning of an output gate consists of making a filter using the values of h_{t-1} and x_t, such that it regulates the values that need to be outputted from the cell vector. In particular, the LSTM learning procedure is computed as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ g_t &= \varphi(W_g \cdot [h_{t-1}, x_t] + b_g) \\ C_t &= g_t \circ i_t + f_t \circ C_{t-1} \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= \varphi(C_t) \circ o_t \end{aligned} \tag{2.10}$$

where (W_g, W_i, W_f, W_o) and (b_g, b_i, b_f, b_o) represent the weights and biases matrices of each type of gate.

The LSTM networks were shown to yield good performance in several real-world applications [14, 15] due to the ability to learn long term relationships in temporal sequences. In this thesis, we leverage the LSTM network for learning the dynamic structure of human movements, and affective computing.

2.5. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network (CNN) [16] is a class of feedforward neural networks specialized in the analysis of visual imagery or any input that has a grid-like topology [1] (Figure 2.5). The most important difference from the other neural network structures is the presence of *convolutional layers*. Convolutional layers are inspired by the human visual cortex which is formed by neurons that have local receptive fields. In other words, neurons are organized in groups that focus on limited regions of the visual field. The receptive fields of different neurons may overlap, and together they cover the whole visual field. Units in the first convolutional layer are not connected to every pixel in the input image like an ANN [1], but instead, the units focus only on pixels in their receptive fields (called also kernels K). Kernels are usually smaller than the input image, and they are

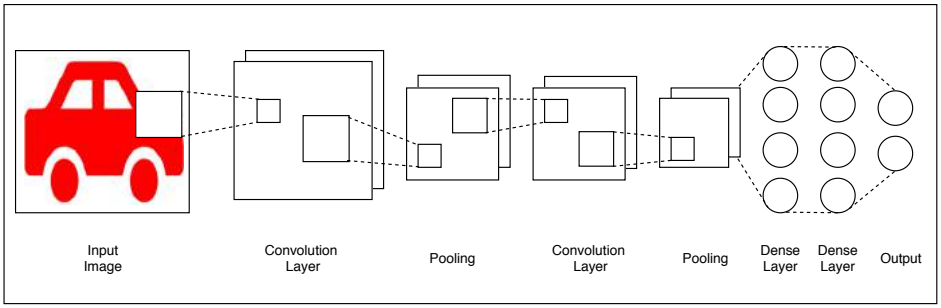


Figure 2.5: Architecture of a Convolutional neural network framework. Convolutional layers are formed by neurons that focus on limited regions of the visual field. Every convolutional layer is summarized using a pooling function. As the CNN contains fully connected layers, the output of the first convolution layer becomes the input of the second. In this way, the network is able to hierarchically learn a set of features.

applied in an overlapping manner to cover the whole input. For example, if we have a 2D image as input of 100×100 pixels, kernels will cover 10×10 pixels in overlapping locations (see the first step in Figure 2.5). This process yields several advantages [1]: Firstly, kernels that embed smaller part of input require fewer parameters reducing the memory requirements of the model. Secondly, this process improves the statistical significance by leveraging the fact that the pixels close to each other are more semantically connected than pixels further away. In this way, CNNs can learn more meaningful features. Lastly, the same kernel parameters are used at every position of the input rather than learning a separate set of parameters for every location. This property makes the convolutions learn features that are robust to translation [1].

The last step of a convolution is called *pooling*. A pooling function is simply a statistical summary of the nearby outputs. For example, the max-pooling operation [17] reports the maximum value within a rectangular neighborhood, or, the mean-pooling operation [18] reports the mean value within the neighborhood. This operation makes the convolutional layers learn robust features invariant to noise.

Figure 2.6 shows an example of the hierarchical learning structure of a CNN architecture. Early Convolution layers learn low-level features such as edges and curves. These features do not carry any semantic information yet, however, as they represent very basic characteristics of objects and images, they can be applied to different/unseen domains (transfer learning). As the CNN contains fully connected layers, the output of the first convolution layer becomes the input of the second one. In this way, the network is able to hierarchically learn a set of features. For example, the low-level features are combined and summarized to learn mid-layer features and so on. Low-level features could be semicircles (combinations of a curve and straight edge) or squares (combinations of several straight edges). As we go through the network and advance to higher convolutional layers, more and more complex features are formed by combining previous layers' features. By the end of the network, more high-level features (semantically connected with the labels) are learned, and ready to be passed to an inference system.

CNN frameworks were shown to yield good performance in several real-world applications [19, 20] due to their ability to learn meaningful and robust image features. In

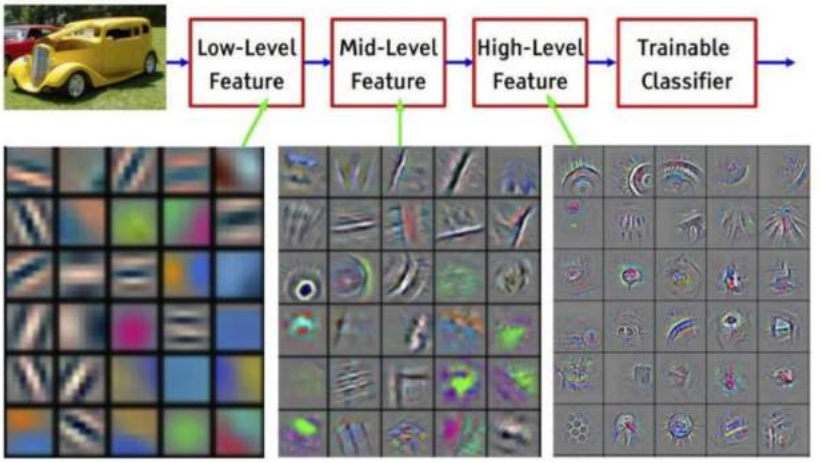


Figure 2.6: Hierarchical learning power of CNN architectures (picture from [21]). As we go through the network and go through more convolution layers, more and more complex features are formed by combining previous lower-level features.

this thesis, we leverage the CNN networks for learning human skeleton motion as well as contextual information from video data.

2.6. LITERATURE REVIEW

In the last decade, machine learning and especially deep learning methods showed impressive results in real-world scenarios. In this chapter, we will focus on describing the state-of-the-art algorithms in the studied domains such as AAL, smart surveillance, as well as personality computing.

2.7. AMBIENT ASSISTED LIVING

Europe has one of the highest portions of elderly population in the world [22]. In 2016, 19% of the European population was over 65 years old. The oldest-old (above 80 years) will increase from 5% in 2016 to 13% of the population in 2070 [23]. This tendency will introduce new challenges in the healthcare system that governments will have to solve. For example, as the senior citizens will continue to grow in number, the need of resources such as facilities and funds will continue to increase across all healthcare settings. Additionally, healthcare professionals will not be enough and, consequently, the quality of their services will decrease. To cope with these challenges, the European Union launched research and innovation programs (e.g. H2020 program) to motivate world-class science and industrial innovation to help in bringing new solutions to the table. In this framework, my PhD was funded by the ICT4Life European project [24] that aims to provide innovative smart ICT services enhancing independent living among the aging population.

Elderly living independently remain most of their time alone, while medical professionals do not have resources to visit them more than few hours a week. As a conse-

quence, clinical pictures required to assess the independent living skills are challenging to make and are often incomplete. Sensor-based technologies systems which quantify Activities of Daily Living (ADL) can be of a great help to existing clinical assessments. Results showed that constant monitoring has several advantages such as early detection of diseases and health related risks [25], and lower cost of medical care [26], among others. Finally, the integration of this information with the current clinical picture of the patient (i.e. integrated care) can help doctors and caregivers to make better decision in the assessment and therapy of the diseases.

Numerous works for ADL recognition in home settings have been reported in the last years with various monitoring technologies: passive infrared motion sensors (PIR), wireless sensor network (WSN), body-mounted sensors, pressure sensors, and video monitoring.

WSN-based systems are usually formed by a set of sensors dispersed in the environment to monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, or motion. They usually rely on wireless connectivity for their communication with a central location. WSNs have the advantage of being non-intrusive and privacy compliant, however, ambient sensors connected through the WSNs usually provide little information about activities. Ambient sensors can be used for the detection of daily activity patterns. For example, authors in [27] studied the daily routine of smart homes inhabitants by placing motion sensors in all the rooms of the house. Binary signals (On/OFF) were sent when the target person was detected by the motion sensors. The temporal sequences of these activations indicated the daily routine patterns of the inhabitants. Similarly, authors in [26] detected the differences in ADL patterns in elderly with dementia, highlighting a significant discontinuity in activities like eating and sleeping in cognitive impaired subjects compared to healthy subjects by using sensors placed on drawers and domestic appliances.

Body-worn sensor systems have the ability to measure body-activity and body-mobility as they are positioned directly on the human body. Accelerometers are an example of this type of technology, as they can be positioned on the human body to detect activity by measuring (linear) accelerations in bodily movements [28]. Applications with accelerometers include fall detections [29], and gait quality detection [30]. In recent years, due to the progress of signals' quality as well as chips' size, wearable sensors have become reliable in the measurements of physiological signals (e.g. heart rate and body temperature). Authors in [31] designed a wearable device to monitor the heart rate and body temperature providing emergency assistance for the elderly. Authors in [32] utilize measurements from wearable sensors such as galvanic skin response (GSR), and skin temperature in conjunction with ambient sensors to measure the patterns of sleep and stress of people affected by dementia.

Video based monitoring is another popular solution for AAL systems. The image signal is rich in information and, therefore, can be used to retrieve important information about daily life [33]. One disadvantage of this technology consists in being difficult to accept by the senior citizens due to its intrusive nature. Video based systems have been proposed with different goals, from broad descriptions of events [34] to fine-grained definition of daily-life actions [35]. For example, authors in [36] use video data for weakly supervised segmentation and detection of ADL in long video sequences. As the human supervision

of long videos is a tedious task, the broad segmentation and recognition of ADL are essential tasks for real-world smart systems. Similarly, authors in [34] study the semantic summary of activities and daily patterns using human motion analysis.

The automatic understanding of daily activity patterns is a crucial task in healthcare as it allows the detection of changes in the subjects' life. For this reason, the ICT4Life project aimed to create smart monitoring systems able to detect abnormal events in elderly with dementia. As society is steadily aging, the number of senior citizens with related diseases is increasing as well. On top of this, a quite common phenomenon is that individuals with early dementia live by themselves or spend many hours alone in hospitals or daily care centers. One of the most common symptoms in dementia and, more in particular, Alzheimer's, is that the patient wanders in a disoriented manner. This phenomenon is very dangerous as it can cause important accidents such as falling down or getting lost outside their home.

Continuous monitoring systems have been shown to be of great help in the detection of these dangerous events. Authors in [37] use trajectory features to detect movements' patterns associated with confusion states. In order to reach a certain destination, signs of wandering can be detected when patients take random and inefficient travel paths instead of a direct one. Similarly, authors in [38] analyzed walking path and motion energy over time for the detection of wandering behaviors. Results showed that walking path of confused subjects contains circular segments and overall, patients showing confusion states walk slower than a person without confusion. In Chapter 3 and Chapter 4 we will explain in detail the monitoring solution built in the ICT4Life platform.

2.8. VISION-BASED SMART SURVEILLANCE

Tackling automatic vision-based surveillance remains a challenging topic within the computer vision community [39, 40]. Many tasks such as object tracking, manual annotations, multiple cameras integration, still remain to be fully solved. Monitored scenes such as busy road intersections or public spaces are highly complex, hence, plenty of works have been proposed using different types of cameras, modalities, and system architectures. An important component of any automatic surveillance system is tracking, which has been addressed using a wide variety of methods. Depth sensors (such as Microsoft Kinect or Intel sensor) were released in order to be able to achieve considerable tracking performance in a convenient and low-cost embedded system. The tracker can detect and track up to 25 human body joints in up to 6 subjects at the same time, and it works by classifying each pixel of the depth image as part of a joint using trained decision forests [41].

Another popular tracking approach is Kalman filtering [42], which estimates the velocity and the unknown state of an object by modelling the statistical characteristics of the system in combination to noise measurements. However, these tracking algorithms rely heavily on people/objects detection and segmentation, and therefore, can be sensitive to noise and occlusions. Optical flow [43] was proposed to track motion dynamics between consecutive frames, by looking at the image intensity as a function of space and time. Optical flow methods showed to be useful for the understanding of moving crowds and when targets' motion is explicit. However, important challenges such as camera motion sensitivity and background motion still need to be tackled [44]. When the tracking

algorithms lose the tracked target, only short consecutive fragments of motion can be extracted. These fragments of information are called tracklets and, tracklet-based methods are a trade-off between object-based tracking and optical flow-based approaches [45]. In the next sections, we will introduce the state-of-the-art algorithms related to our research on behavior motion understanding.

2.8.1. TRAJECTORY BASED ANALYSIS

Once tracking information has been extracted over time, several features have been proposed to describe motion events. A set of simple heuristics can be used on the extracted motion information. Histogram of Oriented Flow (HOF) [46] is probably the most famous example, where flow information is discretized in predefined histograms. The components of the optical flow's output $I^w = (I^x, I^y)$, denoting the estimated motion in the 2D image $w = (x, y)$, are treated independently. The gradient's information is extracted separately, and, through a weighted vote strategy, is discretized into local orientation histograms (in the same way as for the Histogram of Oriented Gradient (HOG) [47]). Following the same concept, the Histogram of Tracklets (HOT) [45] was proposed to quantize the magnitude and orientation information contained in short trajectories (i.e. tracklets). HOT has been shown to create an intermediate feature layer able to capture the dominant motion over a short period of time. Unlike HOF, which estimates motion over two consecutive frames, HOT utilizes longer range motion trajectories estimation. The extraction process is described in Figure 2.7. Tracklets are fragments of complete trajectories generated when the tracking algorithm, due to detection failures, stops to track the target (Fig. 2.7 (a)). More formally, a tracklet can be represented as a sequence of points in time, $tr^n = [p_1^n, .., p_t^n, .., p_T^n]$, where, each p_t^n is a sequence of 2D coordinates (x_t, y_t) over a sequence of frames $t_1, t_2, .., T$ for a certain individual n.

As different regions of the scene usually contain different motion patterns, authors in [45] utilize spatio-temporal cubes $S_x \times S_y \times T$ to encompass all the tracklets in the data. The cuboids do not overlap spatially (S_x, S_y), but only temporally (T). For each cuboid, inspired by HOF, the gradient magnitude (Fig. 2.7 (a)) and orientation (Fig. 2.7 (b)) is computed and quantized independently in histogram bins (Fig. 2.7 (c)). In equations 2.11 and 2.12 the features calculation is shown in detail. To compute the orientation values, the authors use only the entry and exit points of each tracklet in each cuboid (Eq. 2.11). On the other hand, magnitude values are calculated between consecutive points within given cuboids, however, only the maximum value is stored in the final histogram (Eq. 2.12).

$$\Theta^{i,s} = arctan \frac{(y_{end}^{i,s} - y_{begin}^{i,s})}{(x_{end}^{i,s} - x_{begin}^{i,s})} \tag{2.11}$$

$$M^{i,s} = \max_{t,t+1 \in T} \sqrt{(x_{t+1}^{i,s} - x_t^{i,s})^2 + (y_{t+1}^{i,s} - y_t^{i,s})^2} \tag{2.12}$$

where (i, s) refer to the point of tr^n that intersect with the cuboid s, and $x_{begin}^{i,s}$, $y_{end}^{i,s}$ indicate the entry and exit points of tracklet i in/from cuboid s. The authors quantize the obtained orientation values in OR = 8 bins and the magnitude values in MA=3 bins. Finally, the two histograms are concatenated to obtain a feature matrix of size $OR \times MA$

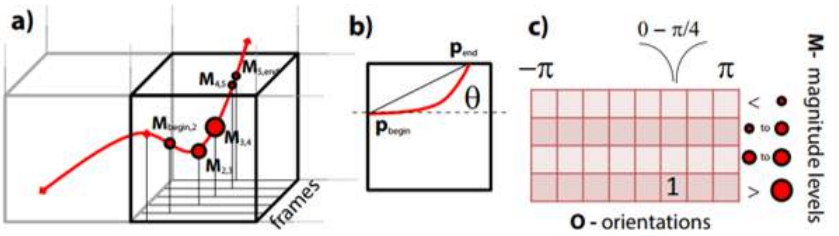


Figure 2.7: Histogram of Tracklets (HOT) proposed by [45]. For each trajectory segment called tracklet (a) The gradient magnitude and orientation is computed (b) Magnitude and orientation values are quantized in histogram bins (c).

(Fig. 2.7 (c)). An advantage of using histograms to learn trajectory motion is that they are built using a discrete fixed size, being independent from the length of the motion events. Several surveillance applications were proposed using these methods, such as Abnormal behavior detection and crowd motion analysis [45]. However, one main disadvantage of histogram-based approaches is that they do not consider the dynamic characteristics of motion.

Modeling the progressive evolution of motion is important because it allows to obtain more fine-grained information about complex movements such as human activities, and behavior intents [48]. There exist several approaches to learn the temporal dynamics of motion, however, in this thesis, we will cover mainly two: Neural Network based temporal learning, and probability based temporal learning.

In the last years, deep-learned motion features have shown valuable performances making use of different architectures such as CNN [49, 50], Autoencoders [51, 52] and Long Short-Term Memory networks (LSTMs) [53, 54].

The most straightforward NN solution when dealing with temporal data is RNN and LSTM networks (described in Section 2.4). Authors in [54] propose an LSTM framework for trajectory prediction and abnormality detection. The authors define the problem as a sequence to sequence prediction. In addition to the trajectory sequence, learned using the LSTM, a set of attention models is added to encode the neighbour's influence on the predicted path of the pedestrian of interest.

LSTM networks showed limitations when learning both short-term as well as long-term relation within sequences, hence, authors in [53] create an augmented long-term memory, proposing a tree memory network that hierarchically selects useful information of the past. Useful information is defined as historical behaviors that show similar contexts and evolution to the current information. Specifically, the authors hierarchically map the memory with a bottom up tree structure. All historical states are represented in the bottom layer of the tree, and, as progressing up the hierarchy, the most significant features are concatenated to predict the output at a particular time step.

In order to encode the spatial and temporal information at the same time, several works proposed the use of convolutional operations to efficiently model them. Authors in [49] built a behavior-CNN to learn the distribution of the past path to predict future pedestrian walking paths. The input to the system is pedestrian walking paths in previ-

ous frames obtained using tracking frameworks. Walking paths are then encoded into a 3D displacement volume $D = X \times Y \times t$, where $[X, Y]$ is the spatial size of the input frames and t indicates a time-window. Behavior-CNN takes the encoded displacement volume as input and predicts an output displacement volume for all the pedestrians simultaneously. A behavior decoding scheme then translates the output displacement volume to future walking paths of all individuals.

Physical obstacles present in the scene obviously influence human walking paths. To tackle this challenge, methods like the one proposed in [50] use the encoding power of CNNs to learn regions suitable for walking. This method proposes two CNNs to encode the spatial as well as the orientation dynamics of moving objects. The first CNN estimates rewards of local regions by comparing similarity between the patch of the target and surrounding patches. The second CNN estimates the future orientation of the object. At testing time, the future path is estimated as an optimization problem of planning a path with the lowest cost (where the cost is determined by obstacles in the scene).

Autoencoders constitute a powerful tool that can be used to learn temporal sequences in an unsupervised fashion. Authors in [52] propose an autoencoder framework to learn regularity in video sequences. The model learns to reconstruct the motion signatures present in regular videos with low error, while, it fails to accurately reconstruct irregular motions. In other words, the autoencoder reconstruction error can be used to model regular dynamics and detect abnormal dynamic changes.

Autoencoders can be also transformed in generative models. Examples of them are Variational autoencoders (VAE) and Conditional Variational Autoencoders (CVAE). Generative AEs have the advantage of estimating the probability distribution of future samples given the past data. For example, authors in [51] use a CVAE aiming to learn a conditional probability distribution of future trajectory outputs given trajectory historical observations.

Modelling future trajectory path as a probability problem is one of the most explored approaches as it allows to mine recurrent activities with much less training data in respect to NN models. Authors in [55] aim to model the motion and interaction between groups of stationary and moving pedestrians. A general energy map is proposed to learn the traveling cost from, and to, each location of the scene. Regions with higher energy values denote that pedestrians can travel through these locations more easily producing greater motion energy. Lower energy values indicate locations with lower occurrence probability. For example, areas near an obstacle or inside a stationary crowd group are difficult to walk through. Finally, for every pedestrian, the most probable walking path based on the energy map is generated simulating the pedestrian decision making process.

The study proposed by [56] uses a hierarchical Dirichlet process to associate co-occurring motion attributes such as speed, orientation, and location, to activities. First, trajectories are clustered into a finite set of activities. Second, the trajectories in the clusters are modeled using the posterior Bayesian probability for applications in anomaly detection and path prediction. In the study of [57], a hierarchical architecture on trajectory data is used for semantic region discovery. The authors adopt the concept of a hierarchically linked infinite hidden Markov model, which can capture the temporal dependency between adjacent observations detecting regions of the scene that have seman-

tic power. As different regions are likely to show different motion patterns, discovering regions connected to certain semantic behaviors can be helpful in several surveillance applications.

Trajectory information has been shown to provide useful cues for several computer vision applications. In this thesis, we propose novel architecture that aim to contribute to the task of motion modelling as well as motion understanding in different scenarios. Within these tasks, there exist several challenges that are far from being solved. For example, in Chapter 3, we will present novel spatio-temporal features that allow us to extract high level motion information such as stationary behaviors (sitting, working at the desk) as well as active behaviors (walking, exiting the space). This information will be mapped to a video-based surveillance system that detects normal vs. abnormal behavior in different situations. In Chapter 4, we will improve the embedding of spatio-temporal motion features using Autoencoders, and we will tackle the real-time prediction of future trajectory paths.

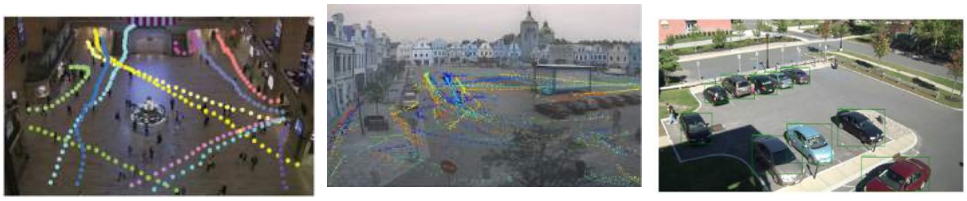
2.8.2. SURVEILLANCE DATASETS USED IN THIS DISSERTATION

Trajectory analysis from surveillance cameras is a topic that received much attention in the last decade. Consequently, several public benchmarks were proposed to tackle challenges like abnormal behavior detection and crowd flow understanding. In this thesis, we take into consideration three different public surveillance datasets: the Long-term Observation of Scenes (with Tracks) or LOST dataset [11] containing outdoor scenarios, the GC dataset [55], collected in the Grand Central Train Station in New York city, and VIRAT Surveillance Dataset Release 2.0 [58]. LOST [11] is a publicly available dataset that includes 24 streaming outdoor web-cams from different locations in the world. Trajectory information as well as the bounding box of moving objects were extracted over a long period of time (more than 1 year). The reason we chose to analyze our proposed methodology on the LOST dataset is because it offers long-term tracks in different outdoor scenarios, while there is limited research work dealing with abnormal behaviour detection on it (Figure 2.8b). We follow the same experimental setting used in [59], by analyzing only two cameras, "camera 001" (Ressel Square, Chrudim, Czech Republic) and "camera 017" (Havlickuv Brod, Czech Republic).

In [55], a large-scale dataset with pedestrian walking routes is described and made available. The GC dataset is collected from the Grand Central Train Station in New York city (Figure 2.8a). A surveillance video of one hour was manually annotated as ground-truth. The data contains 12,684 pedestrians with an average of 123 pedestrians in each frame. For each individual, the complete trajectory from the time point a pedestrian enters in the scene to the time he/she leaves is labeled. This allows long-term trajectory prediction experiments that take into account source and destination of pedestrians.

VIRAT [58] is a public dataset collected in multiple outdoor scenes, where different objects (i.e. vehicles, pedestrians) are recorded in cluttered backgrounds (Figure 2.8c). Data was collected in natural scenes showing people performing normal actions in standard contexts, with uncontrolled, cluttered backgrounds. It contains two broad categories of activities (single-object and two-objects) which involve both humans and vehicles.

As trajectory datasets containing normal/abnormal events in private indoor envi-



(a) New York GC Dataset [55]

(b) Long Observation of the Scene (LOST) Dataset [11]

(c) VIRAT dataset [58]

Figure 2.8: Data examples extracted from the surveillance datasets used in this dissertation

ronments are difficult to find, we recorded a dataset by tracking people in an office room during working hours using the Microsoft Kinect Sensor (SDK 2.0). This dataset (called KIMOFF) was only used for initial experiments in Chapter 3 and, therefore, was not made public. KIMOFF contains data trajectory belonging to the head joint due to the camera position and the context of the experiments (an indoor environment where people are often sitting at their desk and half of the body is occluded). Trajectories from twenty-four working days were recorded from 9 a.m. to 6 p.m., workers were aware of the camera but, as only trajectory points were saved, their privacy was not invaded. The workers acted normal, as the purpose of the recordings was to capture a real-life situation and not artifacts.

2.9. HUMAN MOTION ANALYSIS

Human motion recognition is one of the most important challenges in the computer vision community because of its great applicability in several real-world challenges, such as video surveillance and activity recognition [60]. In Section 2.8.1, we introduced trajectory based methods that provide meaningful insights about motion towards a direction and its associated intents. Although trajectory data is important for general surveillance applications, it does not provide rich information about the articulated motion of the human body. Therefore, in this section, we will introduce human body posture related features (namely, skeleton data), which provide more fine-grained insights about human body motion and actions.

The automatic recognition of handing an object to another person, playing sports, or simply walking in a crowded environment would be extremely challenging without the understanding of how the human body moves [61]. In the past years, several features have been adopted to investigate human body motion, including RGB-based pose features, depth-based pose features and skeleton-based pose features among others. Skeleton pose estimation has been shown to be more robust to noise and occlusions [41]. Additionally, it provides semantic indexing of human body parts, allowing an easier interpretation of actions' movements and sub-movements. Hence, in this thesis, we will base our human motion models mainly on skeleton-based features.

2.9.1. SKELETON-MOTION FEATURES

As human behaviors have a temporal evolution, a plethora of models focus on the investigation of skeleton joints dynamics using temporal models such as LSTMs, and Gated

Recurrent Units (GRUs). The authors in [61] address the short-term motion prediction of skeleton sequences using a sequence-to-sequence (seq2seq) architecture. The standard structure of seq2seq frameworks involves two networks, firstly, an encoder receives the inputs and generates an internal representation and secondly, a decoder network takes the internal state and produces a maximum likelihood estimate for the prediction. Additionally, the authors explore training a single model to predict motion for multiple actions, in contrast to building action-specific models. While modelling multiple actions is a more difficult task than modelling single-action sets, this allows the network to exploit regularities between human motions gaining higher semantic knowledge.
Authors in [62] represent the complex motion of humans over spatio-temporal graphs and model them with an RNN framework. Spatio-temporal graphs are used to impose a high-level structure to the motion (i.e. related motion of arms and legs during walking), while a RNN is used to learn all the relations within the graphs (i.e. edge node connections). This approach has the advantage of utilizing structured components in which domain experts can inject their high-level knowledge in the learning frameworks. Despite the evident temporal power of RNN models, their generalization ability is still a research focus [63]. Specifically, as RNN-based methods use only series of coordinates of skeleton joints, a considerable portion of RGB information is discarded. To overcome this issue, data augmentation and transformation strategies have been proposed. Authors in [63] extend traditional LSTM networks presenting an LSTM-AE framework for spatial-temporal data augmentation. In the LSTM-AE topology, the LSTM network preserves the temporal information of skeleton sequences, while the autoencoder architecture is used to filter irrelevant and redundant information. Similarly, authors in [64] propose an AE framework to learn a nonlinear reduction of movement primitives. The AE constraints force the model to learn smoother and denoised movement representations, while also missing values (e.g. due to occlusions) can be reconstructed using the learned latent space.

As the latter examples show, data augmentation using image information yields an improved representation of temporal dynamics. In this direction, novel strategies have been explored in order to use powerful tools like CNNs for skeleton dynamics learning. An interesting approach is to transform human motion values into indexed images, where the skeleton joints IDs are mapped onto the y-axis, while their temporal evolution is mapped onto the x-axis.

Authors in [65] propose a spatio-temporal representation of skeleton joints called "image clips" for action recognition. Given a skeleton sequence, the skeleton joints of each frame are first arranged as a chain by concatenating the joints of each body part. Four reference joints are chosen, namely, the left shoulder, the right shoulder, the left hip and the right hip, to compute relative positions of the other joints. These relative distances reflect the motions of the other joints. For example, the action "punching" will be represented with an increasing distance between one arm and the stationary reference points on the body. Finally, as shown in Fig. 2.9(A), the distance values are arranged as an image. For every joint ID on the human skeleton, depicted on the x-axis, the frame-by-frame distance evolution is depicted on the y-axis. Each clip represents short-term temporal skeleton sequences, as well as the local interaction between body joints. The long-term temporal structure of the skeleton sequence can be effectively

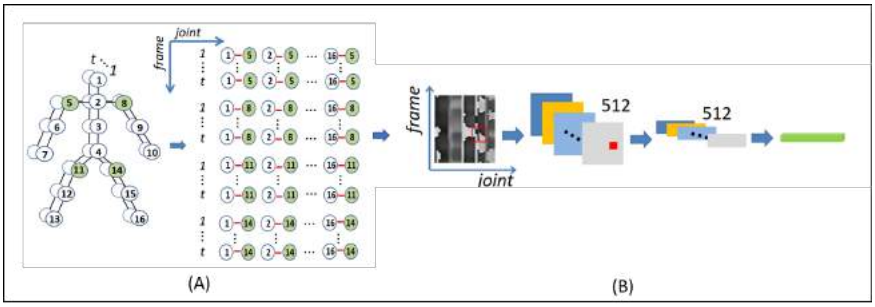


Figure 2.9: The feature extraction strategy proposed by the authors in [65]. (A) Clip Generation of a skeleton sequence. The skeleton joints of each frame are first arranged as a chain by concatenating the joints of each body part. Four reference joints, shown in green, are then used to compute relative positions of the other joints to incorporate different spatial relationships between the joints. (B) A pre-trained CNN model is used as a feature extraction step.

learned by using deep CNNs. This strategy allows the authors to use pre-trained CNN models [66] as a stronger feature extraction tool (Fig. 2.9(B)). Given the generated clips, a deep pre-trained CNN (VGG19 [66]) is leveraged as feature extractor. The output of the convolutional layer $conv5_1$ is used as the representation of the input frame. This is a 3D tensor with size $14 \times 14 \times 512$, i.e. 512 feature maps with size 14×14. Each feature map activation in the $conv5_1$ layer corresponds to local regions in the original input skeleton image. Hence, the skeleton temporal information still resides in the feature maps and it can be extracted from their rows. Specifically, the authors perform a novel pooling strategy, called Temporal Mean Pooling (TMP), applied to generate a compact representation of the row/temporal dimension with a kernel of size 14×1. Finally, the 512-dimensional feature maps are concatenated to form a 7168 ($14 \times 512 = 7168$) feature vector, and fed into the learning network.

Similarly, the authors in [67] constructed a skeleton motion image to benefit from pre-trained CNN models. Skeletal images (called Skepxels) are constructed by organizing a set of distinct skeleton joint arrangements from multiple frames into a single tensor. Unlike previous works where skeleton joints of a frame were arranged in a column, the authors arrange them in a 2D grid to take full advantage of the 2D kernels in CNNs. The temporal evolution of the joints is captured by employing Skepxels from multiple frames of the sequence into one image. The authors exploit a wide variety of existing CNN architectures to effectively process the information in the skeleton frame sequences, such as Inception [68], and ResNet [69].

Authors in [70] create image descriptors containing joint rotations over temporal windows. These descriptors are defined like motion textures and treated in a similar way to images. Each column represents the rotation degree of every joint, and each row represents time information. Authors define these image descriptors as *motion words* embedding a narrow temporal-window of a group of joints. In contrast to single pose feature-sets, motion words represent the pose local evolution in time, facilitating the learning of spatio-temporal properties of the motion within the framework's cost function. Using a deep metric learning based approach, they create an optimized

latent space where similar motion words are placed in the close vicinity, while different motion-words are placed further away.

The multitude of methods, as well as the important amount of available data proposed for human motion recognition opens the way to transfer the acquired knowledge to more general behavior analysis. For this reason, in Chapter 5, Chapter 6, and Chapter 7, we chose to explore the relation between skeleton motion image features and personality attributes.

2.9.2. SOCIAL AND NONSOCIAL INTERACTION

Human motion is influenced both by internal commitments, such as being late for an appointment, as well as by external factors, such as finding groups of people obstructing the path. Several studies investigated the influence of the social surroundings on the prediction of future paths and future behaviors [71, 72]. Authors in [71] propose an LSTM network which can jointly predict the paths of all the people in a scene by taking into account common sense rules and social conventions that humans typically utilize as they navigate in shared environments. In particular, they utilize a "Social" pooling layer which allows spatially related LSTMs to share their hidden-states with each other. This architecture was shown to learn the interaction between trajectories that coincide in time and space. Similarly, authors in [73] construct an LSTM hierarchical model to embed the person, the social, and the scene information for trajectory prediction. The motion information of the target pedestrian (speed, acceleration) is modeled at the same time with the neighbours' motion information. Moreover, the authors take into consideration the spatial affinity between the target and all the other pedestrians. The spatial affinity measures the level of influence of the social context on the target pedestrian.

Social interactions are rich in information, revealing characteristics of both single individuals as well as groups. In [74], the authors propose an RNN based system to model, at the same time, individual and group activities. The action of an individual in a group is influenced by the actions of the other individuals within the group. This phenomenon not only can provide context to recognize the individual actions, but also provides a key information about the group level activities. They proposed a structural RNN formulated as a two-level hierarchy. The lower level predicts individual actions followed by the higher level recurrent network that estimates the group activity.

Similarly, in [75], an LSTM system is proposed to model intra-group dynamics (person within a group) and inter-group (group to group) interactions. To model group-to-group interaction, the authors apply clustering/segmentation method to partition all human tracklets into spatio-temporal consistent groups. After that, a hierarchical recurrent context encoding network is proposed to learn the interactions in the scene. This model aims to encode single human dynamics (change of appearance and movements), the intra-group human interactions (poses and neighbors movements), and finally, the inter-group interactions.

Authors in [72] focus on the relationship between physical and social distance (proxemics) in social gatherings. Distance, as a social relation cue, means that people tend to unconsciously organize the space around them based on different degrees of intimacy. In other words, the more two people are intimate, the closer they get. Authors divide the interpersonal distance in different zones of intimacy finding consistent results be-

tween social group formations and intimacy levels. Similarly, how people use and share their interpersonal space has been shown to be a discriminative cue for personality understanding [76]. The authors identify a set of proxemics features such as: minimum distance to neighbours, velocity, number of persons with whom the target is holding different kinds of relationships, etc. This set of features is used for the prediction of two personality traits: Extraversion and Neuroticism.

As humans are by nature social beings, behavioral displays have been well explored during social interaction. However, few efforts have been made towards the understanding of behavioral patterns in nonsocial contexts. This problem has been shown to be important for applications in domains like Ambient Assisted Living (AAL) and Smart Homes, where it often occurs that individuals are spending a lot of their time at home alone.

When it comes to nonsocial behavior understanding, human-object interaction has been proposed to be an interesting cue that could explain how humans act in nonsocial contexts. In [77], the correlation between actions and context has been explored, showing how actions are constrained by certain scenes. The authors apply bag-of-feature approaches (i.e. SIFT, HOG, HOF) on both human motion and context. Results show that automatically extracted context descriptors improve the action recognition task.

Similarly, authors in [78] propose to model the mutual context between objects and human poses in human-object-interaction activities such that each can facilitate the recognition of the other. Specifically, two contextual pieces of information are considered: 1) The co-occurrence context models the co-occurrence statistics between objects and specific types of human poses within each activity. 2) The spatial context, which models the spatial relationship between objects and different human body parts. Results show that these features improve the performance of both object as well as action recognition.

Following these studies, in Chapter 6 and Chapter 7, we will study the relation between Person-Context interaction dynamics and behavioral patterns in social and nonsocial environments.

2.10. PERSONALITY COMPUTING

Extensive studies in the field of psychology showed that attitude, mood, and personality are directly connected to human behavioral patterns [79]. Since these human characteristics are often subtle, the affective computing field still faces several challenges in order to transfer theoretical models into computational frameworks. Authors in [80], in their introduction to personality computing, define personality as: *"a psychological construct aimed at explaining the wide variety of human behaviors in terms of a few, stable and measurable individual characteristics. In this respect, any technology involving understanding, prediction and synthesis of human behavior is likely to benefit from Personality Computing approaches"*. All personality computing works face three main challenges, namely personality recognition (i.e. recognition of the personality assessed via self-ratings), personality perception (i.e. personality that others refer to a given individual), and personality synthesis (i.e. generation of personality in artificial entities). However, as stated by authors in [81], neither the self nor the other possesses all the necessary info for adequate personality ratings, and both can support each other for the prediction

of important life outcomes.

Personality recognition aims at the automatic recognition of individuals' personality annotated using self-assessment questionnaires. In this approach, accurate self-knowledge of personality, defined as how individuals are aware of their behavioral patterns, is the key assumption. To assess individuals' personality, questionnaires or tests such as the Big Five Inventory (BFI) [82] are traditionally administered. Personality questionnaires usually investigate how much agreement one would give to certain situations, where each situation is associated with a personality dimension. Self-assessed personality labeled datasets are more rare to find in the personality computing field, this is due to the fact that the experimenters not only have to recruit several participants, but they have to find participants willing to share their personality scores.

On the other hand, personality perception is based on the personality labels that other individuals attribute to the target individual. One of the main disadvantages of this approach is the externalization effect [81]. This effect highlights that some personality aspects are expressed more externally than others and, therefore, are likely to be more visible to third-party raters. For example, the Extraversion trait is one of the most interpersonal of the traits, and the way it is defined and measured often emerges in overt behaviors. To assess personality perception labels, researchers assign the personality questionnaire regarding a target individual to external raters. Usually, several voters evaluate the same individuals and agreement analysis is used to obtain the final personality scores.

Several works involving different technologies (i.e. text mining, video analysis, audio analysis) have been proposed for personality computing. However, in this thesis, we will mainly focus on methods that map video features to personality labels. As this thesis aims to extend the understanding of behavior's dynamics and intents, we believe that the study of human attitude and personality can help in the definition and the interpretation of more meaningful semantic behavioral patterns. Hence, in this section, an introduction of state-of-the-art methods that tackled the personality recognition challenge is given.

2.10.1. PERSONALITY RECOGNITION

As video data provides a rich set of information about human behaviors, several researches focused on the extraction and mapping of visual features to personality labels.

In controlled environments, audio features can also be added to visual cues for a multimodal analysis of personality. For example, in the popular challenge named "First Impressions" [83], several works utilizing video and audio features were proposed. The goal of the data was to automatically evaluate the personality of subjects for a job screening application. Video and audio data contain the participants in a portray format (one unique person as foreground at a fixed distance from the camera). Personality traits labels were established using the perceived personality strategy (how others perceive the personality of the target individual). Using this dataset, authors in [84] propose two separate networks to encode video and audio features. Finally, the authors employ a two-step late fusion for personality prediction.

As stated before, audio and face features are best exploited in controlled environments as they are sensitive to camera positions, as well as to privacy restrictions [85].

For example, face detection, and consequently, face analysis becomes very difficult using surveillance cameras due to their high position. To overcome this obstacle, we started to shift our attention towards other indicators of human personality such as body motion as well as body postures and gestures. Body gestures are one of the most important forms of nonverbal communication. They include movements of hands, head and other parts of the body that allow individuals to communicate a variety of feelings, thoughts and emotions [86].

Authors in [87] showed that the upper body motion of public speakers is a good predictor for personality. A set of body motion features such as high/low velocity, variation in motion directions, and movements' amplitudes are extracted from video data. Results showed several correlations between body motion activity and personality traits. Speakers exhibiting periods of low activity interrupted by periods of high activity tended to be perceived as highly agreeable. While speakers showing a high overall motion activity were found to have high values of extraversion.

Body pose features proved their efficiency in discovering the emergent leadership in small groups [88]. For examining social interactions in meetings (such as for emergent leader detection, leadership style prediction and classification of high/low extraversion), the authors utilize short-term motions extracted using optical flow. A hybrid CNN model which takes optical flow images as inputs, is adapted for feature extraction. The obtained final features are used for the detection of various social interactions using the Localized Multiple Kernel Learning (LMKL) classifier for the prediction of High/Low Extraversion labels.

Authors in [89] investigate personality displays during a human-human-robot interaction using multimodal sensors data. Video-based features, as well as audio and physiologic signals were used to study the different interaction settings. This study provides a novel point of view on the personality domain as it aims to predict subjects' personality by comparing two types of interactions: human-human interaction and human-robot interaction. Another work that examines multimodal data for personality recognition is the one described in [90]. In this study, individuals' and groups' affective responses were recorded and analysed while watching emotional videos. This study tackles important challenges in the personality computing field such as comparing the human affective behavior when alone or in groups.

Many researchers have tested the perception of affective states through dance, which allows to have a set of controlled expressive body movements [91]. Authors in [92] report the following cues as highly expressive dance movements: changes in tempo, directional changes in face and torso, frequency of arms up, duration of arms away from torso , muscle tension, and duration of time leaning forward. Authors in [93] suggest that the duration of the movement, the quantity of the movement (the amount of observed movement relative to the velocity and movement energy represented), and contraction index (measured as the amount of body contraction/expansion) play key roles in the perception of affect from dance movements.

Without considering the motion itself, solely the information of body postures and head positions can provide an indication of certain attitudes. Authors in [94] studied face-to-face conversations and free-standing conversational groups (FCGs) in a social environment. Several interesting findings were described. High extroversion trait was

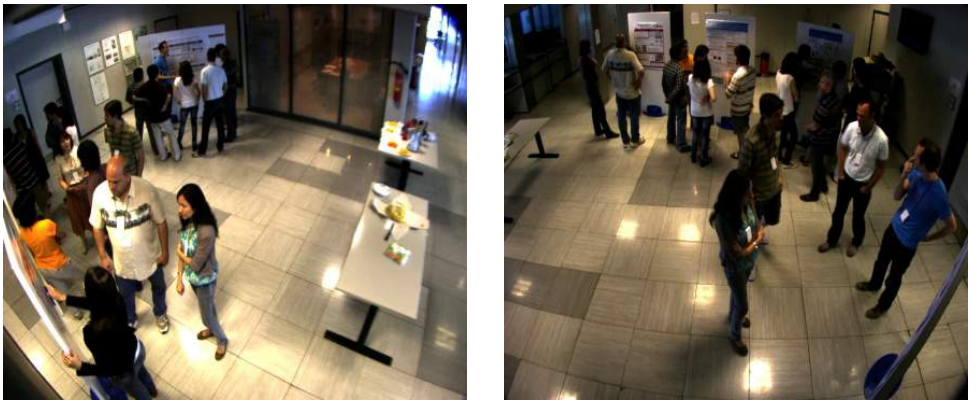


Figure 2.10: Data example showing two camera views included in the SALSA Dataset [94]

found to be highly correlated with body postures predisposed towards face to face conversations. In other words, more extraverted targets appear to adopt a more "open" body pose towards others and therefore, establishing more face-to-face interactions. Agreeable subjects were found to have a tendency towards engaging in face-to-face interactions with a higher number of people within highly connected clusters of people.

Following these studies, in Chapter 5, Chapter 6, and Chapter 7, we aim to study behavioral patterns and their relation with personality assessment scores in diverse real-world situations.

2.10.2. PERSONALITY DATASETS USED IN THIS DISSERTATION

In this thesis, we aim to study human behavioral patterns in unconstrained scenarios. Furthermore, one of the main contributions is to use insights from personality psychology to improve the computational understanding of human intents and behaviors. In this context, the dataset introduced by the authors in [94], called the SALSA dataset, is explored extensively in our experiments. The SALSA dataset is designed for multimodal and Synergetic sociAL Scene Analysis. It contains multimodal data from two social events (30 minutes each) in an unconstrained indoor scenario. Video data was recorded from 4 cameras placed at each corner of the room, and ground truth positions of the subjects' movements were provided every 45 frames (Fig. 2.10). It consists of two parts, the first part was recorded during a poster presentation session, and the second one was recorded during the coffee break, where all the participants were allowed to freely interact with each other (this part is named cocktail party). The two parts contain the same participants and their personality scores were collected using the Big-Five personality questionnaire [95].

The majority of existing methods investigate personality assessment in social contexts, such as crowded places or social events. However, they ignore the role of behaviors as well as personality in nonsocial situations (i.e. activities when individuals are alone). Therefore, in this thesis (Chapter 5), we introduce a novel benchmark dataset, called "Nonsocial dataset", for enhancing personality recognition in nonsocial scenarios. We

also highlight the need for interdisciplinary research, between psychology, computer vision and affective computing, for enhancing the ability to understand and automatically recognize human personality on novel and unconstrained datasets.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[3] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[4] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[5] G. Antipov, *Deep learning for semantic description of visual human traits*. PhD thesis, 2017.

[6] V. Lomonaco, *Continual learning with Deep Architecture*. PhD thesis, 2019.

[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[8] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and helmholtz free energy," in *Advances in neural information processing systems*, pp. 3–10, 1994.

[9] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[10] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the conference on empirical methods in natural language processing*, pp. 151–161, Association for Computational Linguistics, 2011.

[11] A. Abrams, J. Tucek, N. Jacobs, and R. Pless, "LOST: Longterm Observation of Scenes (with Tracks)," in *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 297–304, 2012.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, pp. 194—-197, 2012.

[15] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[16] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism in perspective*, vol. 19, pp. 143–155, 1989.

[17] Y.-T. Zhou and R. Chellappa, "Computation of optical flow using a neural network," in *IEEE International Conference on Neural Networks*, vol. 1998, pp. 71–78, 1988.

[18] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," *arXiv preprint arXiv:1510.07493*, 2015.

[19] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2285–2294, 2016.

[20] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative cnn video representation for event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1798–1807, 2015.

[21] Y. LeCun, "The power and limits of deep learning: In his iri medal address, yann lecun maps the development of machine learning techniques and suggests what the future may hold.," *Research-Technology Management*, vol. 61, no. 6, pp. 22–27, 2018.

[22] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: a survey," *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015.

[23] C. Jaschinski, *Independent Aging with the Help of Smart Technology: Investigating the Acceptance of Ambient Assisted Living Technologies*. PhD thesis, 2018.

[24] "Horizon 2020 eu project (ict4life)." http://www.ict4life.eu/.

[25] J. Habetha, "The myheart project-fighting cardiovascular diseases by prevention and early diagnosis," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6746–6749, IEEE, 2006.

[26] P. Urwyler, R. Stucki, L. Rampa, R. Müri, U. P. Mosimann, and T. Nef, "Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living," *Scientific reports*, vol. 7, p. 42084, 2017.

[27] F. Ordóñez, P. de Toledo, A. Sanchis, *et al.*, "Activity recognition using hybrid generative/discriminative models on home environments using binary sensors," *Sensors*, vol. 13, no. 5, pp. 5460–5477, 2013.

[28] K. K. Peetoom, M. A. Lexis, M. Joore, C. D. Dirksen, and L. P. De Witte, "Literature review on monitoring technologies and their outcomes in independently living elderly people," *Disability and Rehabilitation: Assistive Technology*, vol. 10, no. 4, pp. 271–294, 2015.

[29] T. Theodoridis, V. Solachidis, N. Vretos, and P. Daras, "Human fall detection from acceleration measurements using a recurrent neural network," in *Precision Medicine Powered by pHealth and Connected Health*, pp. 145–149, Springer, 2018.

[30] A. Vienne, R. P. Barrois, S. Buffat, D. Ricard, and P.-P. Vidal, "Inertial sensors to assess gait quality in patients with neurological disorders: a systematic review of technical and analytical challenges," *Frontiers in psychology*, vol. 8, p. 817, 2017.

[31] K. Malhi, S. C. Mukhopadhyay, J. Schnepper, M. Haefke, and H. Ewald, "A zigbee-based wearable physiological parameters monitoring system," *IEEE sensors journal*, vol. 12, no. 3, pp. 423–430, 2010.

[32] B. Kikhia, T. G. Stavropoulos, G. Meditskos, I. Kompatsiaris, J. Hallberg, S. Sävenstedt, and C. Melander, "Utilizing ambient and wearable sensors to monitor sleep and stress for people with bpsd in nursing homes," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 2, pp. 261–273, 2018.

[33] F. Alvarez, M. Popa, N. Vretos, A. Belmonte-Hernández, S. Asteriadis, V. Solachidis, T. Mariscal, D. Dotti, and P. Daras, "Multimodal monitoring of parkinson's and alzheimer's patients using the ict4life platform," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2017.

[34] H. Nait-Charif and S. J. McKenna, "Activity summarisation and fall detection in a supportive home environment," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 4, pp. 323–326, IEEE, 2004.

[35] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pp. 1147–1153, IEEE, 2011.

[36] F. Negin, A. Goel, A. G. Abubakr, F. Bremond, and G. Francesca, "Online detection of long-term daily living activities by weakly supervised recognition of sub-activities," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2018.

[37] E. Batista, F. Borras, F. Casino, and A. Solanas, "A study on the detection of wandering patterns in human trajectories," in *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–6, IEEE, 2015.

[38] Y. Zhang, G. Layher, S. Walter, V. Kessler, and H. Neumann, "Visual confusion recognition in movement patterns from walking path and motion energy," in *International Conference on Smart Homes and Health Telematics*, pp. 124–135, Springer, 2017.

[39] T. v. Kasteren, G. Englebienne, and B. Kröse, "Activity recognition using semi-markov models on real world smart home datasets," *J. Ambient Intell. Smart Environ.*, vol. 2, pp. 311–325, 2010.

[40] T. Nef, P. Urwyler, M. Büchler, I. Tarnanas, R. Stucki, D. Cazzoli, R. Müri, and U. Mosimann, "Evaluation of Three State-of-the-Art Classifiers for Recognition of Activities of Daily Living from Smart Home Ambient Data," *Sensors*, vol. 15, no. 5, pp. 11725–11740, 2015.

[41] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*, pp. 1297–1304, IEEE, 2011.

[42] Z.-A. Deng, Y. Hu, J. Yu, and Z. Na, " Extended Kalman filter for real time indoor localization by fusing WiFi and smartphone inertial sensors," *Micromachines*, vol. 6, pp. 523–543, 2015.

[43] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*, pp. 363–370, Springer, 2003.

[44] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1744–1757, 2011.

[45] H. Mousavi, M., A. Perina, R. Chellali, and V. Mur, "Analyzing tracklets for the detection of abnormal crowd behavior," in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV 2015)*, pp. 148–155, 2015.

[46] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*, pp. 428–441, Springer, 2006.

[47] N. Dalal and B. Triggs, "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, 2005.

[48] D. Xie, T. Shu, S. Todorovic, and S.-C. Zhu, "Modeling and inferring human intents and latent functional objects for trajectory prediction," *arXiv preprint arXiv:1606.07827*, 2016.

[49] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *European Conference on Computer Vision*, pp. 263–279, Springer, 2016.

[50] S. Huang, X. Li, Z. Zhang, Z. He, F. Wu, W. Liu, J. Tang, and Y. Zhuang, "Deep learning driven visual path prediction from a single image," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5892–5904, 2016.

[51] X. Feng, Z. Cen, J. Hu, and Y. Zhang, "Vehicle trajectory prediction using intention-based conditional variational autoencoder," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 3514–3519, IEEE, 2019.

[52] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–742, 2016.

[53] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," *Neurocomputing*, vol. 304, pp. 64–81, 2018.

[54] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.

[55] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3488–3496, 2015.

[56] H. Wang and C. O'Sullivan, "Globally continuous and non-markovian crowd activity analysis from videos," in *European conference on computer vision ECCV 2016*, pp. 527–544, Springer International Publishing, 2016.

[57] Y. Kwon, K. Kang, J. Jin, J. Moon, and J. Park, "Hierarchically linked infinite hidden markov model based trajectory analysis and semantic region retrieval in a trajectory dataset," *Expert Systems with Applications*, vol. 78, pp. 386–395, 2017.

[58] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011*, pp. 3153–3160, IEEE, 2011.

[59] J. See and S. Tan, "Lost World: Looking for Anomalous Tracks in Long-term Surveillance Videos," in *Proc. of the Image and Vision Computing New Zealand (IVCNZ)*, pp. 224–229, 2014.

[60] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[61] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2891–2900, 2017.

[62] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5308–5317, 2016.

[63] J. Tu, H. Liu, F. Meng, M. Liu, and R. Ding, "Spatial-temporal data augmentation based on lstm autoencoder network for skeleton-based human action recognition," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3478–3482, IEEE, 2018.

[64] N. Chen, J. Bayer, S. Urban, and P. Van Der Smagt, "Efficient movement representation by embedding dynamic movement primitives in deep autoencoders," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 434–440, IEEE, 2015.

[65] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 4570–4579, IEEE, 2017.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[67] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," *arXiv preprint arXiv:1711.05941*, 2017.

[68] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[70] A. Aristidou, D. Cohen-Or, J. K. Hodgins, Y. Chrysanthou, and A. Shamir, "Deep motifs and motion signatures," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, p. 187, 2019.

[71] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–971, 2016.

[72] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino, "Towards computational proxemics: Inferring social relations from interpersonal distances," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 290–297, IEEE, 2011.

[73] H. Xue, D. Q. Huynh, and M. Reynolds, "Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1186–1194, IEEE, 2018.

[74] S. Biswas and J. Gall, "Structural recurrent neural network (srnn) for group activity analysis," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1625–1632, IEEE, 2018.

[75] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[76] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pp. 37–42, ACM, 2010.

[77] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929–2936, IEEE, 2009.

[78] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, 2012.

[79] K. Loewenthal and C. A. Lewis, *An introduction to psychological tests and scales*. Psychology press, 2018.

[80] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[81] A. G. Wright, "Current directions in personality science and the potential for advances through computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 292–296, 2014.

[82] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[83] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *European Conf. on Computer Vision*, pp. 400–418, Springer, 2016.

[84] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2018.

[85] K. W. Bowyer, "Face recognition technology: security versus privacy," *IEEE Technology and society magazine*, vol. 23, no. 1, pp. 9–19, 2004.

[86] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, pp. 1–1, 2018.

[87] M. Koppensteiner, "Motion cues that make an impression: Predicting perceived personality by minimal motion information," *Journal of experimental social psychology*, vol. 49, no. 6, pp. 1137–1143, 2013.

[88] C. Beyan, M. Shahid, and V. Murino, "Investigation of small group social interactions using deep visual activity-based nonverbal features," in *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 311–319, ACM, 2018.

[89] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, 2017.

[90] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018.

[91] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 341–359, 2013.

[92] R. T. Boone and J. G. Cunningham, "Children's decoding of emotion in expressive body movement: The development of cue attunement.," *Developmental psychology*, vol. 34, no. 5, p. 1007, 1998.

[93] A. Camurri, I. Lagerlöf, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 213–225, 2003.

[94] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1707–1720, 2016.

[95] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

3

MOTION PATTERN DISCOVERY AND PATH PREDICTION

This chapter is based on the following publications:

- D. Dotti, M. Popa, and S. Asteriadis, "Unsupervised discovery of normal and abnormal activity patterns in indoor and outdoor environments", in *VISIGRAPP (5:VISAPP)*, pp. 210–217, 2017.
- F. Alvarez, M. Popa, V. Solachidis, G. Hernandez-Penaloza, A. Belmonte-Hernandez,S. Asteriadis, N. Vretos, M. Quintana, T. Theodoridis, D. Dotti,et al., "Behavior analysis through multimodal sensing for care of parkinson's and alzheimer's patients", *IEEE Multimedia*, vol. 25, no. 1, pp. 14–25, 2018.

3.1. INTRODUCTION

Automatic monitoring and interpretation of daily motion patterns has gained popularity over the last decade, having applications in ambient-assisted living (AAL), surveillance, and shopping behavior understanding [1, 2]. One of the goals in human behavior understanding consists in detecting deviations from normal behaviors. Once an object's regular activity patterns are learnt, different types of deviations which could be considered abnormal can be detected. This analysis is useful for behavior understanding in varying environments, such as private houses, offices, or public spaces [3]. One of

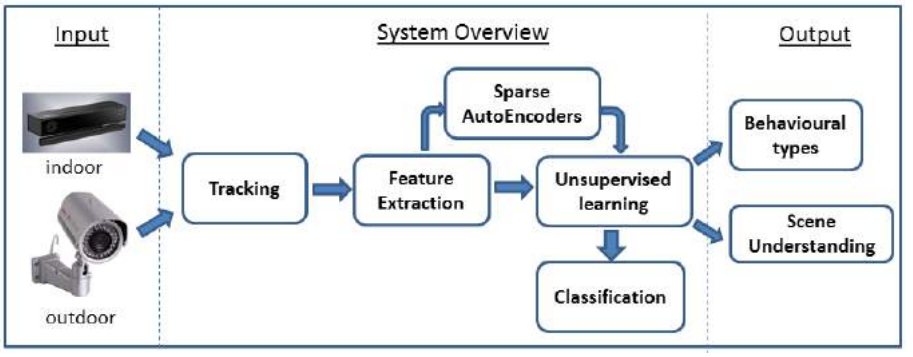


Figure 3.1: Overview of the proposed system.

the major challenges in behavior understanding is that objects' motion and behaviors should be interpreted depending on the context. For example, a pedestrian standing static on the road could be interpreted as a dangerous event, however, if that road hosts a street parade and it is closed to vehicles, then it should be considered as a normal event. As can be noted, behavior understanding presents different goals depending on the scenarios it is applied to, for example, if the data contains behaviors from public spaces (city squares, train stations) the goal could be to detect potentially dangerous situations, such as violence, crashes or aggression [4]. On the other hand, if the data contains behaviors from private environments (homes, offices), the goal could be to detect alterations of the physical or emotional state of individuals for improving their well-being [5].

In this chapter, we propose an adaptive monitoring system, able to work in both indoor and outdoor environments based on two different sensors: the depth sensor Microsoft Kinect v2 and standard surveillance cameras. In an office scenario, we aim to learn repeated patterns of activities, and detect non-expected behaviors [1] (abnormalities). In the outdoor scenario we use the public dataset introduced in [6], where videos are taken from streaming webcams in different public places capturing the same half an hour every day for over a year.

Our approach aims to provide an analysis of the monitored environment by extracting spatio-temporal information on motion trajectories. Spatial information describes the regions which are frequently occupied. Additionally, motion information extracted from these regions contributes to obtaining high level information such as stationary behaviors (sitting, working at the desk) as well as active behaviors (walking, exiting the space) for the indoor scenario. For the outdoor case, motion information is useful at distinguishing between several moving objects (e.g auto-vehicles or pedestrians), as well as for identifying usual spatial-motion patterns for each of the objects (e.g. pedestrians crossing the street in a designated area or not, cars moving on the street and parking in a parking lot). Furthermore, we obtain an improved and efficient feature representation, by applying a sparse autoencoder algorithm on top of trajectory features, which we prove to be useful for representing the expected and unexpected behavioral patterns in

[1] In the remaining of this chapter, we use the term *behavior* to denote a set of activities over a short time interval.

both indoor and outdoor scenarios.

Manually providing annotation to define what is normal and what is abnormal on surveillance data is difficult and time consuming. The data is often unbalanced with 99% of it containing normal events, and only 1% containing meaningful ones. In this chapter, we investigate an unsupervised approach for obtaining data annotations, by performing clustering on the extracted features. Our result is a map of the environment organized in spatio-temporal motion clusters. With this map, we aim to simplify the labeling process as the final users of the system will have to label k activity patterns instead of all the individual trajectories. Furthermore, this process can be very useful when the system needs to be deployed in different environments and the labelling task has to be fast and generalized.

In this chapter, we propose a model trained to distinguish between normal vs. abnormal behaviors using motion information. We compute multiple motion descriptors, which, along with sparse autoencoders, can lead to optimized results in the analysis of the behaviors (Fig. 3.1). As ground-truth labels about normal vs. abnormal classes are difficult to find, for our specific purpose, we facilitate the integration of expert opinion using clustering. Lastly, we propose a system that can learn an environment from scratch and, thus, can be easily deployed in new, unknown settings, both indoors and outdoors.

3.2. FEATURE EXTRACTION

Following the flow of activities presented in Fig. 3.1, we first obtain trajectory data from the tracking algorithm, then, we feed the trajectories to the feature extraction module. The first step in our feature analysis is to split the scene into 2D regions $r_1,..,r_R$, where each region corresponds to a part of the scene (see details in Section 3.4.1). For every region, we extract different types of descriptors which are subsequently used for normal vs. abnormal behavior recognition.

3.2.1. OCCUPANCY HISTOGRAM (OH)

In an indoor environment, often activities are correlated with regions where they are performed. For instance, we usually work sitting at our desk, whereas meetings are organised in the meeting area. In this section, we compute the level of occupancy in each image region and use it as a descriptor for behavior understanding. As a first step, similarly to the analysis described in [7], we count the trajectory points in each non-overlapping spatial patch to form a region based occupancy histogram. Specifically, in a given time interval $t_1,..,t_T$, we detect and count the trajectory points (x,y) that fall in the spatial regions $(r_i), i \in 1,..R$.

3.2.2. ADAPTED HISTOGRAM OF ORIENTED TRACKLETS (AHOT)

Trajectory information heavily relies on detection and tracking algorithms, however, these algorithms may fail in different circumstances such as occlusions, crowded situations, poor lighting conditions and so on. Hence, authors in [3, 8] proposed to analyse shorter motion descriptors called tracklets. A tracklet indicates the movement of a subject, frame by frame, for a short period of time, and it represents only a fragment of the global trajectory, as the tracking information might be terminated due to ambiguities in the scene.

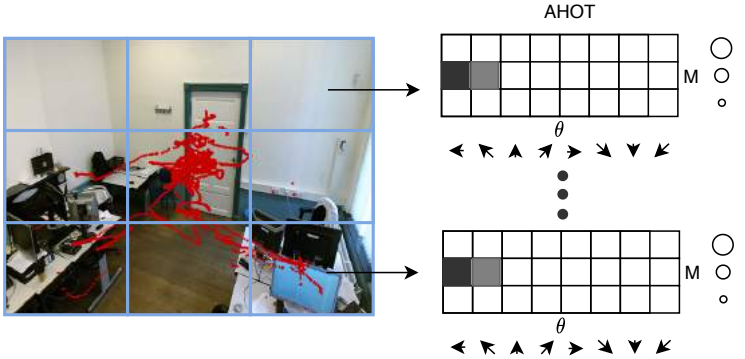


Figure 3.2: AHOT descriptor. Orientation and motion descriptors are extracted from every spatial sector of the scene. The information is quantized in histograms and then concatenated together.

In this chapter, inspired by the Histogram of Oriented Tracklets (HOT) feature extraction algorithm introduced by [3], we propose an adaptation of the HOT features for extracting statistical information of objects' motion in the scene. Differently from the HOT features proposed by [3], where the histogram representation considers only the maximum motion magnitude among all the tracklets inside a spatio-temporal block (see details in Chapter 2 Section 2.8.1), we consider the motion characteristics of every tracklet. As HOT features were proposed for modeling crowd behavior, the authors focused on the maximum motion value in each tracklet. On the other hand, in our approach, we aim at capturing individual motion patterns. Moreover, as we consider indoor scenarios, where the movements are more controlled due to the limited space, we need more fine-grained motion information. The calculation of the descriptor is visualized in Figure 3.2. The scene is spatially divided into non-overlapping regions $r \in r_1,..r_R$. Given a tracklet I, composed by a sequence of 2D coordinates (x_t^I, y_t^I), (x_{t+1}^I, y_{t+1}^I), we compute the magnitude M and the orientation Θ scores as follows:

$$\Theta^{I,r} = arctan\frac{(y_{t+1}^{I,r} - y_t^{I,r})}{(x_{t+1}^{I,r} - x_t^{I,r})} \tag{3.1}$$

$$M^{I,r} = \sqrt{(x_{t+1}^{I,r} - x_t^{I,r})^2 + (y_{t+1}^{I,r} - y_t^{I,r})^2} \tag{3.2}$$

where (I,r) represents the portion of tracklet I that intersects with the spatial block r. Following the original feature structure proposed by [3], $\Theta^{I,r}$ and $M^{I,r}$ are quantized independently into histograms with size $OR = 8$ and $MA = 3$ bins. The bins of an histogram representing the tracklet I is formed by counting the number of $\Theta^{I,r}, M^{I,r}$ combinations in every region r. Finally, every tracklet I is represented by a matrix of size $AHOT^I = R \times OR \times MA$ which encodes its spatio-temporal development.

3.2.3. MOTION DESCRIPTOR SPEED AND CAHOT

To enable a better understanding of the types of behaviors displayed in an outdoor environment, we need as an initial step, to distinguish between the moving objects present

in the scene. This analysis is useful for detecting abnormal behaviors which are different across the various types of involved objects such as pedestrians and vehicles. One intuitive feature that can help in this process is a descriptor that encodes the raw speed of an object. In particular, given a tracklet I that intersect with the spatial block r for a sequence of frames $t_0,..t_T$, we calculate the velocity value as $vel^{(I,r)} = \frac{(x_{t_T}^{I,r}, y_{t_T}^{I,r}) - (x_{t_0}^{I,r}, y_{t_0}^{I,r})}{T}$, the acceleration value as $acc^{(I,r)} = \frac{vel_{t_T}^{I,r} - vel_{t_0}^{I,r}}{T}$, and the curvature value as $k^{(I,r)} = \frac{1}{rad}$ where rad is the radius between the entry point $x_{t_0}^{(I,r)}, y_{t_0}^{(I,r)}$ and the exit point $x_{t_T}^{(I,r)}, y_{t_T}^{(I,r)}$ of the tracklet I in/from the cuboid r. Finally, for every tracklet I in every cuboid r, we concatenate these three values obtaining a final matrix of size $SPEED_I = R \times 3$.

Lastly, as the SPEED descriptor is formed by raw values, namely vel, acc, and k, we form another descriptor named $CAHOT$, by quantizing the raw values of SPEED. Specifically, 8 bins were found for the curvature values, 3 bins were found for the velocity values, and 3 bins were found for the acceleration values, forming a descriptor of size $CAHOT = 8 \times 3 \times 3$.

3.2.4. SPARSE AUTOENCODERS (SAE)

In this chapter, every tracklet is transformed into mid-level descriptors that aim to describe its spatio-temporal dynamics. However, the spatial as well as temporal information inside the descriptors is not equally distributed. For example, there exist spatial blocks that contain more data than others, or blocks that do not have any data at all (e.g. blocks on the top part of the image). Therefore, our idea is to obtain a more compact and meaningful descriptor using Sparse Autoencoders [9]. An autoencoder is a technique which aims to minimize the reconstruction error between the input and the output in an unsupervised way (see Chapter 2 Section 2.3 for the theoretical background). It is useful at estimating the underlying data distribution, and by placing constraints on the network like sparsity [10], the algorithm can learn meaningful structures in the data.

In particular, given a tracklet I transformed into any of the descriptors (i.e. $AHOT, SPEED, CAHOT$), denoted by X for simplicity, we aim at finding a more compact representation to use for the final abnormality detection task. We use the encoder and decoder structure explained in Chapter 2, Eq. 2.8 and Eq. 2.9 with a SAE loss function L that can be described as follows:

$$L_{sparse}(X) = L(X, \hat{X}) + \beta \sum_{j=1}^{J} KL(\rho, \rho') \tag{3.3}$$

where $L(X, \hat{X})$ represents the loss function with the goal of minimizing the error between the input data X and the reconstructed data $\hat{X}$. J is the number of neurons in the hidden layer, and the index j is summing over the hidden units in our network. $KL(\rho, \rho')$ indicates a measure of the difference between two probability distributions, and it is used to add the sparsity constraint to the AE. The sparsity constraint is used to force most of the hidden units to be close to 0, reconstructing the input using as few features as possible. In this case, the penalty will be applied on ρ' when it will deviate too much from ρ. Finally, parameter β controls the weight of the sparsity penalty.

3.2.5. UNSUPERVISED LEARNING

The goal of this study is to develop a system useful for detecting normal and abnormal behavior patterns in unknown environments, in an unsupervised manner. As we do not have labels for normal/abnormal events in the data, we aim at using an unsupervised approach to obtain the labels without human intervention. This is particularly important for surveillance data where the process of detecting abnormal events is very tedious, as usually the abnormal events are much more rare than the normal ones. In particular, we perform clustering analysis on our spatio-motion descriptors X to obtain a clear separation between different behavioral patterns.

Moreover, as the analyzed datasets contain several moving objects (i.e. pedestrians, vehicles, etc.), we can use the clustering analysis to detect the different moving objects in an unsupervised way. We assume that different types of moving objects have to obey at different regulations, therefore, the definition of the object is critical for the detection of possible abnormalities.

Finally, the labels obtained in this module are used for training and testing the next system's component in a supervised way using the Logistic Regression.

3.3. EXPERIMENTS

The main goal of this chapter is to test our model on the abnormality detection task on two surveillance datasets (introduced in Chapter 2 Section 2.8.2). The Long-term Observation of Scenes (with Tracks) or LOST dataset [6] contains trajectory data recorded in outdoor public spaces such as city squares or road intersections. This dataset provides trajectory coordinates as well as bounding box information of the detected objects. However, the objects' labels are missing.

Consequently, for the LOST dataset, we first apply a clustering technique on the computed features to differentiate between pedestrians and auto-vehicles data in an unsupervised way (Section 3.3.1). This step gives us the objects' labels for the definition and detection of normal/abnormal events (3.3.2).

Furthermore, in this chapter, we present KIMOFF (kinect-monitoring-office), a dataset created by monitoring an office environment, during working hours, for twenty-four days. KIMOFF contains trajectory coordinates belonging to the head joint of the tracked workers.

3.3.1. PEDESTRIANS VS. AUTO-VEHICLES LABELING IN THE LOST DATASET

Defining an abnormal behavior model on the LOST dataset can be very challenging given the big changes of the analysed environments during long-term recording. Imagine a public square that in different times of the year hosts different events (food festival, Christmas market etc) and therefore, special regulations are defined. Moreover, the variety of moving objects also has to obey different regulations. For example, cars are not allowed to drive in the center of the square, whereas bikes and pedestrians can. Our first task is to separate the trajectories belonging to pedestrians from the trajectories belonging to vehicles using the feature descriptors introduced in Section 3.2.

As there do not exist ground-truth annotations that indicate the types of object in the scene, we aim at estimating them using the bounding boxes information provided in the

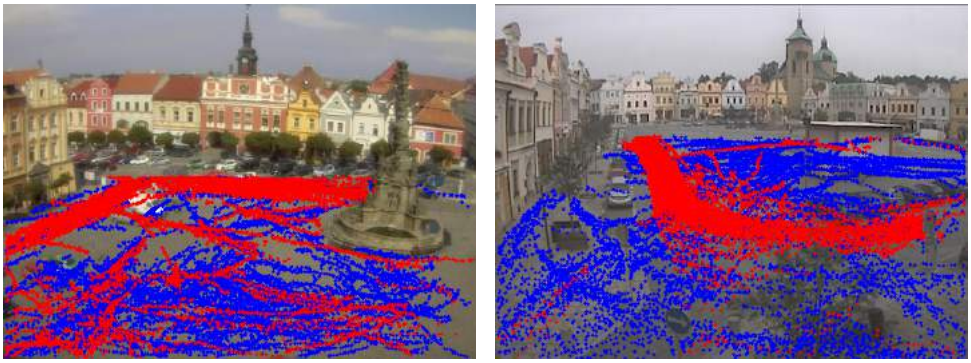
dataset. In particular, we compute the aspect ratio of each bounding box to assign the label "vehicle" if the longer side is horizontal and we assign the label "pedestrian" if the longer side is vertical.

Table 3.1 shows the prediction accuracy between pedestrians and vehicles using different descriptors and the Logistic Regression. In "Camera001", the SPEED descriptor obtains the best result, as it embeds information which is more robust to the cluttered scenario. Orientation information embedded in the AHOT and CAHOT descriptors becomes less crucial when the vehicles are allowed to go almost everywhere, during the events. On the other hand, in "Camera017" CAHOT and AHOT descriptors perform better than SPEED because vehicles follow the same path, information which is captured by the orientation and curvature features. Next, because abnormality has different meanings for each of the classes, we will treat them separately as input to the abnormal behavior detection module.

Descriptors	Camera001	Camera017
SPEED	**83.5%**	85.5%
AHOT	82.7%	87.6%
CAHOT	82.8%	**87.7%**

Table 3.1: Pedestrian vs. Auto-vehicles prediction accuracy

Fig. 3.3 depicts the separation between vehicles and pedestrians in the two analyzed scenarios. In Fig. 3.3a the spatial separation between the two classes is less clear than in 3.3b due to the many events that take place in the square. In fact, during these events, trucks are allowed to enter the square for commercial or construction purposes.



(a) "Camera001" (b) "Camera017"

Figure 3.3: Two scenes belonging to the LOST dataset. Color red indicates the vehicles trajectories, color blue indicates pedestrian trajectories.

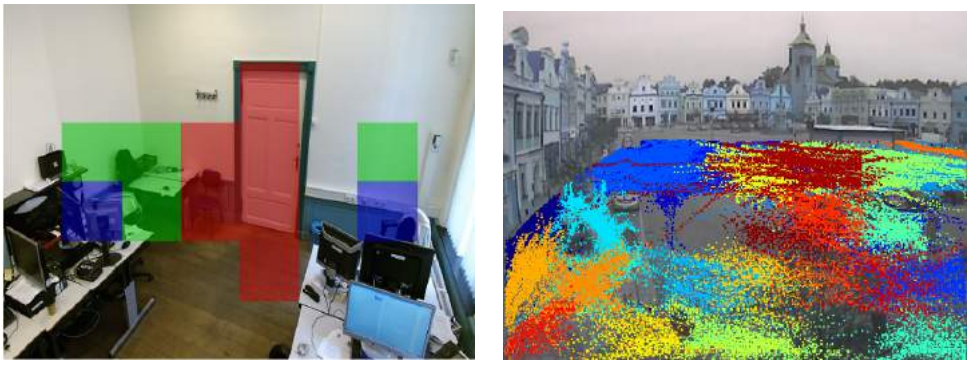
3.3.2. LABELING OF NORMAL AND ABNORMAL EVENTS

For each dataset, for each object's type, we apply mean-shift clustering on every descriptor. The best number of clusters is chosen applying user knowledge, as the clusters have to reflect the human interpretation of the scene. This is a key point in our system, as instead of manually labeling each trajectory sample, we allow the users of the system to validate the clustering results as well as defining what is normal and what is abnormal for the considered scenario.

In Fig. 3.4 we present an example of the obtained clustering result on both datasets, using the AHOT descriptor. The different colors belong to different activity patterns. For the KIMOFF dataset depicted in Figure 3.4a, for visualization purposes, we only show the 3 most populated clusters. The red color indicates the regions of the scene where big movements are found (e.g. corridor area and the door), which are transition areas. Green indicates the regions of the scene where light movements are detected, including areas close to the desks, where activities such as standing up, sitting down, or stretching are observed. Finally, blue indicates the regions where no-movement is detected, being restricted to the regions close to the computers, where usually people do not move too much because they are focused.

Figure 3.4b shows the clustering result of the pedestrian trajectories in the outdoor scenario. Pedestrians can be observed in all regions of the scene. However, they should follow the road regulations performing actions like crossing the street only in the permitted areas. Following this regulation, we chose the clusters that contained deviations from the permitted behavior and labeled them as abnormal behaviors.

The results obtained are satisfactory, as they can be interpreted in a meaningful way, highlighting that both spatial regions and motion information are important to define activity patterns. Given the semantic interpretation of these clusters, we define a binary classification problem (normal vs abnormal), selecting clusters that present unusual motion patterns as abnormal. A binary Logistic Regression is used for the training and testing of the model.



(a) Indoor scenario

(b) Outdoor scenario

Figure 3.4: Unsupervised semantic interpretation of the scene.

3.4. EXPERIMENTAL RESULTS

3.4.1. ABNORMAL BEHAVIOR PREDICTION PERFORMANCE

In this section we present the analysis of the performed experiments, for detecting normal vs. abnormal activity patterns, using the features described in Section 3.2. An important parameter in our analysis is represented by the spatial division of the scene in R regions. As the indoor dataset (KIMOFF) was recorded using Microsoft Kinect V2, the scene was divided using three dimensions (i.e. height, width, depth). The best division was $R = (8 \times 6 \times 2)$. In the outdoor scenario, the scene was divided using two dimensions (height and width), and the best division was $R = (8 \times 6)$. In the binary classification experiment, 80% of the data is used for training, and 20% is used for testing in a 5 fold cross validation setting

Table 3.2 displays the results obtained for the proposed feature descriptors using Logistic Regression. As expected, the motion related descriptors obtain higher results in the outdoor scenario than in the indoor one, and vice versa, density based Histogram (OH) obtains the highest result in the indoor scenario. The best result in both scenarios is obtained by applying the Sparse Autoencoder algorithm (SAE) on top of the adapted histogram of oriented Tracklets (AHOT). The feature representation obtained using the learned hidden layer parameters, introduced in Section 3.2.4, is beneficial as it helps at increasing the accuracy of the classification method in relation to the raw features. In fact, in Table 3.2 we highlight that the augmented features obtained by applying the SAE algorithm, reach higher accuracy than raw features in all the cases. Moreover, once trained, the autoencoder algorithm is useful at compressing the feature vectors, by estimating the underlying feature distribution and decreasing the processing time in the case of real-time applications. The best results are obtained for the SAE algorithm, using $J = 100$ hidden units, hence drastically decreasing the size of the AHOT and CAHOT raw descriptors. The number of hidden units was found experimentally, using 10-fold cross validation.

Descriptors	KIMOFF Dataset	LOST Dataset
SAE(AHOT)	**98.4%**	**98.7%**
AHOT	96.5%	97.5%
SAE(CAHOT)	86.1%	98.3%
CAHOT	85.2%	94.2%
SPEED	80.1%	97%
OH	97.4%	–

Table 3.2: Abnormal behavior prediction F1 accuracy on the considered datasets.

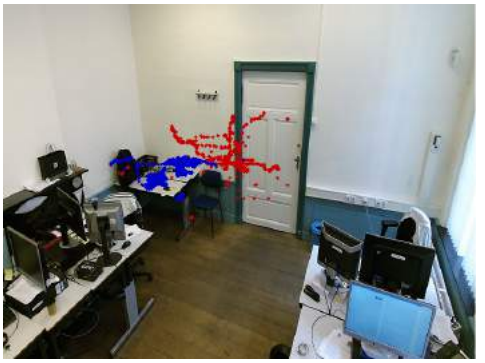
3.4.2. QUALITATIVE RESULTS

Examples of discovered normal and abnormal patterns are shown in Fig. 3.5, normal behavior patterns are defined by trajectories colored in blue, whereas abnormal behaviors are colored in red. Fig. 3.5a depicts the most common behavior pattern in an office, as we expect that most of the time people are in front of the computer, creating big clouds

of tracking points in the desk regions. For the outdoor scenario, Fig. 3.5b shows tracks of pedestrians walking on the appropriate location: sidewalk. On the other hand, in Fig. 3.5c one possible abnormal behavior in an office is shown; a person is standing up (red trajectory clouds) being close to the worker sitting at the desk (blue points), which might indicate an interaction pattern for a long period. In Fig.3.5d pedestrians are crossing the road in dangerous areas where zebra crossing signs are not present, therefore we defined these actions as abnormalities.

(a) Working at the desk

(b) Walking on the pedestrian sidewalk



(c) Converging in the middle of the room

(d) Crossing the road in a dangerous area

Figure 3.5: Examples of normal and abnormal behaviors from the two analysed datasets.

3.5. ABNORMAL BEHAVIOR IN HEALTHCARE: THE ICT4LIFE PLATFORM

3.5.1. MOTIVATION

Automatic monitoring systems have been shown to be beneficial also in the healthcare domain helping to increase the security of the aging population living independently. Consequently, several research projects have been proposed to target different goals such as detection of abnormalities or automatized robotic assistance (see Section 2 for more details). In this context, the system proposed in the previous sections was adapted for monitoring aging individuals in indoor environments. Elderly living independently remain most of their time alone while medical professionals usually assist them only for a few hours a week. As a consequence, sensor-based technologies and systems which quantify Activities of Daily Living (ADL) can add new dimensions to existing clinical assessments [11, 12].

However, sensor-based applications working in private indoor environments present several challenges. Firstly, privacy in environments like private houses has to be well preserved. In this sense, monitoring frameworks must avoid recording sensitive data and they must be as unobtrusive as possible to not alter people's life. Secondly, monitoring systems should cover the most important areas of the house where abnormalities are more likely to occur. Given the privacy constraints explained before, this process requires the collaboration of several professionals from different fields for adequate installation plans. Thirdly, multimodal sensors covering a diversity of functionalities were shown to yield to a more general and complete picture of human behaviors. However, fusing different types of information is never a trivial task. Lastly, as the aging population suffers from heterogeneous comorbidities, different medical professionals should collaborate for the definition of specific objectives that automatic systems should cover.

In this framework, my PhD thesis supported the development of the ICT4Life European project [2], bringing a substantial contribution to the development of indoor monitoring systems. The project was composed of a consortium that included universities, companies, as well as hospitals to provide innovative ICT smart services for people affected by Alzheimer's and Parkinson's diseases. In the next sections, we will explain in detail the platforms and the provided algorithms for abnormal behavior detection.

3.5.2. INTRODUCTION

With an increasingly growing population in Europe, cognitive impairments is a major social and health issue. According to the World Alzheimer Report released by Alzheimer's Disease International (ADI) [13], dementia, including Alzheimer's disease, remains one of the biggest global public health challenges our generation is facing. In 2019, ADI estimated that there are over 50 million people living with dementia globally, a figure set to increase to 152 million by 2050. The old people in Europe want to live at their home: only 3,3% of the population older than 65 years live in an institutional center and also in Europe, 50% of people older than 80 years old, live alone, and 35% live as a couple [13]. However, aging brings several difficulties that are difficult to manage alone, demanding additional care from the elder's family and the healthcare professionals.

[2] https://cordis.europa.eu/project/id/690090

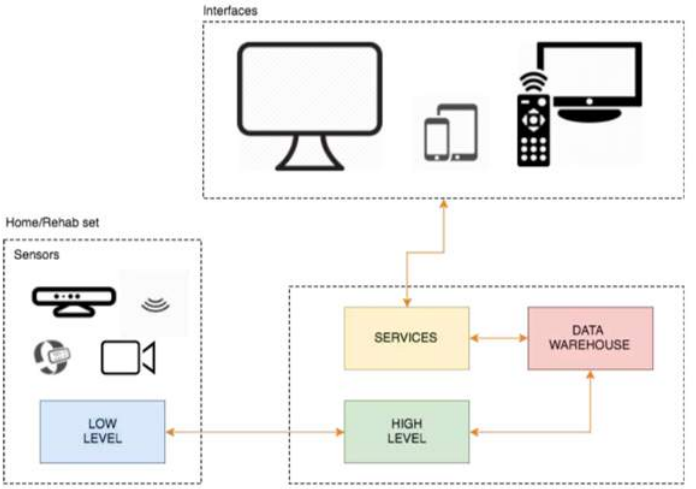


Figure 3.6: ICT4Life architecture. The Sensors module is responsible for extracting and analysing data extracted from sensors deployed in the environments. Services will utilize the information generated through the ICT4Life algorithms to generate personalised information for supporting senior citizens. Interfaces allow the interaction between the back end and front end of the systems, targeting senior citizens, caregivers, as well as doctors.

From the caregiver side, relatives are the other victim of the disease. As the impairment increases, the families need to dedicate much more time, as well as mental and physical effort. In many cases, the carer is the patient's partner who also is an old person with associated health problems. The patients' children are also very implicated in the care of their parents. This poses a challenge for healthcare organizations to find solutions that ensure the future financial sustainability of healthcare systems.

In this scenario, the ICT4Life European project aims to implement a platform integrating a series of innovative services, targeting aging people with cognitive impairments. A series of sensors such as depth sensor cameras, wearable sensors, and ambient sensors were exploited with the goal of providing proactive and patient-centered care to formal as well as informal caregivers. One of the main objectives of this platform is to provide advanced monitoring of patients using multisensor-based analytics and its integration with biomedical devices. Patients' activity patterns are recognized using two parallel channels: 1) Indoor daily activity analysis employing passive sensors such as depth cameras and ambient sensors, and 2), Health status estimation through the use of wearable sensors.

Figure 3.6 shows the architecture of the proposed platform. ICT4Life is formed by a set of independent modules connected between them in a logical manner, allowing easy information exchange, processing and reasoning. The sensors, as well as the interfaces are the inputs to a central platform. Sensors are in charge of acquiring the data related to the patient's movements, medical information and interactions with the environment. The interfaces allow the interactive communication between the end-users and the core services provided by the platform. High-level reasoning and inferences regarding the pa-

tient's health condition and daily activities represent the final product of ICT4Life, obtained by advanced analysis and multimodal fusion of the sensory data and the patient's medical data. A typical scenario in which the ICT4Life platform could be applied is the home monitoring scenario. Sensors, mounted in specific areas of the house (e.g. living room, bedroom, kitchen), monitor the activities of daily living. The low level module analyzes the data stream to detect clear and urgent deviations from the normal activity's patterns (i.e. abnormalities), for example, falling down or agitations states. Every week, the data is summarized and passed to the high level module. The high-level module focuses on fusing the information coming from the sensors as well as from the medical files to provide high level insights on the patient's health condition. For example, if the medical files indicate that patient x had some medications changed, and the low level module provides an increase of abnormalities occurrences, the high level module is responsible to investigate the connection of these two events.

During this PhD, my main role in relation to the ICT4Life platform was to develop intelligent algorithms for abnormal behavior detection. Therefore, only this module will be described in detail in the following paragraphs.

3.5.3. ABNORMAL BEHAVIOR DETECTION IN HEALTHCARE

This module focuses on analyzing sensing information to identify meaningful behaviors and inform any interested party (professionals, formal/informal caregivers) about the patient's situation. Specifically, one of the main goals is to provide the ICT4Life platform with the capability of assessing *real deviations* from the expected daily conduct of target users. For example, if the target senior shows a decrease in daily motion activity observed over a long period, the platform is able to detect an abnormal pattern, in this case, apathy.

In Figure 3.7 we show the data flow (from left to right) of the low-level subsystem. Firstly, passive sensors such as ambient sensors and cameras are installed to monitor the selected indoor environment (private home or rehabilitation center). Secondly, the data is analyzed using two different levels of processing. The first level is performed in real-time and it involves the detection of life threatening abnormal events, like falling down and agitation. In these cases, the algorithms are set to immediately send a warning notification to the patient's caregivers. The second level of analysis is performed to compute daily behavior analytics, such as overall daily motion or average blood pressure. These analytics are accumulated for a certain period of time (decided by the medical professionals) and summarization/visualization techniques are executed to be used as support evidence in the patient's clinical picture.

In recent years, due to new cutting-edge technologies, the range of equipment and services available to help elderly people safely stay in their home has substantially grown. The tools that have been chosen in the ICT4Life platform are a series of sensorial means that, in an unobtrusive manner, can track the old person's activities and behaviors. In this direction, depth sensors like the Kinect camera are used for the detection and tracking of human movements. By using the skeleton detection and tracking functionalities embedded in the sensors, fine-grained information on human activities can be exploited in a low-cost way. Furthermore, the use of depth images yields other benefits. As shown by authors in [14], depth sensors are robust to illumination changes. As illumination

Figure 3.7: Data flow in the ICT4Life platform. Sensors deployed in the environment are responsible for extracting behavioral data. Behavior analysis is performed with two levels of frequency. Life-threatening events like falling down and confusion states are analysed and detected in real-time. Behavior analytics over a longer period of time are summarized and shown to the professionals as support evidence in the patient's clinical picture.

in private homes is not always constant, having a detection and tracking algorithm that handles special cases such as when light is turned off is very important. Another important benefit consists of ensuring the privacy protection. The 3D depth information does not allow the recognition of the individuals' identity, fully preserving the privacy of the inhabitants. An example of the depth image and skeleton tracking information is shown in Figure 3.7 (left).

Indoor tracking using video data is challenging due to the inherent structure of those environments, which often include small separated rooms. Consequently, due to the presence of walls and furniture, the cameras coverage range cannot be maximally exploited. One solution would be to install a camera for every room, however, this will result in a very intrusive monitoring system. Hence, in the ICT4Life framework, binary ambient sensors were included due to their advantages such as cheap cost and low intrusiveness. Specifically, magnetic sensors were deployed in the environment for the detection of opening/closing events of objects like doors and cabinets. Magnetic sensors are composed of two magnets. When the two magnets separate, an internal reed switch inside the sensor will activate sending a binary event signal. For example, if the sensor is placed on a door, one magnet would be placed on the external frame, and the other magnet would be placed on the actual door. Opening the door causes the magnets to separate, and an event would be triggered, while closing the door would also trigger an event as the two magnets are reunited. Although the information provided by the sensor is limited, i.e. 1 indicates the opening event, and 0 indicates the closing event, by combining the binary events with time-stamps, we can obtain the inhabitants' movement patterns in areas where cameras are too intrusive such as bedrooms and bathrooms.

To the best of our knowledge, datasets containing abnormal behaviors such as confusions and agitations are rare to find in the computer vision field. Therefore, a novel dataset was proposed within the frame of my PhD. Specifically, in the next paragraphs, we will describe a novel dataset designed, developed and used to investigate the detection and the understanding of abnormal behaviors related to Alzheimer's disease.

3.5.4. MULTIMODAL DATASET FOR ABNORMAL BEHAVIOR DETECTION

Activities of Daily Living (ADLs) refer to a set of daily self-care activities, which people perform habitually and universally. ADL performance decline can often be a sign of mild cognitive impairment and early dementia, and, its assessment as well as its detection is critical to provide help promptly. Therefore, in the last decades, many datasets have been recorded for the purposes of generating automated solutions for Activity Recognition (AR) [15]. These datasets usually contain each ADL in a separate video, in a controlled environment (i.e. only one person in the video, and the camera placed in a frontal position).

Over the course of my PhD research, we designed, developed, analyzed and proposed a dataset which encourages the analysis of human behaviors in unrestricted settings for the discovery of generic patterns from spontaneous activities. One of the contributions of this dataset is that it contains elicited behaviors and not acted ones. In this dataset, we asked the participants to perform predefined tasks without instructing them on how these tasks should be performed. In this way, spontaneous behaviors were recorded. Additionally, datasets containing abnormal agitation and confusion behaviors from video data are rare to find as well as difficult to collect. Hence, with the help of medical professionals, we designed certain tasks which might provoke trajectory data similar to the ones created by people in confusion states. Doing so, we aimed to train our abnormal detection models on this dataset, and, by applying a transfer learning procedure, test them directly in hospitals with real patients.

The experimental design was inspired by both ADL datasets, as well as problem-solving based psychological tests. To create an unconstrained environment, no time limit nor know-how was given to complete the proposed six tasks. The experimental room was furnished with tables, chairs, a tea corner with a water kettle, and two office cabinets, having each drawer filled with many different objects. All around the room, boxes, and cases also containing objects, were spread to challenge our subjects for the completion of the experimental tasks (Figure 3.8).

We recorded the dataset using two sensors: Kinect V2 and binary magnet sensors. The Kinect sensor was placed on a closet in the corner of the experiment room, while the switch sensors were set on the entrance door and drawers.

In total, 19 participants were recorded, each of them had to perform 6 tasks and on average, each experiment lasted for 15 minutes. For each participant, the Kinect sensor recorded around 25000 frames (30 fps), each frame containing body joint coordinates (x,y,z). To preserve the privacy and to comply with the ethical principles of the project, the RGB image was not saved; instead, we saved only the depth (greyscale) where the human face is not recognizable. For each participant, the magnet sensors were activated on average for 30 times, sending a binary signal (a value of 1 for the open event and a value of 0 for the close event). As the participants were asked to leave the room after the completion of every task, the magnet sensor placed on the entrance door was used for automatically segmenting the data for each task.

Next, we present the tasks assigned to the participants and the respective explanations:

1. **Instructions:** *"Look for an item in the room".*
   Description: During this task, the participants were inspecting the content of dif-

Figure 3.8: Experimental settings in an university room. The experimental design was inspired by both ADL datasets, as well as problem-solving based psychological tests.

ferent boxes and cabinets for finding the indicated item. It is inspired by ADL situations, like searching for the television remote or the car keys. Additionally, we aimed at obtaining a result similar to the psychological test called the "key search Test" [16]. This test is used to diagnose cognitive functions in aging individuals by asking the participants to draw their searching strategy to find a set of keys. Similarly, our goal is to record and investigate the participants' searching strategy. Eventually, the hidden item is found, hence, we consider the data of this task as an intact and successful strategy to find the object, compared with the next task, which will be set as the distorted strategy.

2. **Instructions:** *"Look for an item (nonexistent) in the room".*
   Description: The task description is the same as above, while this time, the item is not present in the room. Therefore, after looking for the item everywhere and for a certain amount of time, all the participants started to get confused. As noticed by our collaboration with medical professionals, this situation elicits similar behaviors with those of a person with an early stage of Alzheimer's, who might not remember the placement of things. Compared to the previous task, we label the recorded searching strategy as "distorted" or "abnormal".

3. **Instructions:** *"Inspect and try to memorize the content in each cabinet. You can only open one drawer per time, per cabinet".*
   Description: Repetitive behaviors are among the most common and burdensome of the behavioral and psychological symptoms of Alzheimer's disease and yet, little research has been conducted into their manifestations. Therefore, one of the objectives of this dataset is to investigate the automatic detection and recognition of this abnormal event. We asked the participants to inspect the content of each drawer of the cabinets placed on opposite sides of the room. Additionally, a con-

straint was given, the participants could open one drawer per time, per cabinet. For example, once the first drawer of the first cabinet is explored, the participant has to investigate the first drawer of the other cabinet. This constrain forced the participants to walk repetitively to the two cabinets with the goal of eliciting repetitive behaviors.

4. **Instructions:** *"Answer the questionnaire regarding the content of the office cabinets".*
   Description: In this task, we asked the participants a set of specific questions about the items present in each drawer. For example, we asked to recall the number of books in the first cabinet or to recall in which drawer the pencils were. The participants' answers were not important to us, however, some questions were purposefully impossible to answer to force the participants to go back to the cabinets and check the content again. In this way, a rich set of actions including sitting down, writing, standing up, and searching were recorded.

5. **Instructions:** *"Prepare a cup of tea".*
   Description: The participants were asked to prepare a cup of tea, while the needed ingredients (e.g. tea bags, water, sugar, mug, and spoon) were distributed in the room at various locations. Similar to task number 1, it is inspired by ADL activities. This task has the goal of recording the participants performing an everyday activity that involves unconstrained sub-actions.

Taking advantage of the structured tasks, we can assign a different ground-truth label to each task. Specifically, task 2 and task 3 denote confusion and repetitive behaviors respectively, and they are labeled as abnormal behavioral patterns, while the other tasks are labeled as normal behavioral patterns. In the next paragraph, the trajectory analysis and abnormal behavior recognition experiments are explained.

TRAJECTORY ANALYSIS USING THE PROPOSED DATASET

The goal of our trajectory analysis is the recognition of abnormal events like confusion (task 2) and repetitive (task 3) behaviors. Clinically, confusion behaviors are usually observed when a patient experiences an episode of memory loss, and he/she gets agitated by this uncertain state [17]. Movements are not fluid and the walking patterns are not logical. The repetitive behaviors consist of repeating the same activity or going through the same locations in the room several times in a row, due to memory loss. Since these two abnormal behaviors have a significant safety as well as social impact, it is extremely important to build a framework able to recognize them.

For the data analysis, we employ the AHOT spatio-temporal descriptor explained in Section 3.2.2. However, in this scenario, we aim at extracting statistical information from each spatial block of the scene over short time intervals of $T = 2$ seconds. Differently from the application explained in the first part of the chapter, in which tracklets could be analysed in their entirety (offline), in this scenario, the goal is to detect abnormalities as soon as possible (real-time). Therefore, in this part of the chapter, we propose a Bag-of-Words(BoW) technique for aggregating the AHOT descriptors over time intervals for the detection of abnormal behaviors.

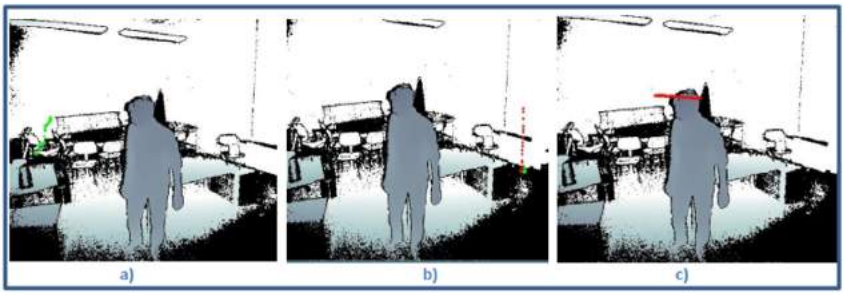


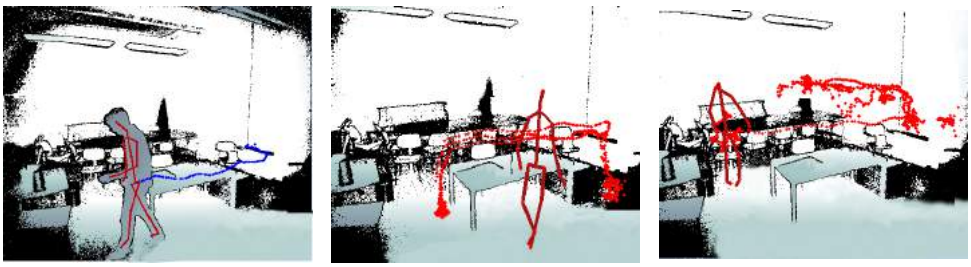
Figure 3.9: Examples of discovered trajectory words. Trajectory words are basic spatio-temporal elements that, aligned together, form higher semantic movements.

Our motivation is that the studied abnormal behaviors become distorted after being extended over a longer period of time. In other words, if a normal behavior is repeated sequentially multiple times, it may become abnormal. For example, confusion behavior may start as a normal search, however, if the pattern is extended and it becomes a random search, we may assume that the subject is confused and lost a logical strategy.

In the ICT4Life framework, trajectories information is extracted using the AHOT descriptor on $T = 2$ seconds time-windows. In this framework, we do not apply the autoencoder technique to keep the magnitude and the orientation bins well separated. Then, all the descriptors are clustered in semantic groups called " trajectory words". We aim at obtaining a clear separation between different spatio-temporal patterns, which are seen as a combination of motion patterns and spatial regions. We use the mean-shift clustering technique [18] to create a dictionary D of descriptive spatio-temporal words of size k. Mean shift clustering aims at discovering "blobs" located through the maxima of a density function. One advantage of this method is that we do not have to specify a priori the number of clusters k. We experimentally found that $k = 30$ gives stable results for the discovery of semantically meaningful spatio-temporal words.

The obtained trajectory words are short-term spatio-temporal motions that describe certain patterns common to multiple participants. In Figure 3.9, we depict some examples of trajectory words. Figure 3.9a represents a slow vertical movement around the tea kettle area, indicating that this word is mostly adopted during Task 6, where participants are asked to prepare a cup of tea. Figure 3.9b represents a fast vertical movement towards the right cabinet. This word is probably part of a searching pattern. Finally, Figure 3.9c represents a fast horizontal movement in the center of the room, the area used by the participants to walk from one part of the room to the other.

Given a set of discrete words $k \in D$, we aim at investigating their distribution in the recorded tasks. In particular, following the BoW strategy, we count the number of each word appearing in a task, making a frequency histogram from it. Finally, we use a Logistic Regression to train our Abnormal Behavior Detector (ABD) model. At training time, each task is represented as a BoW, with the task type (i.e. 1 to 6) used as ground-truth labels. At testing time, we use a 10-fold cross-validation strategy, randomly selecting 10% of the data as testing data for 10 times. The F1 classification accuracy obtained for each behavior pattern was 63% for repetitive, 69% for confused and 99% for the nor-



(a) Direct and clean movement provoked by task 1

(b) Repetitive movement provoked by task 3

(c) Random and illogical searching strategy provoked by task 2

Figure 3.10: Qualitative examples of the elicited behaviors recorded in the dataset

Testing Location	Normal Activity	Confusion	Repetitive
Psychiatry clinic (Hungary)	71,1%	93,7%	83,6%

Table 3.3: Abnormal behaviors recognition results obtained using real patients data

mal activities. Figure 3.10 shows three qualitative examples of trajectories belonging to the dataset. In Figure 3.10a, the depicted trajectory is quite direct, indicating a successful strategy for finding the hidden item. In Figure 3.10b, the depicted trajectory shows a repetitive movement from one region to the other, this movement was correctly predicted as repetitive behavior. Finally, in Figure 3.10c, the depicted trajectory shows a more twisted and complex trajectory, which indicates that the participants could not find the item and they started to wander randomly. As this experiment showed promising insights, we transferred the experimental paradigm to real healthcare locations to test our system on participants affected by Alzheimer's disease.

TRAJECTORY ANALYSIS USING REAL PATIENTS DATA

In this section, we describe the testing of the system with real Alzheimer's patients carried out in the clinic of Psychiatry and Psychotherapy in Pecs (Hungary). In order to have a comparable test with healthy participants, the same experimental paradigm was used. With the help of caregivers and doctors, four elderly participants, affected by Alzheimer's disease, were asked to perform the same 6 tasks. The experimental room was designed in the same way, Kinect camera was placed at the top of a closet and two cabinets were placed respecting the spatial arrangement of the original experimental room. To test the generalisability of our model, we used the Logistic Regression trained on the Multimodal Dataset (Section 3.5.4), and tested on real patients.

Results are showed in Table 3.3. The results from the psychiatry clinic show that normal activities were recognised with a lower accuracy then abnormal activities. This indicates that healthy participants perform normal daily activities in a different way, probably by being more efficient and fast. However, when it comes to detecting the abnormal behaviors, the algorithm shows better accuracy in the unhealthy population. This result highlights that the searching strategies of old citizens produce real contorted and random patterns that are easier to spot. These results are also confirmed by the confusion

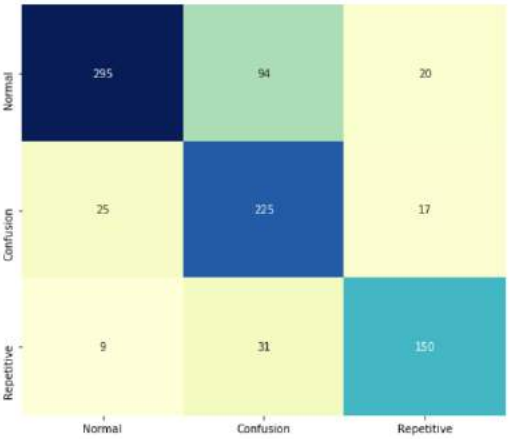


Figure 3.11: Confusion matrix obtained from the test on four Alzheimer's patients carried out in the clinic of Psychiatry and Psychotherapy in Pecs (Hungary).

matrix in Fig. 3.11.

Even though the real scenarios brought several challenges and the proposed algorithm was not always stable in the results, these experiments highlighted that the proposed models were flexible enough to cope with new environments and participants of different ages. This proves the generalizability of the proposed methods in adapting to new conditions. These results are showing that our methods can work in new environments, without a pre-training step, which could otherwise be difficult to be performed at each new scenario/ elderly house.

THE ICT4LIFE INTERFACE

In a scenario involving real patients, being able to detect abnormalities in real-time can be fundamental for their safety. Therefore, we proposed a variation of the system to send real-time feedback to their caregivers. Specifically, at testing time, every time the frequency histogram is updated with a new spatio-temporal word, the pre-trained Logistic Regression evaluates it. In other words, the classifier assesses the histogram every time it is updated ($T = 2$ seconds). Then, we set the classifier to return both the classification prediction, as well as the classification confidence. In our case, the classification confidence is the signed distance of the test sample to the hyperplane, if the returned confidence for a certain class is greater than 0, it means this class would be predicted. The confidence value can be used as a metric to understand how sure the classifier is about the analyzed test samples. On this data, we experimentally determined that the classifier confidence greater than 5 is a good threshold to have robust decisions.

Two seconds of data provide little information and the results of the classifier cannot be very reliable, therefore, the ICT4Life system stores the classification result until the classification confidence reaches the robustness threshold (greater than 5). If one of the two abnormal behaviors (e.g. confusion or repetitive behavior) exceeds the threshold, a notification is raised to call for professional help.

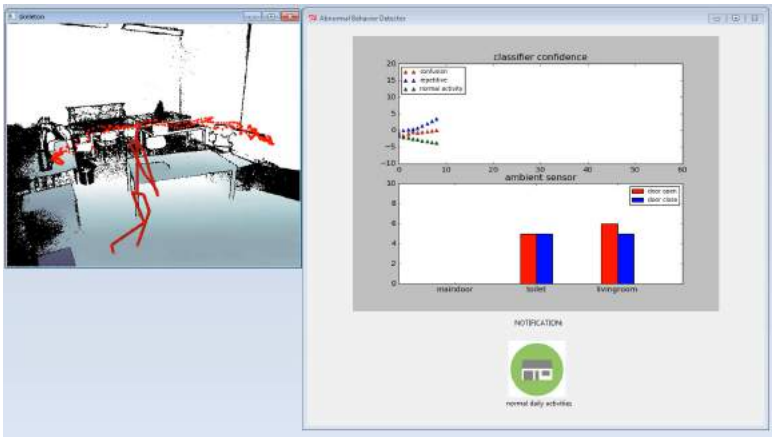


Figure 3.12: The ICT4Life Interface is composed by two main components. On the left of the screen, the skeleton tracking is displayed, while, on the right of the screen, the outputs of the machine intelligence algorithms are displayed. The output is updated every two seconds and a classification confidence value is displayed. If the value exceeds a certain threshold, a warning notification is sent to the caregivers.

In Figure 3.12, we depict the interface of the Abnormal Behavior Detection system. On the left side, the real-time video of the detected skeleton of the patient and his/her trajectory movement are displayed. Real color images are not recorded in order to respect the privacy of the participants. On the right side, the Logistic Regression classifier confidence values (updated every two seconds) are displayed. Note that the confidence value in the case of the Logistic Regression is the distance of the samples to the hyperplane. In this example, the classification confidence for the confusion label reached the threshold and the notification is raised immediately sending a message through the ICT4Life app to the doctors or caregivers.

Finally, in the ICT4Life platform, the output of the low-level subsystem is used by higher-level modules to perform multimodal fusion of different information and extracting higher-level inferences on the patient's health condition. As explained above, these higher-level modules are outside the scope of this dissertation and not in the focus of my PhD research.

3.6. CONCLUSIONS

In this chapter, we proposed a new system for detecting normal and abnormal human behaviors from surveillance data. Our approach is based on a spatial-temporal method which analyzes trajectories over a spatial grid. One important aspect of our work relies in the flexibility and generalization ability of the proposed system. Our feature extraction and clustering algorithms offer useful insights regarding the underlying distribution of the data obtained in an unsupervised way. This new feature representation enables the discovery of semantic regions based on the users' behavior over long periods of time, facilitating the annotation task. In the first part of this chapter, we tested our model on two public datasets. The obtained results prove the efficacy of our method, as we are

able to correctly classify normal vs. abnormal behavior in over 98% of the cases in both scenarios, while sparse autoencoders improve the classification accuracy by at least 1% in comparison to the raw spatial and motion descriptors.

In the second part of this chapter, the designed features were implemented in a healthcare system (ICT4Life platform), and an abnormal behavior recognition method based on BoW was tested on a novel dataset as well as in hospitals with real patients. The ICT4Life platform was tested for 4 months in real life pilots reaching the following objectives: 1) It demonstrated the feasibility and the utility of the ICT4Life platform as a global ecosystem to support independent and healthy life. 2) It led to practical field tests and user feedback/validation for issuing recommendations with regards to specific approaches more targeted towards a senior public. 3) It gathered conclusions from users' feedback in order to improve the product design.

REFERENCES

[1] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3488–3496, 2015.

[2] M. Popa, L. Rothkrantz, Z. Yang, P. Wiggers, R. Braspenning, and C. Shan, "Analysis of shopping behavior based on surveillance system," in *2010 IEEE International Conference on Systems, Man and Cybernetics*, pp. 2512–2519, IEEE, 2010.

[3] H. Mousavi, M., A. Perina, R. Chellali, and V. Mur, "Analyzing tracklets for the detection of abnormal crowd behavior," in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV 2015)*, pp. 148–155, 2015.

[4] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Int. Conf. on Computer Analysis of Images and Patterns*, pp. 332–339, 2011.

[5] Z. Saenz-de Urturi and B. Garcia-Zapirain Soto, "Kinect-based virtual game for the elderly that detects incorrect body postures in real time," *Sensors*, vol. 16, no. 5, p. 704, 2016.

[6] A. Abrams, J. Tucek, N. Jacobs, and R. Pless, "LOST: Longterm Observation of Scenes (with Tracks)," in *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 297–304, 2012.

[7] K. B.-Y. Wong, T. Zhang, and H. Aghajan, "Data Fusion with a Dense Sensor Network for Anomaly Detection in Smart Homes," *Human Behavior Understanding in Networked Sensing*, pp. 45–73, 2014.

[8] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," *Lecture Notes in Computer Science (LNCS)*, vol. 6311, pp. 577–590, 2010.

[9] J. Masci, U. Meier, D. Ciresan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *21th Int. Conf. on Artificial Neural Networks (ICAN'11)*, pp. 52–59, 2011.

[10] J. Ngiam, A. Khosla, and M. Kim, "Multimodal deep learning," *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, pp. 1–9, 2010.

[11] P. Urwyler, R. Stucki, L. Rampa, R. Müri, U. P. Mosimann, and T. Nef, "Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living," *Scientific reports*, vol. 7, p. 42084, 2017.

[12] T. Theodoridis, V. Solachidis, N. Vretos, and P. Daras, "Human fall detection from acceleration measurements using a recurrent neural network," in *Precision Medicine Powered by pHealth and Connected Health*, pp. 145–149, Springer, 2018.

[13] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[14] C. Zhang, Y. Tian, and E. Capezuti, "Privacy preserving automatic fall detection for elderly using rgbd cameras," in *International Conference on Computers for Handicapped Persons*, pp. 625–633, Springer, 2012.

[15] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[16] J. M. Oosterman, M. Molenveld, M. G. OLDE RIKKERT, and R. P. Kessels, "Diagnostic utility of the key search test as a measure of executive functions," *Psychogeriatrics*, vol. 10, no. 4, pp. 173–178, 2010.

[17] H. Chinaei, L. C. Currie, A. Danks, H. Lin, T. Mehta, and F. Rudzicz, "Identifying and avoiding confusion in dialogue with people with alzheimer's disease," *Computational Linguistics*, vol. 43, no. 2, pp. 377–406, 2017.

[18] M. ComaniciuD, "A robust approach toward feature spaceanalysis [j]," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 313–329, 2002.

4

A HIERARCHICAL AUTOENCODER LEARNING MODEL FOR PATH PREDICTION AND ABNORMALITY DETECTION

This chapter is based on the following publication:

- D. Dotti, M. Popa, and S. Asteriadis, "A hierarchical autoencoder learning model for path prediction and abnormality detection", *Pattern Recognition Letters*, vol. 130, pp. 216–224, 2020.

4.1. INTRODUCTION

Understanding and predicting human motion in complex real world scenarios has a vast number of applications, from designing intelligent security systems to deploying socially-aware robots [1]. In the previous chapter, we presented a framework that enables the discovery of semantic motion regions based on the users' behaviour using trajectory data. Our method analyzed short motion trajectories (i.e. tracklets) for the discovery of normal and abnormal spatio-temporal behaviors. Analyzing human motion trajectories is challenging in different aspects, as human motion is affected by internal

needs as well as by external environmental factors. Internal needs are for example, going to the supermarket to do grocery shopping, while environmental factors that may affect human' path are for example, moving obstacles along the way. Therefore, one of the biggest challenges for a smart surveillance system, is to embed these two factors in order to understand what is going on in the scene, and more importantly, to predict what can happen next.

A popular approach towards the prediction of future pedestrian trajectories consists in modeling their interactions in crowded spaces. Interaction can be defined by a set of hand-crafted rules like social forces [2], stationary crowds influence [3], or neighbors movements [4]. Although these models are useful for estimating the forthcoming motion status, we argue that the destination goal of an object does not change according to the people around it. Specifically, even though the local motion patterns may vary according to the social interactions, the global motion descriptors need to be invariant to small deviations, for efficiently predicting the long-term evolution of the trajectories. This degree of anticipation can help in inferring targets' route [5], intents [6], and destinations [7] of moving objects. For example, in outdoor scenarios like the ones described in [8], it is important to detect immediately whether a car lost control and is approaching a pedestrian walking area, or to forecast a pedestrian's walking path to detect if they are moving into a restricted area.

In this chapter, we propose a hierarchical model capable to learn and predict future motion trajectories in real-time. The lower levels in our hierarchical architecture aim at modeling the ambiguity of short spatio-temporal trajectories. Like stated above, short trajectories may deviate abruptly due to possibly moving obstacles such as other pedestrians on the way. On the other hand, the higher levels of the model aim at capturing statistically significant combinations of short spatio-temporal trajectories. Being able to learn more robust motion characteristics is meaningful for revealing more global characteristics such as the final goals/destinations of the objects.

Figure 4.1 shows a higher level description of the proposed method. Given the original trajectory depicted in red, we firstly encode direction and velocity of short motion patches, using the lower levels of the hierarchy. Then, in the higher levels, we combine motion patches to learn meaningful combinations of short individual patches. Finally, we aim to model temporal motion transitions using Bayesian probability, by inferring the future trajectory step, given the current motion descriptor. The predicted motion patches (yellow lines) are the result of the inference layer applied on the observed motion patterns (blue lines). Our system predicts the orientation direction, as well as the velocity of the future pattern. Speed and orientation information are useful to summarize the analyzed behavior; in this case, the target object is quite fast when it enters the scene (bottom right of the figure), and it slowly reduces its speed when approaching a possible obstacle, the stairs (top center of the figure).

Several hierarchical architectures have been proposed for modeling long-term motion evolution [9], [4]. For example, Convolutional Neural Network (CNN) frameworks [10] often obtained impressive results that are highly competitive, however, due to the large number of parameters to fine-tune, they require large amounts of labeled data. In this chapter, we propose an unsupervised hierarchical model based on autoencoders [11], that has the advantage of requiring less data and less computational complexity,

Figure 4.1: Our predictions of future trajectory motion patches (colored in yellow) contain direction orientation as well as speed information. Orientation describes the object's movements in the scene, while speed (embedded using our new feature representation on the right) is important for behavior understanding, for example to distinguish between a pedestrian that walks calmly and a pedestrian rushing somewhere. The observed trajectory state is colored in blue and the ground truth is colored in red.

while it is still powerful enough to capture both local and higher semantic motion information. We motivate our design choices using three public datasets (i.e. New York Grand Central station introduced in [3], LOST dataset proposed in [8], and VIRAT dataset proposed in [12]) , where our proposed framework is comparable with state-of-the-art deep neural network methods.

Our learning process is data-driven, and it aims at obtaining an improved feature representation by employing sparsity and a cross-entropy based optimization approach, in combination with autoencoders. One contribution of our work consists in designing a novel feature representation inspired from object recognition. 3-D spatio-temporal motion patches (x, y, t) are mapped to 2-D image patches (x, y), translating the temporal information into pixel intensity values. Profiting from the ability of sparse autoencoders to learn meaningful patterns from 2-D gray-scale images, the bottom level of the hierarchical model learns basic spatio-temporal patches. The discovered local patterns are useful to describe small movements but lack global descriptive power. Thus, on the next hierarchical level, we encode longer parts of trajectories by increasing the spatial grid. More distinctive motion patterns are learned, representing statistically meaningful combinations of low-level elements as well as preserving local information like orientation and speed. The new descriptors, due to the non-linear transformation using first level weights, are more invariant to small shifts or deviations, and incorporate more discriminative power for long-term trajectory prediction. Finally, we use a Bayesian probabilistic framework to model the dynamics of the learned features for trajectory path forecasting, direction prediction, and abnormality detection. We tested the flexibility of our approach in both outdoor and indoor environments, using The Long-term Observation of

Scenes (with Tracks) or LOST dataset [8], the New York Grand Central dataset (GC) [3], and the VIRAT surveillance dataset [12] (see the datasets' introduction in Chapter 2).

The contributions of this work are the following: First, we build a new hierarchical model that learns, in the bottom layer, local motion patches, and in the subsequent levels it learns statistically meaningful co-occurrences of local patterns, to form higher semantic motion features. Its power resides in its inexpensive learning needs in combination with a general and flexible learning mechanism. Our model is different from the classical deep autoencoder [13], as we increase the size of the receptive field in every hierarchical level. Second, a novel motion feature representation which embeds motion speed and orientation is created, and fed as input into an autoencoder learning framework. Third, to demonstrate the efficiency of our method in both outdoor and indoor scenarios, we apply our model in various applications like trajectory path prediction, destination prediction, and abnormality detection.

The remainder of this chapter contains an introduction of our proposed hierarchical model in Section 4.2, along with the description of feature representations and learning approaches. Section 4.3 contains the description of high level inference, while in Section 4.5 the obtained results are presented and evaluated against state-of-the-art methods. Finally, the conclusions and the directions for future work are included in Section 4.6.

4.2. MODEL DESCRIPTION

In this chapter, we propose a hierarchical model for representing and learning trajectories in an unsupervised approach. Surveillance cameras collect huge amounts of data over weeks, and it takes a great deal of human effort to store, process, and analyze the data to detect rare abnormal situations. In this context, unsupervised methods tackling the understanding of surveillance data are critical in supporting and optimizing vigilance tasks.

Inspired from the language processing field [14, 15], we introduce a word-sentence analogy to trajectory composition towards the intended destination. A trajectory can be compared to a "sentence", as its elements (trajectory units - words) are combined in a logical way, depicting the movement goal. Firstly, the bottom layer learns minimal spatio-temporal units of trajectories (words), and secondly, higher layers learn combinations of these units resulting in the representation of bigger spatio-temporal patterns (sentences). These higher concepts are used to extract the semantic structure of the data, as in natural language processing, where the relations between words determine the meaning of a sentence. For example, if the data contains a diagonal road like in Fig. 4.2, the first layer of the model will extract diagonal units of trajectories, that will strongly correlate in higher layers with other diagonal units, creating a stronger diagonal movement.

Fig. 4.2 depicts the proposed hierarchical architecture. For visualization purposes, our model is formed by 2 layers, however, the number of layers is a parameter that can be changed depending on the data. Every hierarchical layer is composed by two components: the feature extraction component and the learning component. The feature extraction component in the bottom layer aims to embed local trajectory features such as orientation and speed from short temporal windows patches. To do so, we propose a new spatio-temporal descriptor called "Trajectory units" u. These trajectory units are

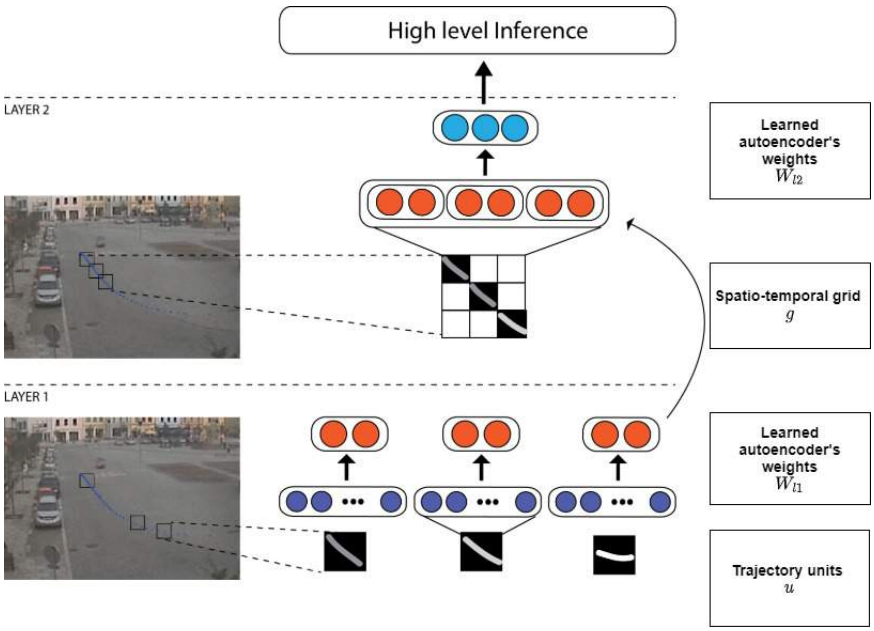


Figure 4.2: The architecture of the proposed model. Layer 1: Trajectory fragments are transformed into 2-D patches and fed into the bottom layer autoencoder. Layer 2: combinations of trajectory fragments encoded in layer 1 are learned in the second level autoencoder. Finally, an inference layer is added for trajectory prediction and abnormality detection.

the input of the bottom layer's learning component made of a greedy autoencoder. The first autoencoder aims to encode atomic motion patterns, yet, the learned representation is limited to a short spatio-temporal scale, and it is not powerful enough to be used for behavior discrimination. Therefore, inspired by the hierarchical feature learning for object recognition [16], [17], we build a second hierarchical layer to encode bigger trajectory parts. A key point of the proposed framework is that the feature extraction component in the second layer consists of trajectory units encoded using the first layer's autoencoder. In this way, we aim to transfer the knowledge acquired from the previous layer to the subsequent layer. Furthermore, we aim to enhance the knowledge of our framework by encoding bigger trajectory patches using spatio-temporal grids. These spatio-temporal grids are the input of the second layer learning component made of a second greedy autoencoder. The autoencoder in the second layer aims to learn the correlation between adjacent atomic motion patches, obtaining a better understanding of bigger trajectory parts. This second layer in the hierarchy bears the following benefits: 1) bigger spatio-temporal features can lead to a better understanding of behaviors, and 2) the new motion representation is invariant to small shifts or deviations of trajectory points. Due to the modular learning strategy, depending on the trajectories length in the analyzed scenario, additional layers can be added to encode increasingly larger spatio-temporal motion patterns. A final learning layer is added on top to model the probability distribution of trajectory units for applications such as path prediction, and abnormality

detection. In the following sections we describe the proposed approach in detail.

4.2.1. GREEDY HIERARCHICAL LEARNING

An autoencoder (AE) is a neural network which aims to minimize the reconstruction error between the input and the output in an unsupervised way [11]. The process of mapping from input X to a reconstructed representation $\hat{X}$, is useful for estimating the underlying distribution of the data and for obtaining an improved feature representation. We propose a hierarchical autoencoder architecture, trained in a greedy layer-wise fashion [13], where, at every layer of the hierarchy, we encode trajectory patches using a single hidden layer autoencoder (see Chapter 2 for an introduction of Autoencoders).

Each hierarchical level $l = [l_1, \ldots, l_i]$ is trained in a sequential manner. Once the first autoencoder layer l_1 is trained, we can use the learned weights $W_1^{l_1}$ as a pre-trained initialization for a new Autoencoder in the next level l_i of the hierarchy.

4.2.2. BOTTOM LAYER FEATURE EXTRACTION

In this section, we propose a new spatio-temporal descriptor that aims to provide a new representation of trajectory points, embedding local features like orientation and speed. We start by extracting trajectory points within random window patches of size $s \times s$. Encouraged by the impressive results of autoencoders in image reconstruction, we transform the spatio-temporal patches containing the trajectory points (x, y, t) into 2-D gray-scale images (x, y). Single trajectory points are too sparse to be reconstructed in an efficient way and the information they convey is limited, hence, every point inside a patch is connected using image processing techniques. The distance between points is computed and used as pixel intensity values for the connecting line. The resulting gray-scale values will represent the distance between points, with bigger distances between points (higher speed of the object) reflected by higher pixel values. The rest of the pixels are set to a value equal to zero and are considered as background.

Fig. 4.3 shows an example of first layer trajectory units u, where local motion is represented by pixel intensity values (object speed) and the orientation of the stroke. Speed information is very important in motion analysis because it allows the discrimination between different objects such as pedestrians and vehicles in an outdoor scenario, or different behaviors such as running versus walking. Moreover, the orientation describes object movements in the spatial dimension, and is useful for exploring the path of the trajectories. Finally, the extracted images are vectorized and fed into the Autoencoder.

4.2.3. BOTTOM LAYER LEARNING

The input to the first layer autoencoder is represented by $s \times s$ gray-scale patches containing trajectory units u of different orientations and speed. See Eq. 2.8 from Chapter 2 for the Autoencoder formulation.

An important parameter for autoencoders is represented by the size of $h(x)$ which is optimized based on the minimum reconstruction error (see Fig. 4.12). In Fig. 4.4, we start by visually inspecting the set of weights learned by an autoencoder with m hidden units on the LOST dataset.

Orientation patterns are easily visible, describing the structure of the data. There are more diagonal filters than horizontal and vertical ones, meaning that diagonal move-

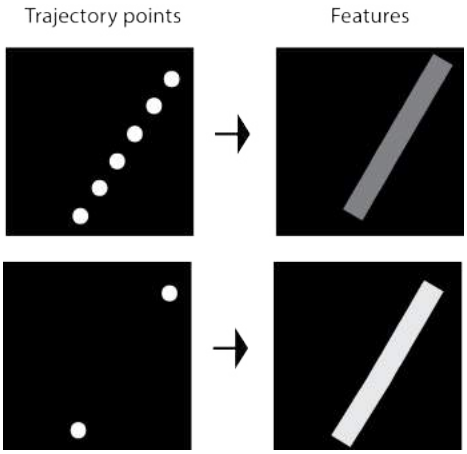


Figure 4.3: First Layer feature extraction. Consecutive trajectory data is transformed into our proposed trajectory patch. Trajectory patches embed motion characteristics such as speed, represented by gray intensity values, and orientation.

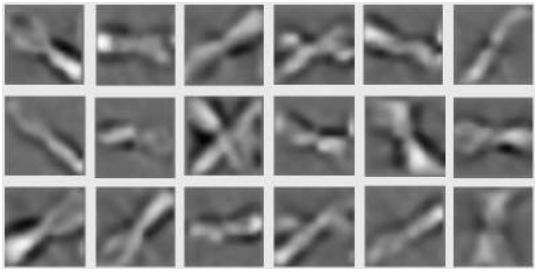


Figure 4.4: Visualization of a set of weights trained in the first hierarchical layer.

ments are more common in this dataset. Moreover, speed is embedded in these weights, as the different trajectory stroke gray-scale values depict different speed rates of the objects. The learned weights represent basic trajectory units, describing their local motion. To obtain more descriptive and global motion features, in the next layers, we encode bigger parts of trajectories by aggregating their atomic descriptors using a spatial grid.

4.2.4. SECOND LAYER FEATURE EXTRACTION

Motion information differs in every scenario, for example, outdoor scenes usually contain longer and more linear trajectories, while, indoor scenes contain trajectories that are more fragmented and curved due to limited space. Therefore, based on the scenario, a different number of hierarchical layers may be needed to create global spatio-temporal features. In this section, we start by describing the construction of the second hierarchical layer l_2, as additional hierarchical layers can be built following the same methodology.

Fig. 4.5 shows the feature extraction process: given a bottom layer trajectory unit

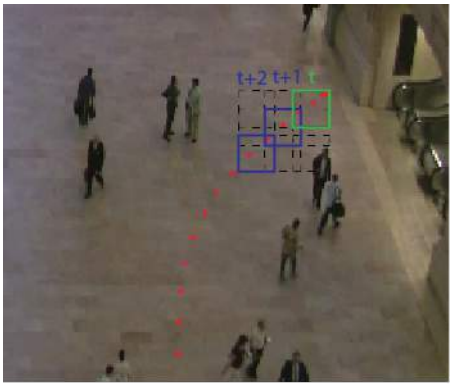


Figure 4.5: Second Layer feature extraction. A spatio-temporal grid composed of $g = 3 \times 3$ patches is used to cover for all the possible trajectory directions (blue color) given the current patch at time t (green color).

u^{l_1} at time t (colored in green), we create a spatial grid formed by $g = 3 \times 3$ patches, to account for all the possible future directions. In this way, our grid encodes the future trajectory units at time $t+1$ and $t+2$ (colored in blue). Every patch in the grid is encoded using the weights ($W_1^{l_1}$) learned on the previous layer l_1, and concatenated (Eq. 4.1).

$$X^{l_2} = [h^{l_1}(j)]_{j=1,\ldots,(g)} \tag{4.1}$$

where $h^{l_1}(j)$ represents the patches in the grid g encoded using the previous autoencoder l_1.

The encoding process allows to have higher order features more invariant to noise (i.e. trajectory points shift). Furthermore, their concatenation preserves the temporal and spatial relationship between neighboring patches, creating feature vectors X^{l_2} that capture higher semantic information. Finally, the new feature vector is fed to the second autoencoder.

4.2.5. SECOND LAYER LEARNING

The input to the second autoencoder consists of feature vectors X^{l_2} of size $(m \times g)$. Trajectory units embedded using the first autoencoder (composed by m hidden units), are disposed in a spatial grid of $g = 3 \times 3$ patches and fed to the second autoencoder. The advantage to have a second autoencoder and not a single deep autoencoder lies in the fact that we can increase the input size in every hierarchical level to learn wider trajectory patches on every level.

As described above, the input to this autoencoder is represented by higher order feature vectors ($h^{l1}(j)$) and not 2-D images like in the bottom layer. Therefore, in order to visualize the learned weights, we need to use the current decoder weights W_2^{l2} as well as the previous autoencoder decoder weights W_2^{l1} to map the features back to the original input patches (Eq. 4.2).

$$\hat{X}^{l_2} = g_0^{l2}(W_2^{l2} h^{l_2}(X^{l_2}) + b^{l2})$$
$$\hat{X}^{l_1} = g_0^{l1}(W_2^{l1} \hat{X}^{l_2}(j) + b^{l1})_{j=1,\ldots,(g)} \tag{4.2}$$

where g_0^{l2} and g_0^{l1} are the decoder functions of the two autoencoders in the hierarchical layers (Eq. 2.9 in Chapter 2). W_2^{l2} and W_2^{l1} are the decoder weights, and b^{l1} and b^{l2} are the bias weights. g_0^{l2} reconstructs the entire grid g, then, we use g_0^{l1} to reconstruct every trajectory patch in g.

Fig. 4.6 shows the reconstruction results of the weights $W_2^{l_2}$ learned in the second hierarchical layer. More distinctive motion patterns can be noticed, remarkably preserving low-level information such as orientation and speed. The encoded information captures statistically significant combinations of the trajectory units u^{l_1} used as input in the bottom layer. For example, all the units have the same orientation (first row, first column from the left), depicting a straight movement, or units have different orientations (second row, second column from the left), depicting changes in direction.

Ascending in the hierarchy of the model, we are able to learn more complex spatio-temporal structures of trajectories. The aggregation and co-occurrences of minimal trajectory patches generate more global and discriminative motion features, that can be used for behavior understanding. Since trajectory information varies depending on the scenario, additional hierarchical layers can be built following the same methodology.

4.3. HIGH LAYER INFERENCE

4.3.1. OVERVIEW

We demonstrated that our model is able to learn discriminative motion patterns in an unsupervised manner using a hierarchical architecture based on autoencoders. In this section, we explain how the discovered pattern dynamics are modeled for trajectory path prediction and abnormality detection. We aim to model temporal transitions using Bayesian probability, by inferring the future trajectory step, given the current motion descriptor. Nevertheless, modeling all the small deviations that trajectories can have would make the probability framework very inefficient. Therefore, we quantize the motion features using k-means clustering. Motivated by [18], we take advantage of the non-linear embedding learned by our hierarchical autoencoder architecture, to run the k-means algorithm on the encoded output of our last layer h^{l_2}. In the next sections, we show how the clustering method is useful at finding a dictionary D^l of significant motion descriptors on layer l. Then, we describe our Bayesian framework built using sequences of dictionary words, and finally, we explain our training framework for applications like path prediction, as well as abnormality detection.

4.3.2. MOTION DICTIONARY CONSTRUCTION

We aim to create a high level dictionary D for finding a set of meaningful k clusters of motion descriptors. Inspired by [18], we propose first to use the hierarchical autoencoder to encode spatio-temporal patches, and then run the k-means clustering algorithm on the obtained embedded feature representation $h^l(j)$ on each layer l. Our motivation is

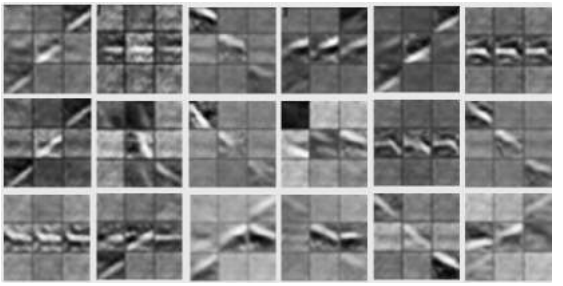


Figure 4.6: Set of weights trained in the second layer of the hierarchy. Higher layers of the hierarchy learn statistically significant combinations of local motion patches.

two-fold: 1) Since the dimension of the hidden layers is lower than the input layer, the network is bounded to keep only the most representative motion trajectory units, removing the part of the data that is not important (outliers). This reduces the variability of the data making the clustering problem less hard. 2) It has been shown that using constraints like sparsity, the hidden units specialize on different types of input.

Therefore, our hypothesis is that different subgroups of hidden units are activated for specific motion patterns. For example, a diagonal motion element will activate different hidden units from a horizontal one; in the same way, the activation pattern will be different for a trajectory unit containing high pixel values (high speed) than for the one with low pixel values (slow speed).

We found the number of clusters k on each layer l using the optimization of the intra-clusters variance with respect to the total variance. Similar motion patterns are grouped together, creating a dictionary of words $k \in D^l$ representing different spatio-temporal motion patterns. Modeling the dynamics of these words enables the understanding of longer parts of trajectories for destination prediction and abnormality detection.

4.3.3. PATH PREDICTION USING BAYESIAN NETWORKS

Given a set of discrete words $k \in D^l$, where D^l is the dictionary of words in layer l, we aim to model their temporal sequence using a probabilistic graphical model. Bayesian Networks (BNs) are chosen due to their ability of modeling joint uncertainties and complex information without requiring a vast amount of training data. The probability distribution of the two words $P(k_i, k_j)$ is computed to learn significant motion dynamics in the data. We maintain the spatio-temporal grid g described in Section 4.2.4 as a structure for our BNs, however, we map every encoded trajectory unit h^l in the grid to a word $k \in D^l$.

BNs are represented by directed acyclic graphs (DAGs), where nodes represent random variables and links between them represent their direct dependency. The most natural task using a BN is to compute the posterior probability using the Bayesian theorem. At the same time, DAGs determine the joint probability distribution (JPD) over all the variables. Once the JPD is calculated over all the positions in the spatio-temporal grid, we can infer the structure of the data by discovering the conditional probability between nodes.

In the context of our framework, we estimate the probability of the trajectory element

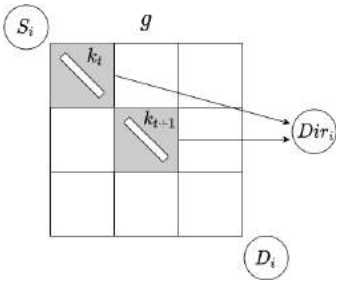


Figure 4.7: Example of data construction for Destination prediction. Source S_i and Destination D_i are plugged into the spatio-temporal grid g. Dir_i is automatically obtained using the positions of k_t and k_{t+1} in the grid.

at time $t+2$, $P(k_{t+2})$, conditioned on the previous elements k_t, k_t+1 as $P(k_{t+2}|k_t, k_{t+1})$, using:

$$P(k_{t+2}|k_t, k_{t+1}) = \frac{P(k_t, k_{t+1}|k_{t+2})P(k_{t+2})}{P(k_t, k_{t+1})} \tag{4.3}$$

where the prior probabilities $P(k_{t+2})$ and $P(k_t, k_{t+1})$ are estimated using the probability distribution of the pair of elements in the dataset. In this way, the label k^* of the trajectory element k_{t+2} is calculated given the evidence, using the Maximum a Posteriori (MAP) principle included in Equation 4.4:

$$k^* = \underset{\hat{k} \in D^l}{\operatorname{argmax}} P(k_{t+2} = \hat{k}|k_t, k_{t+1}) \tag{4.4}$$

4.3.4. LONG-TERM PATH MODELING FOR DESTINATION PREDICTION

The prediction of long-term spatio-temporal intervals is an important task for smart surveillance cameras. In this work, we aim at learning the entire trajectory progress using their source and destination information. Similar to [3], ten Source/Destination region blobs $S_i, D_i (i \in [1,10])$ are annotated on the GC Dataset (Fig. 4.14). Every trajectory is assigned to the pair Source/Destination by using their pixel coordinates. For example, in Fig. 4.14 (a), the trajectory starts in blob 3 (Source label), and finishes in blob 8 (Destination label).

To model a long-term trajectory path, we use the same architecture as described in Section 4.3.3, where the spatio-temporal grid g is fed into the BN framework. In addition to k_t and k_{t+1}, in this experiment, we plug into our BN network their source label S_i as well as their direction information dir_i. Their direction information is obtained in an unsupervised way using g. Figure 4.7 contains an example of how we construct the training data in this experiment. The spatio-temporal grid g contains k_t and k_{t+1}, and, we can automatically extract their direction thanks to their positions in the grid. In this example, the occupied blocks in the grid have an angle of 315 °, therefore, $dir_i = 315$.

Finally, given a trajectory words k_t and k_{t+1}, we aim at predicting the future trajectory word at time $t+2$ as well as their destination D_i conditioned on the previous elements $P(k_{t+2}, D_i|k_t, k_{t+1}, dir_i)$ using the Maximum a Posteriori (MAP) principle included in the Eq. 4.4.

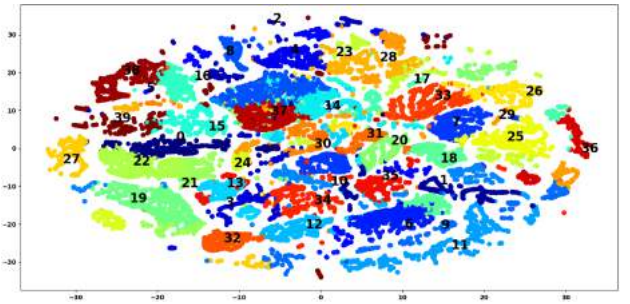


Figure 4.8: Visualization of the embedding space learned by the autoencoders on the first hierarchical layer on the GC dataset.

4.3.5. ABNORMALITY DETECTION MODELING

Autoecoder reconstruction error has been previously exploited to define abnormalities [19], following the idea that regular motion patterns are reconstructed with a lower error than unexpected ones. In this chapter, atypical observations are reported by both the autoencoder and the BN module. Unforeseen motion dynamics (e.g. sudden speed or direction change) generate a high reconstruction error in the autoencoder modules. Additionally, word sequence dynamics which do not respect Eq. 4.4 in the Bayesian framework, are reported as abnormalities.

4.4. ANALYSIS OF THE HIERARCHICAL LEARNING MODEL

In this section, we begin with a set of ablation studies to validate the efficiency of the proposed architecture using the GC dataset [3], which contains trajectory data extracted from individuals in the Central Station of New York. We randomly select 80% of the trajectories as training set, 10% as validation set, and 10% as test set. An advantage of the proposed framework is that it is modular, hence, the future trajectory predictions can be computed from every layer. In this section, experiments are conducted on each layer separately as it was the highest in the hierarchy. Note that the reported results are computed on the validation set.

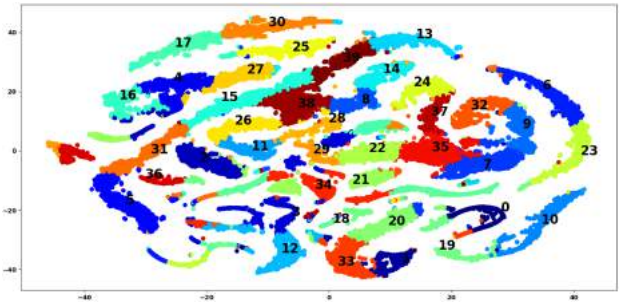


Figure 4.9: Visualization of the embedding space learned by the autoencoders on the second hierarchical layer on the GC dataset.

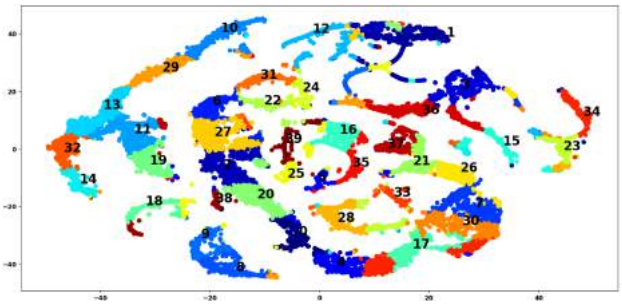


Figure 4.10: Visualization of the embedding space learned by the autoencoders on the third hierarchical layer on the GC dataset.

In Fig. 4.8, Fig. 4.9, and Fig. 4.10, we show the autoencoder embedding subspace for $l = 3$ hierarchical layers using the t-sne visualization tool [20]. Fig. 4.8 indicates the embedding subspace of the features learned in the first hierarchical layer l_1, Fig. 4.9 indicates the embedding subspace of the features learned by the second hierarchical layer l_2, and finally, Fig. 4.10 indicates the embedding subspace of the features learned by the third hierarchical layer l_3. As shown, higher hierarchical layers provide a significantly more discriminative subspace compared to the bottom layers, making the clustering assignment more definite.

However, as the higher hierarchical layers embed larger motion patches, they are more sensitive to the amount of training data as well as to the trajectory lengths. Therefore, in Fig. 4.11, and Fig. 4.12, we display the effect of important parameter variations on the prediction performance using the GC dataset. In Fig. 4.11, the effect of the spatio-temporal window size $s \times s$ on the prediction performance of future trajectory state is displayed. The window size is established in the bottom layer (i.e. Layer1), then, on each subsequent layer, the window size is extended (e.g. Layer1=10×10, Layer2=30×30, and Layer3=60×60) to encode the previous layer temporal grid $g = 3 \times 3$ (Fig. 4.11). Results show that the performance of the third layer is lower than the one obtained from the first two layers. This indicates that there is not enough data to train three hierarchical layers, hence, in this chapter, we use the proposed model to learn a two-layer hierarchy. Due to diverse speed of tracked objects, our spatio-temporal grids embed different amounts of trajectory points, allowing the system to learn meaningful dynamics of trajectory patches without a fixed time window. Specifically, the first layer encodes patches ranging from 2 to 5 seconds of data, while the second layer encodes patches ranging from 5 to 10 seconds of data. Finally, a window size of size 20×20 provides the best results and it will be used for the final experiments.

In Fig. 4.12, the variation of the size of autoencoder hidden layers is shown. In this experiment, we aim to minimize the reconstruction error (denoted as mean squared error - mse) on $\hat{X}_i$ as well as to maximize the prediction accuracy. The x-axis indicates the ratio between the input and the hidden layer size. Finally, all the parameters considered in our approach are presented in detail in Table 4.1.

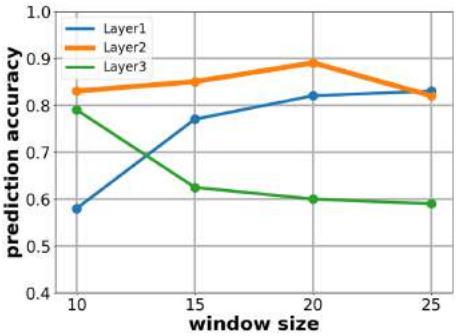


Figure 4.11: Sensitivity analysis to understand the window size parameter effect on the GC dataset. X-axis indicates the window size for the bottom layer, then, for every subsequent layer the window size is extended to encode the previous layer temporal grid (e.g. Layer1=10 × 10, Layer2=30 × 30, and Layer3=60 × 60). The first two hierarchical layers provide the best results given the dataset size, therefore, for the final prediction experiment, a two-layer framework is built. Window size 20 × 20 provides the best results.

4.5. EXPERIMENTS

The general goal of our experiments is to apply the proposed approach in various applications such as short and long term path prediction as well as abnormality detection using three challenging surveillance datasets, comparing its efficiency with other state-of-the-art methods.

Table 4.1: Model parameters used in the feature extraction and learning phase, for the considered datasets (LOST, GC, and VIRAT dataset).

Parameters	LOST	GC	VIRAT
Patch size on l^1, (s)	18	20	20
Patch size on l^2, (r)	56	60	60
Grid size (g)	9	9	9
Dictionary size $(k \in D^l)$	40	40	20
AE first layer $(\|W_1^1\| = m^1)$	144	169	169
AE second layer $(\|W_1^2\| = m^2)$	400	625	625

4.5.1. EXPERIMENT 1: PATH PREDICTION

In the previous section, we assessed the hierarchical learning structure of our model, while, in this section, we evaluate whether the hierarchical structure improves the prediction performance. In the evaluation phase, given the first two patches of the spatio-temporal grid at time t and $t+1$, we aim to predict the cluster label k^* that describes the next motion stroke at time $t+2$. F1 prediction accuracy results are presented in Table 4.2, where the best prediction result (96.9%) is achieved by computing the inference using Layer2 features. Moreover, in the LOST dataset, prediction results are significantly higher than the results obtained in the GC dataset. This is due to the nature of the different behaviors in the recorded scenarios. Specifically, the chosen scene in

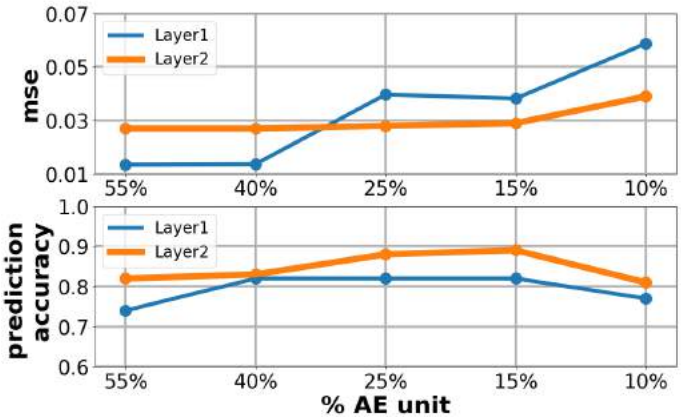


Figure 4.12: Sensitivity analysis to understand the variation of the autoencoder hidden layers' size on the GC dataset. The hidden space size for the two hierarchical layers is set to minimize the reconstruction error (mse) and maximize the prediction accuracy. X-axis indicates the percentage of data preserved after the dimensionality reduction.

the LOST dataset presents a road intersection with pedestrians and vehicles. Trajectories are mostly straight lines, indicating objects' intention to reach a specific destination using the shortest route. On the other hand, in the GC dataset, numerous examples of twisted trajectories, created by people walking without an evident goal (i.e. spending time waiting for the train) are recorded. This characteristic of the data seems to affect the prediction of Layer1 (82.2%), whereas Layer2 features prove to be more robust (89.1%).

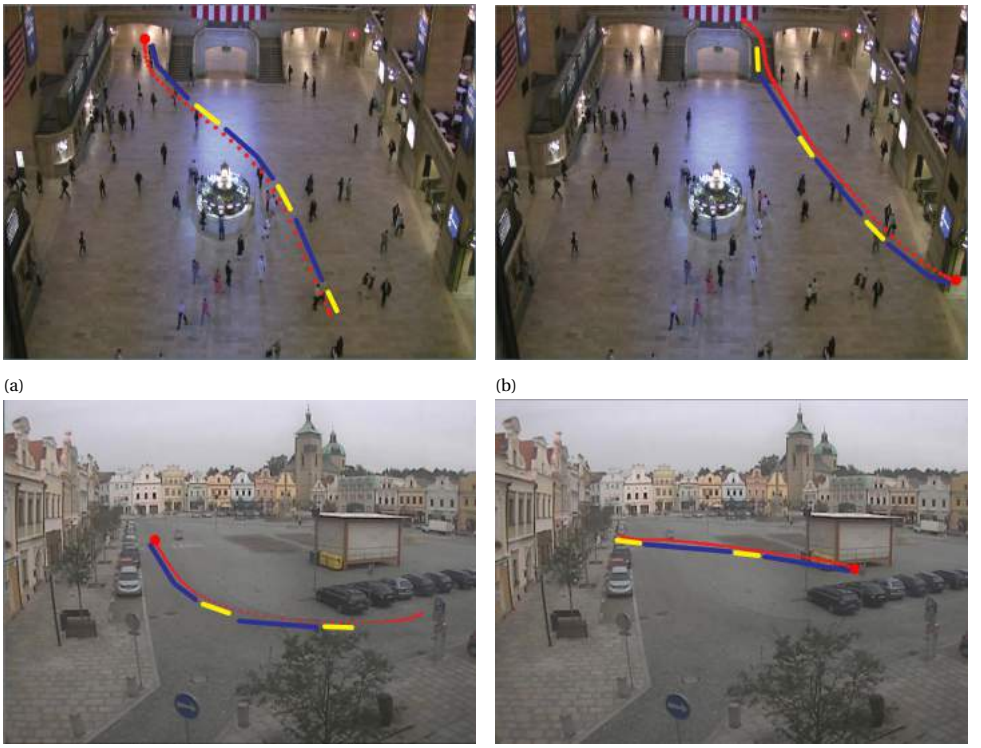
Table 4.2: Trajectory cluster label prediction accuracy.

Hierarchical layers	LOST dataset	GC dataset
Layer1	88.5%	82.2%
Layer2	**96.9%**	**89.1%**

The multi-stage learning approach shows to be more invariant to local distortions of trajectories, as autoencoders tend to learn a smoother representation of the data, canceling local noise effects. Overall, Layer2 inference obtains the best result, proving that by ascending the hierarchy, we obtain more global and discriminative features, useful for behavior prediction and understanding.

4.5.2. PATH PREDICTION: QUALITATIVE RESULTS

Examples of trajectory prediction are shown in Fig. 4.13, where, the ground truth trajectory is colored in red, blue depicts the observation at time t and $t+1$, and our prediction of the future trajectory state at time $t+2$ is colored in yellow. Fig. 4.13a and Fig. 4.13b show motion patterns of pedestrians in New York Train Station taken from our analysis of the GC dataset. The data is very challenging because trajectories are full of deviations



(a) (b)

(c) (d)

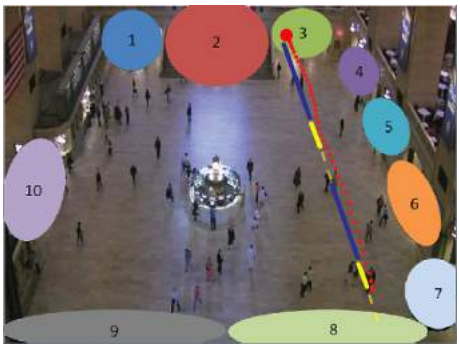
Figure 4.13: Path prediction in indoor and outdoor scenarios. Predicted trajectory elements are colored in yellow, observed trajectory patches are colored in blue, and the trajectory ground truth is colored in red.

and small shifting due to the crowded scene. Our system shows to be robust to these challenges, being able to correctly predict the objects' motions. On the other hand, different movement patterns caused by different objects are displayed in Fig. 4.13c and Fig. 4.13d taken from our analysis on the LOST dataset. In Fig. 4.13c, our system predicts correctly the motion of a car transiting on the road, while in Fig. 4.13d, the trajectory of a pedestrian walking on the zebra crossing is correctly predicted.

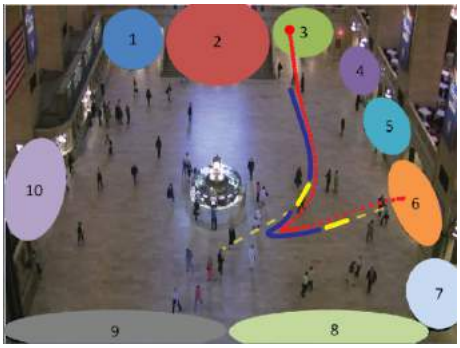
4.5.3. EXPERIMENT 2: DESTINATION PREDICTION ON THE GC DATASET

In order to demonstrate that our model is able to predict long trajectory paths, in this section, we perform the destination prediction experiment using both the first layer as well as the second layer's features. We follow the experimental procedure described in [3], to predict the destination area of every trajectory (see details in Section 4.3.4).

We randomly select 80% of the trajectories as the training set, 20% as test set in a K-Fold cross validation experiment. We discarded very short trajectories which do not have any source/destination labels. Given a test trajectory, our model recursively analyzes every motion patch at time t and $t+1$, generating the next motion state at $t+2$, as well as the destination region D_i. Similar to [10], the predicted trajectory patch is fed back into



(a) Pedestrians take small path deviations (due to the crowded space) without affecting their final destination. Our model correctly predicts the trajectory progress as well as the final destination.



(b) Drastic modification of trajectory direction and destination. Our method encodes the change of direction and, at the second iteration, correctly predicts the final destination.

Figure 4.14: Qualitative examples depicting the destination prediction in the indoor scenario. Predicted trajectory elements are colored in yellow, observed trajectory patches are colored in blue, and the trajectory ground truth is colored in red.

the hierarchical framework to predict the next states, and our accuracy is given by the number of correct predictions of both the motion path and the final destination along the entire trajectory path. As explained in Section 4.4, for both scenarios, Layer1 predicted motion state ranges from 2 to 5 seconds of data, while Layer2 predicted motion state ranges from 5 to 10 seconds of data. Finally, following the evaluation procedure described in [10], top n predictions (the ground truth is within the top $n = 2$ predicted destination labels with the highest probability) are reported in Table 4.3.

Table 4.3: Destination prediction results. We report the accuracy of the top 2 predictions, e.g. the ground truth is within the top $n = 2$ predicted destination labels with the highest probability.

Methods	Top 1	Top 2
Layer2 inference (our method)	**77%**	**80%**
Layer1 inference (our method)	**74%**	**79%**
Spatio-Temporal Perspective [21]	62%	78%
Behavior-CNN [10]	53%	72%
EMM [3]	48%	69%

Both our hierarchical layers outperform the existing approaches, being able to increase the prediction accuracy (accounting for the Top 1 prediction label).

Our results outperform the best two approaches listed in Table 4.3 (see explanation of these methods in Chapter 2). Behavior-CNN [10] as well as the Spatio-Temporal Perspective [21] jointly model the dynamics of individual trajectories as well as crowd movements for predicting the future path as well as destination labels. However, the distribution of motion patterns and their correlation with the source/destination labels are not considered. The method Energy Map Model(EMM) [3], takes into account the influ-

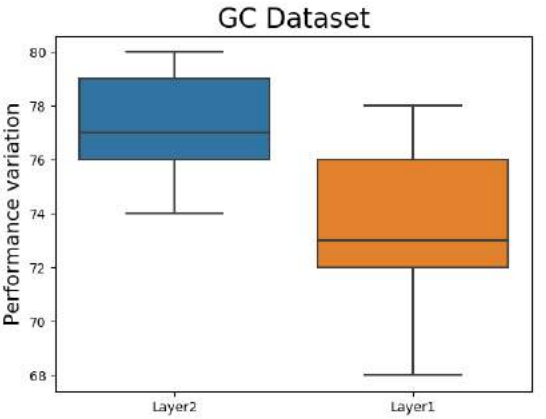


Figure 4.15: The accuracy distributions of Layer1 and Layer2 methods obtained using the different experimental folds in the destination prediction experiment.

ence of stationary objects (e.g. crowd groups). Although stationary objects can modify the path of a person, we believe that they do not influence the choice of destinations. Instead, our approach reveals that after a few steps in the scene, most of the pedestrians know their destination and modify their motion patterns accordingly. Finally, in Fig. 4.15, we investigate the performance range/variation obtained in the experimental folds. The accuracy distributions of the two best methods (i.e. Layer1 and Layer2) are displayed in box plots. Results show that Layer1 inferences (standard deviation = 3.6) present a larger variation in performance compared to Layer2 (standard deviation = 2.1). As Layer1 embeds shorter trajectory patches, the prediction of long-term destinations is more challenging, hence the greater variation.

The obtained higher accuracy for Layer2 shows the efficiency of our hierarchical model, which benefits from a powerful new feature representation learned in a hierarchical structure. By augmenting progressively the encoded spatio-temporal features, representations that are more invariant to small path shifts/deviations are learned, making the distribution of motion patterns towards a certain destination more distinctive.

4.5.4. DESTINATION PREDICTION: QUALITATIVE RESULTS

Examples of destination prediction are displayed in Fig. 4.14, where 10 Source / Destination blobs are added to the image. Trajectory ground-truth is colored in red, the observed trajectory units at time t and $t+1$ are colored in blue, our path prediction is colored in yellow and, finally, the dashed yellow line represents the predicted long-term destination. Due to the crowded scene, pedestrians take small path deviations as shown in Fig. 4.14a. However, the final goal destination remains the same most of the time. Our hierarchical autoencoder framework is able to smooth out these small shifts, and correctly predict the final destination. On the other hand, pedestrians may change their intentions, drastically modifying direction and the goal destination. In Fig. 4.14b, our

first prediction is the destination region number 9, which is where the observed and predicted motion seems to aim, but suddenly the trajectory takes a turn towards destination number 6. The new trajectory observations at time t and $t+1$ allow the system to recover and predict the new correct destination.

4.5.5. EXPERIMENT 3: ABNORMALITY DETECTION

The proposed framework can be adopted for detecting abnormal trajectories for surveillance applications in indoor as well as outdoor scenarios. Therefore, in this section, we perform the abnormality detection experiment using the GC dataset.

Following the experimental procedure proposed by [1], we selected 500 trajectories and we asked three annotators to mark abnormal trajectories on the GC dataset. All three annotators agreed on 48 abnormal trajectories (inter-rater agreement score was 87.2%), that are used to evaluate our framework for abnormality detection. Abnormal trajectories are detected in an unsupervised way using the framework explained in Section 4.3.5. We compare our results with the abnormality detection framework proposed in Chapter 3, where the best results were obtained by the SAE(AHOT) method. As the Hierarchical Autoencoder detects abnormal trajectories in an unsupervised way, we remove the classification module from SAE(AHOT), using the output of the meanshift clustering algorithm as an abnormality detection indicator. In particular, similar to the concepts explained in Section 4.3.5, if the meanshift clustering cannot assign a given sample to any of the found kernels, we consider it as "orphan" (i.e. abnormal). The experiment is repeated five times to make sure to address the randomness of the clustering assignment, the average accuracy is reported.

Table 4.4: Abnormality detection results.

Descriptors	Precision	Recall	F1
SAE(AHOT) [Chapter 3]	44%	60%	50%
Hierarchical AE [Chapter 4]	70%	77%	73%

Results are displayed in Table 4.4. The obtained results show that the Hierarchical AE reaches higher abnormality detection accuracy than SAE(AHOT), showing higher precision as well as recall results.

Furthermore, in Fig. 4.16, we visually inspect the detected abnormal trajectories. In the GC dataset [3], not all of the annotated trajectories aim to take the shortest path to reach a certain goal. In fact, in a train station, pedestrians can arrive early to catch their train, or they could be waiting for someone, deciding to spend their time walking around without a clear destination.

4.5.6. EXPERIMENT 4: PATH PREDICTION OF DIFFERENT OBJECTS ON THE VIRAT DATASET

In this chapter, we propose a system suitable for surveillance applications in both indoor and outdoor scenarios. As in outdoor scenarios there exist several moving objects with different features and rules, it is important for a surveillance system to be able to distin-

Figure 4.16: Qualitative examples depicting the detected abnormalities in the GC dataset. Abnormalities are detected in an unsupervised way when trajectory sequence dynamics (e.g. sudden change of speed or unusual turns) are observed.

guish between classes of objects. In this experiment, we test our system using trajectories coming from different objects (i.e. pedestrian vs car) using the VIRAT dataset Release 2.0 [12], as it provides frame by frame manual annotations for car as well as pedestrian trajectories (see the dataset's description in Chapter 2). In order to compare our system with the state-of-the-art algorithms [22, 23], we focus on scene A. Scene A contains cars as well as pedestrians events in a parking lot, Fig. 4.17b.

In Section 4.2, we describe how our hierarchical autoencoder framework is able to learn discriminative motion patterns that encode speed and orientation features. Therefore, in the first experiment, we evaluate the discriminative power of the encoded feature patches in both the first as well as the second hierarchical layers for pedestrian vs. car trajectory classification. A non-linear SVM classifier [24] is trained using the output provided by our two hierarchical layers ($h^{l_1,l_2}(x)$). Classification results are reported in Table 4.5. Results show that our system is able to learn discriminative motion features in every hierarchical layer. As Layer2 results are higher than Layer1, embedding bigger spatio-temporal features in a hierarchical way helps to improve the classification results. We also compared our results with the nonrealtime features (AHOT) proposed in Chapter 3 using the same experimental settings. This result confirms that embedding bigger spatio-temporal features in a hierarchical way is beneficial for motion understanding.

Table 4.5: Pedestrian vs. cars F1 classification on VIRAT dataset

Methods	Pedestrians	Cars
AHOT [25]	83%	66%
Layer1 (ours)	88%	61%
Layer2 (ours)	**91%**	**70%**

Fig. 4.17a shows examples of correctly classified objects where our feature representation and learning framework capture the different motion patterns of the two objects,

in which cars have faster and more linear trajectories (blue color) than pedestrians (red color).
Given the classification results, the second experiment consists in predicting the path of the selected object trajectories. Following [22], we use two metrics for evaluating our trajectory prediction. First, we compute the modified Hausdorff distance (MHD) between the ground-truth trajectory patch u and the predicted patch u^* as:

$$\begin{aligned} MHD(u,u^*) &= \max(d(u,u^*), d(u^*,u)), \\ d(u,u^*) &= \frac{1}{|u|}\sum_{x\in u}\min_{x^*\in u^*}|x-x^*|. \end{aligned} \tag{4.5}$$

This metric returns the distance measured in pixels between the two trajectories. Note that this metric assumes that the output of our model is a set of trajectory points that can be compared to the "true" trajectory points. Nevertheless, the output of our model is cluster labels k, where every label represents a group of similar trajectory patches. Therefore, we set the predicted patch u^* as the trajectory patch closest to each cluster center.

Then, since our method relies on Bayesian probability, we compute the negative log-loss (NLL) between the ground-truth trajectory word at k_{t+2} and the predicted label (Eq. 4.4) as:

$$NLL(k) = -\sum_{i}^{n} k_{t+2}^{i}\log(\widehat{k_{t+2}^{i}}). \tag{4.6}$$

where n is the number of samples, k_{t+2}^{i} is the true label and $\widehat{k_{t+2}^{i}}$ is the predicted label.

Note that for this experiment, we use the results from Layer2. Since the annotations for the official test set are not provided, in this experiment, we adopt a leave-one-out strategy, in which we leave one recorded day out as test set, training our model on the others, and we repeat the process for every recorded day. As explained in Section 4.4, we encode and predict, in a recursive way, patches of 5 seconds of data. Table 4.6 shows the path prediction results as a weighted average between the two target objects (i.e. pedestrians and cars).

Table 4.6: Path prediction on the test trajectories from VIRAT dataset.

Methods	MHD	NLL
Hierarchical AE (ours)	21.67	**1.357**
ALM [22]	16.7	1.476
hMDP [23]	-	1.594

Our results in Table 4.6 show that our model is comparable with the state-of-the-art models, furthermore, NLL results demonstrate that quantizing the discovered pattern dynamics in the High-Layer Inference (Section 4.3) helps to improve the state-of-the-art. On the other hand, the MHD metric shows that our predicted trajectory patches have grater distance to the ground-truth than the comparison method (i.e. ALM [22]). This was expected, as taking the trajectory patch closest to the center of the predicted

(a) (b)

Figure 4.17: (a) Pedestrians vs Cars trajectory classification based on motion patterns. The color blue indicates that the trajectory belongs to a car and color red indicates that the trajectory belongs to a pedestrian. (b) Destination prediction on VIRAT dataset scene A. Predicted trajectory elements are colored in yellow, observed trajectory patches are colored in blue, and the trajectory ground truth is colored in red.

cluster as the most representative sample is not always the most accurate solution. On the other hand, this result indicates that the trajectory patches variance within the found clusters is low as the MHD metric is comparable with the results obtained by state-of-the-art methods.

Finally, in Fig. 4.17b, we show an example of correctly predicted trajectory, where, the ground truth trajectory is colored in red, blue depicts the observations at time t and $t+1$, and our prediction of the future trajectory state at time $t+2$ is colored in yellow.

4.6. CONCLUSIONS

In this chapter, we proposed a novel hierarchical framework for modeling trajectories, both locally, by discovering spatio-temporal patterns, as well as globally, by learning statistically meaningful combinations of low-level elements leading to higher semantic descriptors. The proposed feature representation proved to be useful at capturing spatial and temporal characteristics of trajectories, such as orientation and speed variations. Furthermore, we demonstrated its power in the discrimination between different classes of moving objects (i.e pedestrians vs. cars). The effectiveness of our approach was proven in both indoor and outdoor scenarios for short-term as well as long-term path prediction. Furthermore, the presented approach is also suitable for abnormality detection, computed using the likelihood of the conditional probabilities. Pedestrian behavior modeling is an important task for surveillance applications. However, human behavior is expressed in several ways which cannot be captured efficiently using only trajectory information. Hence, in the next chapters, we will extend our research and methods on assessing motion information that involves the entire human body.

REFERENCES

[1] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.

[2] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[3] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3488–3496, 2015.

[4] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–971, 2016.

[5] A. Bera, S. Kim, T. Randhavane, S. Pratapa, and D. Manocha, "Glmp-realtime pedestrian path prediction using global and local movement patterns," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pp. 5528–5535, IEEE, 2016.

[6] D. Xie, T. Shu, S. Todorovic, and S.-C. Zhu, "Modeling and inferring human intents and latent functional objects for trajectory prediction," *arXiv preprint arXiv:1606.07827*, 2016.

[7] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2871–2878, IEEE, 2012.

[8] A. Abrams, J. Tucek, N. Jacobs, and R. Pless, "LOST: Longterm Observation of Scenes (with Tracks)," in *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 297–304, 2012.

[9] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5308–5317, 2016.

[10] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *European Conference on Computer Vision*, pp. 263–279, Springer, 2016.

[11] C. P. Marc'Aurelio Ranzato, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," *Advances in neural information processing systems*, pp. 1137–1144, 2007.

[12] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011*, pp. 3153–3160, IEEE, 2011.

[13] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, *et al.*, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[14] W. Luo, B. Stenger, X. Zhao, and T.-K. Kim, "Automatic topic discovery for multi-object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[15] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models," *International journal of computer vision*, vol. 95, no. 3, pp. 287–312, 2011.

[16] J. Masci, U. Meier, D. Ciresan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *21th Int. Conf. on Artificial Neural Networks (ICAN'11)*, pp. 52–59, 2011.

[17] F. J. Huang, Y.-L. Boureau, Y. LeCun, *et al.*, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[18] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 5747–5756, IEEE, 2017.

[19] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–742, 2016.

[20] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[21] Y. Li, "A deep spatiotemporal perspective for understanding crowd behavior," *IEEE Transactions on multimedia*, vol. 20, no. 12, pp. 3289–3297, 2018.

[22] D. Xie, T. Shu, S. Todorovic, and S.-C. Zhu, "Learning and inferring "dark matter" and predicting human intents and trajectories in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 7, pp. 1639–1652, 2018.

[23] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *European Conference on Computer Vision*, pp. 201–214, Springer, 2012.

[24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[25] D. Dotti, M. Popa, and S. Asteriadis, "Unsupervised discovery of normal and abnormal activity patterns in indoor and outdoor environments.," in *VISIGRAPP (5: VISAPP)*, pp. 210–217, 2017.

5

BEHAVIOR AND PERSONALITY ANALYSIS IN A NONSOCIAL CONTEXT DATASET

This chapter is based on the following publications:

- D. Dotti, M. Popa, and S. Asteriadis, "Behavior and personality analysis in a nonsocial context dataset", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2354–2362, 2018
- F. Gibellini, S. Higler, J. Lucas, M. Luli, M. Stallmann, D. Dotti, and S. Asteriadis, "Towards approximating personality cues through simple daily activities", in *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 192–204, Springer, 2020.

5.1. FROM MOTION TO BEHAVIORS USING PERSONALITY BASED MODEL

In Chapter 3, we introduced an indoor dataset for studying abnormal behaviors elicited by problem-solving as well as daily activities tasks. When designing the experimental tasks, we purposefully did not include any time constraints or know-how on the way

to solve these tasks. Our goal was to let the participants express freely their strategies and solutions without any impositions. Then, in the context of the European project called ICT4Life, we analyzed the participants' trajectories aiming to detect patterns that could be connected to confusion and repetitive movement patterns, common in people affected by Alzheimer's disease. During the analysis, we also noticed that the recorded data contained much more information than trajectories. Due to the unconstrained setting of the experiment, the participants felt free to use their own strategy to complete the given tasks (see Chapter 3, Section 3.5.4 for the tasks explanation). Some participants took their time to search for the hidden items, while others felt pressured to complete the tasks as soon as possible. Some participants were very precise in counting and listing all the objects in the cabinets, being very meticulous in the screening of the boxes and cabinets. While other participants were not concerned to respond to the questionnaire in an inaccurate way. To summarize, we noticed that our participants expressed behavioral and interaction patterns that were worth to explore more in-depth.

Therefore, part of the research during my PhD was conducted to study how commonalities as well as differences between behaviors can be learned by computational models. In this direction, personality psychology provides several models that aim at categorizing how we, as individuals, tend to behave in ways that are broadly consistent over time. Personality psychology is a branch of psychology aiming at capturing individual characteristics that explain and predict observable behavioral differences [1]. Personality models were shown to be accurate in the prediction of behavioral tendencies as well as life outcomes. Hence, we believe that the automatic prediction of personality can be also beneficial to the computing community, for example, in applications involving human behavior understanding and forecasting.

In this chapter, we will explain how we enhanced the multimodal dataset explained in Chapter 3, Section 3.5.4, using personality scores collected following a first-person (self-assessment) strategy. Participants were asked to fill in, in an anonymous way, the short version of the Big Five Inventory (BFI-10), introduced in [2]. This short version measures the Big-Five traits, namely, Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness, using 10 questions (i.e. two questions per personality trait) and can be completed in less than one minute.

In this study as well as the next ones, we will deepen our understating on the relation between human body postures and expressions and behavioral patterns. We will use the knowledge of trajectory motion as well as spatio-temporal dynamics learned in the previous chapters to investigate if there exist commonalities and motion "blueprint" between individuals, and, if these low-level motion signatures are connected to higher-level behavioral interpretations schema like personality traits. Specifically, by using spatio-temporal features and by exploiting neural networks to predict personality labels as ground truth, we will strive to reduce the gap between theoretical personality psychology models and computational models.

5.1.1. PERSONALITY RECOGNITION USING TRAJECTORY PATTERNS

Authors in [3] showed that even simple actions like walking in public environments involve several complex decisions that are handled in different ways. How pedestrians interact with the environment was shown to be connected to personality attributes such

as aggressive or conservative traits [4]. For instance, aggressive behavior is common in pedestrians who prefer to walk directly towards their destinations, despite other people moving along their path. On the other hand, conservative behavior is represented by pedestrians who prefer to take the longer way in order to avoid contact with other people. Hence, given our previous work on trajectory understanding (Chapter 3 and Chapter 4), we firstly conducted an exploratory experiment analyzing the participant trajectories in relation to their personality trait scores.

We selected the Hierarchical Autoencoder model (Chapter 4) for this exploratory experiment. In the previous chapter, we demonstrated that our model is able to learn discriminative motion patterns in an unsupervised manner using a hierarchical architecture based on autoencoders. The model utilizes a final learning layer to study temporal transitions between motion descriptors using Bayesian probability. On the other hand, in this paragraph, our goal consists in investigating the mapping between motion information and personality recognition. Hence, the Bayesian component is replaced with a learning component that allows us to embed the personality labels. Due to the impressive results shown by the LSTM networks for both movement learning as well as multiclass classification, in this experiment, we employ an LSTM network (introduced in Chapter 2). A softmax classification layer is added on top of the LSTM output with the goal of mapping trajectory motion patterns to the participants' personality scores.

In Figure 5.1, we introduce the new architecture of the Personality based Hierarchical Autoencoder (PHA) used for this experiment. To extract as well as learn short-term motion patterns, we start by building the hierarchical architecture using Layer 1 (Chapter 4, Section 4.2.2 and Section 4.2.3). As short-term motion descriptors contain little semantic information, we extract and learn longer trajectory sequences by building a second layer (Layer 2) in our hierarchical model (Chapter 4, Section 4.2.4 and Section 4.2.5). As the second layer is the highest layer in the hierarchy, we utilize its encoded trajectories output as input to the High level inference Layer.

Following the implementation details of the Hierarchical Autoencoder model, we encode trajectory patches of 2 seconds of data for Layer 1, and 5 seconds of data for Layer 2. Given the encoded output of our model at time t, we create a spatial grid formed by $g = 3 \times 3$ patches, to account for all the possible motion directions (Figure 5.1). In this way, our grid encodes the three consecutive motion patches at time t, $t+1$, and $t+2$. Finally, the spatial grid g is fed to the LSTM learning framework with the goal of mapping sequences of motion patches to personality trait scores. In this experiment, we use a 1-layer LSTM implementation with hidden layer size of 256, and a sequence-to-one input-output structure, where, t and $t+1$ are used as sequence input, and $t+2$ as prediction.

Following previous studies on personality recognition [5], for each participant, we binarize the score of each trait using the trait's median. If the score is higher than the median, we assign the value 1, otherwise we assign the value 0. This exploration experiment is evaluated using each personality trait independently. We divide the data into training and test sets (80% of the data for training and 20% for testing) using a 5-fold cross validation approach. In order to investigate the temporal power of our motion descriptor sequences, we compare the LSTM recognition performance to a learning model that does not include any temporal information. We chose a standard Random Forest

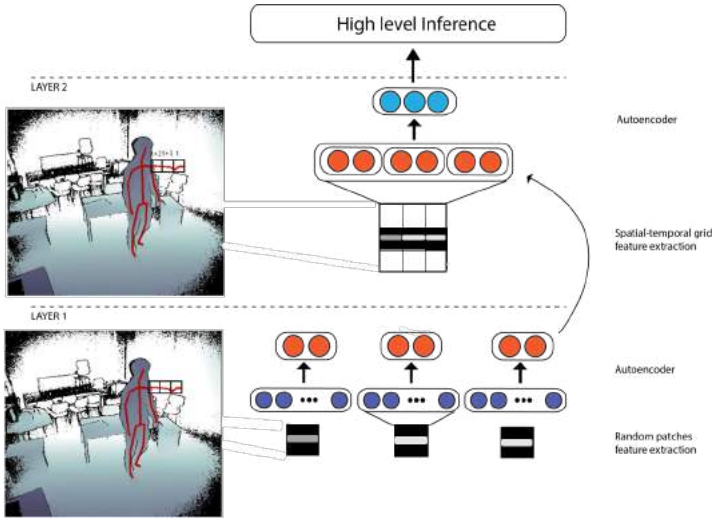


Figure 5.1: Personality-Based Hierarchical Autoencoder (PHA) architecture used for personality prediction. We employ the model introduced in Chapter 4, with the addition of an LSTM network as the inference component because of its ability in capturing temporal dependencies.

classifier as baseline comparison. Furthermore, as baseline, we added a dummy classifier that generates predictions uniformly at random.

Table 5.1: F1 accuracy of traits scores, E= Extraversion, A= Agreeableness, C= Conscientiousness, N= Neuroticism, OE= Open to Experience

Hierarchical layers	E	A	C	N	OE
Dummy Classifier	50.2%	46.1%	49.4%	51.7%	50.1%
RFC	53.4%	45.8%	50%	55.6%	55.1%
LSTM	**58.2%**	**51.3%**	**51.5%**	**59.8%**	**57.1%**

F1 recognition accuracy is reported in Table 5.1. Both classification methods report results significantly above chance (p-values less than 0.05) for Extraversion, Neuroticism and Openness traits. Note that the chance level is represented by the dummy classifier's scores. The results show that there is a link between motion patterns and the Extraversion, Neuroticism, and Openness traits. Moreover, these results indicate that the information extracted from participants' trajectories such as speed and orientation are connected to some personality attributes. On the other hand, results indicate that, trajectory information learned by the proposed PHA are not enough to recognize Agreeableness as well as Conscientiousness traits. A possible explanation may be that these traits cannot be fully captured by features that do not consider the contextual information. For example, in order to infer patterns of the Agreeableness trait, the context should be such that interactions among individuals are supported, probably in collaborative settings. Finally, the LSTM network outperforms the RFC classifier, demonstrating that the tem-

poral dynamics are critical when studying behavioral patterns.

This exploratory experiment showed us that there exists a connection between motion behavior and personality traits. However, as the proposed model (PHA) was not designed to encode the complex behavioral patterns present in the data, the model was unable to grasp in-depth behavioral patterns connected to personality. Hence, in this chapter, we will explore novel descriptors encoding nonverbal behavioral cues such as body postures and body gestures. Moreover, as the LSTM network showed promising results, we will further investigate its temporal power to build a framework designed to encode behavioral patterns connected to personality attributes.

5.2. INTRODUCTION TO PERSONALITY RECOGNITION USING BEHAVIORAL CUES

Personality recognition using behavioral observations is a challenging task due to psychological, as well as technical modeling reasons [6]. First of all, underlying human mechanisms for emotion and personality understanding are still mostly obscure to the psychology community. Additionally, human judgment regarding personality evaluation of others is often too unstable, due to many possible interpretations of human expressive power [7]. Research in Affective Computing has shown big improvements over the last years, where verbal and nonverbal behavioral cues have been studied for a variety of applications, such as Human-Computer Interaction (HCI) and Ambient Assisted Living (AAL) [1].

Modeling human behavioral cues requires a deep understanding of several components like facial expressions, gaze, hand gestures, body postures, and conversation dynamics. Facial expressions have been shown to provide reliable behavioral and affective information [8]. However, the detection of faces in unconstrained environments remains challenging when face sizes are small [9] (i.e. when surveillance cameras are placed at a distance from the recorded scene), and, furthermore, data of people's faces may raise privacy concerns. On the other hand, nonverbal behaviors of human bodily cues have been shown to be robust, as well as an important predictor for personality [10].

In light of the fact that individuals' interactions with others are shaped by their personality [6], nonverbal behavioral cues have been widely studied in social situations. For example, as shown in [11], extrovert individuals tend to engage in more face-to-face positions during conversations, and shy individuals tend to avoid walking too close to their neighbours. Nevertheless, psychological research showed that components such as social pressure, may affect natural personality displays in social contexts [12]. Furthermore, for applications in AAL where a considerable number of people are living alone, there is a concrete need for systems able to understand behaviors and personality in nonsocial contexts. For example, authors in [13] highlighted the need of real-time detection of affective states in HCI to reduce users' frustration during the interaction with machines. Imagine a smart home environment where all the sensors are tuned to enhance the user experience. If the system understands that a person has a neurotic personality, it will modify all the settings in the house in order to avoid as much as possible the tension to arise.

In this direction, personality computing can enhance the understanding of subtle

human behaviors to make the interaction as natural as possible. Hence, in this chapter, we propose a nonverbal behavior analysis based on skeleton motion features using Histograms of Oriented Tracklets (HOT) [14], spatial heat-maps, as well as body posture features extracted in an unsupervised way using Autoencoders [15]. Moreover, as behaviors have a dynamic nature, temporal sequences are investigated in an LSTM framework, for personality recognition (Fig. 5.2).

Additionally, the dataset presented in the previous section is extended with an additional 27 participants, reaching a total of 46 participants [1]. Personality labels were collected using the BFI-10 personality questionnaire, which is the shorter version of the Big Five Inventory (BFI). To the best of our knowledge, this is the first dataset that provides sensory data, skeleton tracking information and self-assessed personality labels [2], for behavior and personality modeling in an unconstrained indoor environment.

The contributions of this chapter are as follows: Firstly, using the BFI-10 personality questionnaire, we collect the personality scores of all the subjects recorded in the Multimodal dataset (Chapter 3). Then, in contrast with the dimensional approach widely present in the literature where traits are considered independently within individuals, we follow a data-driven approach to investigate new personality labels that represent the configurations of all the traits within individuals. Lastly, we propose a novel framework for personality recognition that encodes spatio-temporal features as well as body postures using an LSTM-AE model. The proposed model aims to encode body posture representations in an unsupervised way using an Autoencoder module as well as temporal behavior dynamics connected to personality scores using an LSTM network.

5.3. BEHAVIOR AND PERSONALITY IN A NONSOCIAL CONTEXT DATASET

While human behavior has been largely studied in social environments [11] and crowded places [17], few efforts have been made towards the relation between human behavior and personality in nonsocial situations, e.g. performing individual tasks. In order to provide a new benchmark to further study the relation between human behavior and personality recognition, in this study, we released the Behavior and Personality in a nonsocial context Dataset. [2]

5.3.1. THE NONOSOCIAL DATASET: AN EXTENSION OF THE INDOOR MOTION DATASET

The multimodal dataset (Chapter 3) was recorded filming 19 participants performing 6 tasks in an unconstrained indoor environment. As the exploratory experiment described in Section 5.1.1 showed promising results when mapping motion patterns and personality traits scores, we decided to extend the dataset with more participants.

The experimental room and the experimental design was kept the same to ensure data compatibility between the two sessions. In the second session, 27 participants were

[1] The extension of this dataset was part of a project for master students under our supervision. The project resulted in a publication titled: "Towards Approximating Personality Cues Through Simple Daily Activities" [16]

[2] Dataset is available at `https://project.dke.maastrichtuniversity.nl/personality/`

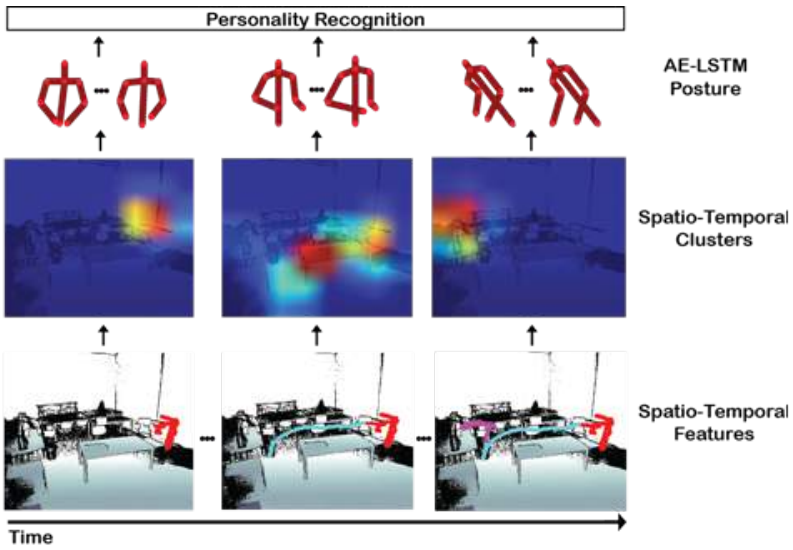


Figure 5.2: Architecture of the proposed system. First, spatio-temporal clusters are derived from skeleton motion features and spatial heat-maps. Second, from every cluster, an LSTM network is used to map posture sequences to personality recognition.

involved to perform the same 6 tasks [3]. The data was recorded using the Kinect depth sensor V2 placed at around two meters distance from the ground. Ambient magnetic sensors, composed by two magnets and a sensor which fires when the state of the two magnets changes, were placed on two office cabinets and on the entrance door, to detect opening/closing events.

Personality labels were collected through the same procedure employed in the first recording session (Chapter 3), i.e. using the BFI-10 personality questionnaire. Even though shorter than the BFI-44 [18], the BFI-10 was shown to provide reliable scores because it was built by preserving the questions that best correlate with the results of the original inventory. In this study, we followed a first-person (self-assessment) strategy, where participants were told that all questionnaires would remain anonymous, in order to respect the privacy regulations imposed by the GDPR directives.

The data was cleaned from outliers and recording errors, and finally, the two sessions were merged to create a novel dataset containing problem-solving ADL activities. The tasks were designed in a way that only one participant had to be in the room during the experiment. In this way, the participants could not get any external help on how to solve the task, and they were "forced" to use their own strategies. Moreover, from a technical point of view, recording only one participant per task helped to maximize the skeleton detection and tracking results. As the data contained only single individual behaviours, the dataset was titled "nonsocial Dataset".

[3]The extension of this dataset was part of a project for master students under our supervision. The project resulted in a publication titled: "Towards Approximating Personality Cues Through Simple Daily Activities" [16]

5.4. AN AE-LSTM FRAMEWORK FOR PERSONALITY RECOGNITION

In this chapter, we investigate nonverbal behavioral cues for personality recognition on the introduced "nonsocial dataset". Due to the position of the camera, upper body joints are the most robust to occlusions and noise, and, therefore, in our feature extraction phase, we consider only eight joints: head, spine, left-right shoulders, elbows and wrists. For each frame, behaviors are represented in terms of joints relation (posture), motion, as well as global spatial location (heat-map of the room). As shown in Figure 5.2, first, we aim at finding areas of the scene where similar activities occur by clustering spatio-motion features. Second, in every cluster, we aim at extracting and learning meaningful upper-body sequences for personality recognition using an AE-LSTM framework.

5.4.1. SPATIO-TEMPORAL FEATURES (HEATMAPS)

In every scenario, behaviors are correlated with areas where they are performed in, for example, the activity of making tea takes place at the tea corner, whereas walking behaviors happen in the walking area. In this study, we use a data-driven approach to discover the main spatio-temporal patterns in the dataset in an unsupervised way (Fig. 5.3). For the spatial descriptor, we employ the method introduced in Chapter 3 to compute the level of occupancy in each image region. Firstly, the video scene is divided into 3D not overlapping patches, where every cube is of size $h \times w \times d$, namely height, width and depth. Secondly, the heat-map histogram descriptor SP is built by counting the occurrences of the upper-body skeleton joint coordinates inside each not overlapping patch for every time window $t_1, .., t_T$. Every time-window t contains 1 second of data. Our heat-map descriptor is important as it indicates the current location as well as the locations visited by the subject. If the subject is moving, trajectory points can be found in more than one patch, whereas if the subject is stationary, points can be found in only one patch (SP descriptor in Fig. 5.3). In addition to SP, which provides global understanding of the movements in the scene, we aim at building a motion descriptor OM which extracts the local motion information from each trajectory. Specifically, in every scene patch, an adaptation of the Histograms of Oriented Tracklets [14] is used to encode the magnitude and orientation of each upper-body joint in a time window t. For every frame i in t, we compute the magnitude and orientation of each upper body joint $J = [j_0, j_1, ..., j_n]$ with respect to the previous frame. The motion induced by the body joints is highly correlated, for example, when we walk towards a direction, the head joint as well as as the arms joints may present the same motion information. Nevertheless, we are not really using our arms nor our head for the activity, all the joints are moving because they are connected to the whole body. In this case, we could have a biased information that we are moving the arms in areas of the scene in which there are not interesting objects. Therefore, to remove this correlation, we subtract the magnitude value of the head joint (j_0) from all the other upper body joints ($j_1, .., j_{n-1}$). Our goal is to highlight discriminative movements that do correspond with an activity of daily living versus general movements like walking or sitting. The obtained motion values are quantized in $MA = 3$ bins for magnitude and $OR = 8$ bins for orientation (see Chapter 3, Eq. 3.2). Every histogram descriptor OM_j is of size $OM = OR \times MA$ (OM descriptor in Fig. 5.3).

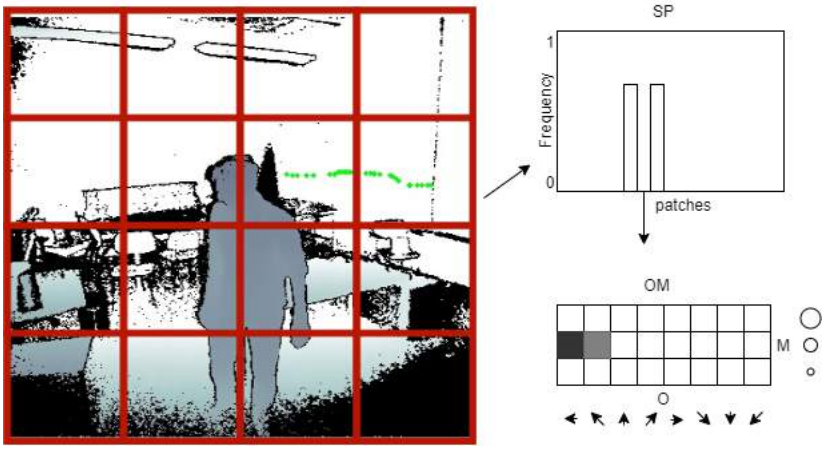


Figure 5.3: Spatio-temporal heatmap SH used to discover the main spatio-temporal patterns in the dataset in an unsupervised way. We compute the motion descriptor OM on the tracklets that populate every scene patch in SP.

Finally, the spatial and motion information are concatenated to form the final feature vector of size $SH_t = (h \times w \times d) \times OM \times J$.

5.4.2. SPATIO-TEMPORAL CLUSTERS

Using the Spatio-temporal heatmap described above, we aim at finding areas of the scene where similar activities occur in an unsupervised way. Hence, we employ the Gaussian Mixture Models (GMM) clustering technique [19] on the heatmap features SH_t to find a set of clusters that defines a clear separation between behavioral patterns. For example, searching for an object produces different spatio-temporal information than the activity of making tea, as they are happening in different regions of the scene and they are characterized by different motion magnitudes. Because GMM is a probabilistic model, we can treat the clustering task as "soft" assignment, being more robust to outliers and noisy data. In the next sections, we will explore the posture dynamic inside each discovered cluster using an AE-LSTM framework.

5.4.3. UNSUPERVISED POSTURE REPRESENTATION

In this section, we introduce a novel approach to learn upper body posture representations using autoencoders [15]. Our approach consists of two parts: posture extraction and posture learning.

Skeleton-based representations have shown promising results in encoding spatial and temporal relations among body joints for the task of action recognition. Towards this goal, various techniques have been proposed in bibliography, while the use of artificial neural networks has received significant attention [20, 21]. Therefore, we propose a new descriptor which is optimized for an autoencoder framework and it aims to learn in an unsupervised way the skeleton joints local relation through posture.

For every frame, we build a new binary image of size $s \times s$ around the upper body

skeleton data, where the pixels corresponding to the eight joints of interest are set to a value equal to one, and the rest of the patch is considered as background, with pixel values equal to zero. The image size is selected to account for all possible situations in which the joints could appear (e.g. when the arms are wide open the overall posture size is bigger than when the arms are closed). Single skeleton coordinates x, y are too sparse to be learned in an efficient way and, moreover, the pose information conveyed is limited. Hence, following their natural physical connections (i.e. left shoulder and right shoulder), related joints are connected by a line with a value equal to one (Figure 5.4).

Even though we built the descriptor considering only the x, y information, the z coordinate is not lost. In fact, as showed in Figure 5.4, the body information inside our descriptor is not normalized, nor centralized, resulting in a body representation that varies its size according to the actual distance to the camera (i.e. z coordinate). For example, the frames where the skeleton is far away from the Kinect sensor (high value of z), will contain a body posture with a smaller size than the ones in which the skeleton is closer to the sensor (Fig 5.4). The advantages of our descriptor are the following: 1) we preserve the local spatial relation between joints for posture learning, and 2) we reduce the learning problem complexity by using a binary image, where the desired information is set to a value equal to one, and the background is set to zero.

Encouraged by the impressive results of autoencoders in image reconstruction, a deep autoencoder is trained to minimize the input reconstruction error [4]. For each autoencoder layer $L = [l_1, ..., l_n]$, the encoder f^L and decoder g^L functions are designed to reconstruct the input data X, represented as a vectorized set of input features $X = [x_1, \ldots, x_n]$, as good as possible in an unsupervised way. Given our binary image as input data X, the encoding step is obtained using the encoder function f^L, while the mid-level representation is denoted by h^L and the decoding step is captured by the function g^L for obtaining the reconstructed inputted denoted as $\hat{X}$(see more details about Autoencoders in Chapter 2 Section 2.3).

The optimization goal is to minimize the error between the input data X and the reconstructed data $\hat{X}$, using stochastic gradient descent with adaptive learning rate and cross-entropy (CE) (Chapter 2 Eq. 2.5).

In Fig. 5.4, we start by visually inspecting the reconstructed images using the deep autoencoder (with $L = 2$) learned weights with $h^{l_1} = 900$ hidden units in the first layer and $h^{l2} = 225$ units in the second layer. Different postures are clearly visible, where the 2D skeleton information (x, y) embeds the spatial relation between the joints. Moreover, the proposed autoencoder shows to be robust to varying skeleton sizes, proving that our framework in able to embed the z values (i.e. depth values).

5.4.4. POSTURE DYNAMIC MODELING AND PERSONALITY RECOGNITION USING LSTM

Although the posture features described above have been shown to be important indicators for personality recognition [22], [23], the temporal nature of human behaviors also plays an important role in personality recognition [24]. For this reason, we propose to use an LSTM network to learn behavior dynamics. By adding a classification layer on top

[4] For our experiments we used NVIDIA Titan X GPUs.

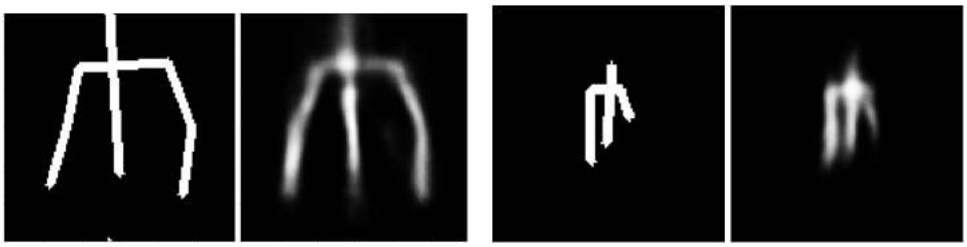


Figure 5.4: Raw posture descriptors and corresponding reconstructions. The autoencoder weights learn the 2D joints relation, while depth information is retained through the relative size of the skeleton.

of the LSTM output, recognition of the participants' personality type is performed.

LSTM networks have shown good accuracy for movement learning, however, in cluttered and noisy scenarios, a single LSTM network lacks learning capacity [17]. Hence, as described above, cluster analysis is performed on the spatio-temporal heatmap (Section 5.4.2) to reduce motion ambiguities given by the unconstrained experiment scenario. Subsequently, inspired by [25], an AE-LSTM framework is trained in each cluster independently for posture sequence learning and personality recognition (Fig. 5.2). Posture features are extracted from each generated cluster and encoded for every frame using the deep autoencoders described above (Section 5.4.3). A limitation of our method is that some clusters do not have enough data to encode posture sequences of 30 fps. For this reason, we empirically down-sampled the skeleton sequence to 8 fps, without loss of information. Finally, for each cluster c, an LSTM network is trained to map posture sequences $[x_1,..,x_t]$ of length $T = 8$ to personality type $y, y \in 1,\ldots,y_n$, where each sequence item x_i is encoded by the deep autoencoder and contains 225 features as described in Section 5.4.3. Our training objective is two-fold: firstly, we aim to learn the posture dynamics in each spatio-temporal cluster and, secondly, we aim to capture the relations between behavior display and the associated personality label. In this work, we implemented an LSTM network as in [26], that uses memory cells h^t for each time slot, with input gates i^t, forget gates f^t and output gates o^t, applied on the input node g^t, and as output, a dense layer followed by a softmax activation function for enabling a multilabel classification (see LSTM introduction in Chapter 2).

5.5. EXPERIMENT AND RESULTS

In this Section, the experiments are explained in detail. Firstly, data analysis on participants' personality scores provides the ground truth for the task of automated personality recognition making use of the features and methodology discussed above.

Secondly, personality recognition using an LSTM framework is carried out to investigate the relation between low level nonverbal behavioral features and personality display.

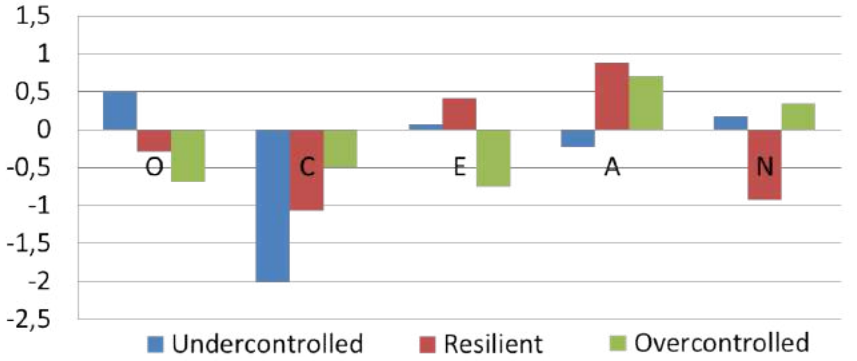


Figure 5.5: Z-score values of the personality traits in each personality types (Undercontrolled, Resilient, Overcontrolled).

5

5.5.1. PERSONALITY DATA ANALYSIS

The field of personality recognition is dominated by the dimensional approach of the Five Factor Models [1] in which the score of each trait is considered in isolation [27]. However, this approach fails in considering the traits' configurations and their dependencies within a person. In fact, combinations of traits can provide more information than interpreting one scale at a time. For example, if we have to describe an individual as being "friendly", we have a higher chance to be close to the real behavior representation if we use high scores on both the Agreeableness and Extraversion scales. By contrast, an individual with high Agreeableness but low Extraversion will tend to be perceived as docile or conformist [27].

In this work, we hypothesize that, by using configurations of personality traits, our model can learn behavioral patterns that are closer to how individuals describe their personality. In this perspective, personality types are described in terms of unique combinations of traits. There have been several propositions of personality types, e.g. the 16 personality types proposed by [28], or the types based on temperament proposed by [29]. Numerous studies confirmed the theory proposed in [30], in which all the personality traits can be organized in three major types: Resilient, Undercontrolled and Overcontrolled. Resilient personality type showed below average Neuroticism, and intermediate or above average for the rest of the traits, the Undercontrolled type usually scores high in Neuroticism and Extraversion and, finally, the Overcontrolled type scores below average on Extraversion and above average on Neuroticism.

In this study, we propose to follow a data-driven approach, applying a clustering technique on the participants personality traits scores, to find a higher convergence of traits. The results of the short BFI version [2] give the score of the five traits on a 1-10 scale, hence, every subject is represented by a vector $v = 1 \times 5$. Hierarchical clustering was applied, and the best inter-class intra-class similarity was given by $n_y = 3$ clusters.

In order to provide semantic descriptions of the discovered clusters, following the same procedure as in [27], we display the z-score of the three clusters compared to the population mean provided by [31]. In Fig. 5.5, we show the results which are consis-

tent with the theory proposed in [30], as the resilient personality type has a lower neuroticism score and a higher extraversion score than the population mean. The undercontrolled type exhibits extraversion as well as neuroticism above the population mean, and finally, the overcontrolled personality type showed a low score in extraversion and a high score in neuroticism. The resilient cluster was found to be the most populated with 21 participants, the overcontrolled cluster contains 15 participants and, finally, the undercontrolled cluster contains 10 participants. These findings provide a new way of labelling personality, and will be used as ground truth for our personality recognition experiments. The novelty of the approach is given by the fact that our model is able to learn directly different combinations of the big five traits using only three labels.

5.5.2. SPATIO-TEMPORAL CLUSTERING OF NONVERBAL BEHAVIORS

In the proposed study, we aim to map nonverbal behavioral features to personality labels for personality recognition. The proposed benchmark is very challenging due to its unconstrained structure, where participants could adopt any strategy to complete the 6 tasks (introduced in Chapter 3). Furthermore, analyzing the task independently would considerably reduce the training data making the clustering results and the LSTM prediction too task-specific. Therefore, the six tasks were organized in three semantic activity types called: "searching-activities" (A1), "problem-solving activities" (A2), and "daily-routine activities" (A3). The first set of tasks (1- Look for an item in the room, and 2- Look for an item hidden in the room) constitutes the "searching-activities", they have a medium problem solving difficulty, as the participants are required to search for predefined objects hidden somewhere in the room. The second set of tasks constitutes the "problem-solving activities" (3- search for an item that was not in the room, and 4- memorize the content in all drawers of the two cabinets), and have a high problem solving difficulty, as the participants need to firstly search for objects that are not in the room, and then they have to try to remember the content of each cabinet in the scene. Finally, the last set of tasks constitutes the "daily-routine activities" (5- sit at the table to complete two questionnaires, and 6- make tea) have a low problem solving difficulty, as the participants are asked to perform simple tasks, like making tea.

For each activity type, we aim to obtain spatio-temporal clusters representing different behavioral patterns. Spatio-temporal descriptors (Section 5.4.1) are extracted at an interval of $t_\tau = 1$ seconds and GMM clustering is applied to find K clusters. The optimization of the intra-clusters variance with respect to the total variance is used to select the correct number of clusters. For the searching-activity task, $K = 17$ clusters were found, for the problem-solving activity task $K = 19$ and, finally, for the daily-routine activity tasks, $K = 16$ clusters were found.

In Figures 5.6 - 5.8, we show the top three most populated clusters in each activity type. To visualize the content of the clusters, we average the values of the samples in each cluster. Pixel values range from dark blue, which shows low activity, to dark red, which shows high activity. For displaying purposes, in these figures, we deconstruct the feature vector SH back into the spatial features SP and the skeleton temporal features OM separately. In the first row, the spatial heat-map information SP is displayed. The second row shows the skeleton motion information OM in the respective clusters. For displaying purposes, we show only the $M = 3$ magnitude bins on the y-axis, for each of

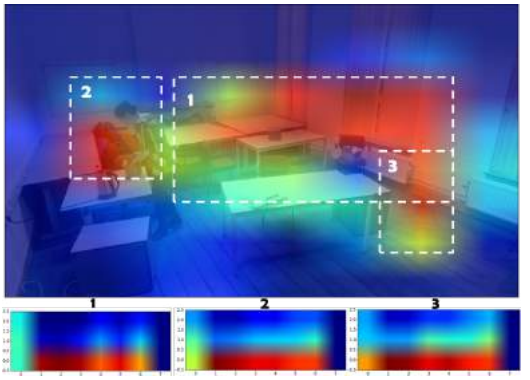


Figure 5.6: Top three most populated clusters on the **searching-activity** tasks. The first row shows the heat-map of the spatial information, while the second row shows the mean values of the skeleton motion information in the respective clusters.

the $j = 8$ upper-body joints on the x-axis. The body joints are plotted in the following order: head, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist and spine.

In the searching-activity tasks shown in Fig. 5.6, participants walked around looking for items, covering many parts of the scene. This behavior is reflected in the spatial as well as motion information. The spatial information (top row) from the three clusters presents high values covering almost the entire room. Likewise, the motion information (bottom row) reports high motion from all the joints. Additionally, the motion information is very informative for defining meaningful sub-behaviors of searching activities. For example, the first cluster contains high movement of the head (joint number zero), but slow movement of the other joints, which characterize a walking behavior. On the contrary, in the second and third cluster, motion cues from the other joints can be noticed, showing that the participants were exploring the content of the scene in order to complete the tasks.

In the problem-solving activities (Fig. 5.7), walking patterns are more focused on specific areas of the scene. As participants were challenged to fulfill the tasks with a high problem solving difficulty, they needed to focus more when searching for the hidden object or when studying the content of the cabinet. Hence, the spatio-motion clusters depicted in the picture have high peaks in specific areas, like the cabinet or the table, where participants had to perform the activity. As the participants had to move from one area to the other, we still have medium-valued spatial information (azure colors) in several spatial locations.

Finally, the daily-routine activities (Fig. 5.8) are characterized by a different spatio-temporal salience map. Participants were mainly concentrated in the areas of the table (filling questionnaire task) and the tea corner (making tea activity). In the first cluster, movements of the head and arms (elbow and wrist joints) indicate that participants were performing the activity of making tea, whereas in the third cluster, where participants were sitting at the table, the reported joint movement is little. Our cluster analysis provided spatio-temporal separations between behavioral patterns in an unsupervised

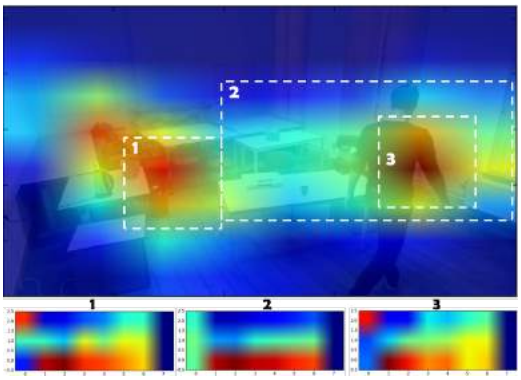


Figure 5.7: Top three most populated clusters on the **problem-solving** tasks. The first row shows the heat-map of the spatial information, while the second row shows the mean values of the skeleton motion information in the respective clusters.

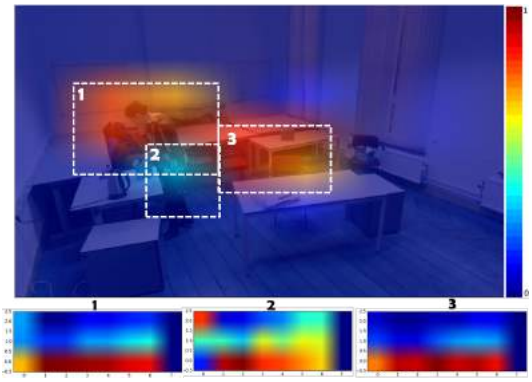


Figure 5.8: Top three most populated clusters on the **Daily-activity** tasks. The first row shows the heat-map of the spatial information, while the second row shows the mean values of the skeleton motion information in the respective clusters.

manner. Hence, for each cluster, an LSTM network is trained using the posture features introduced in Section 5.4.4.

5.5.3. PERSONALITY RECOGNITION

In this section, we test the strength of our proposed framework on personality recognition experiments. In particular, our goal is to compare three methods: 1) The proposed method which exploits the combination of LSTM networks and clustering. Specifically, an LSTM is trained in each spatio-temporal cluster found in the activity types (PR_{CL}) (Section 5.4). 2) A standard LSTM is trained without clusters in each activity type (PR_{AT}). 3) Personality based Hierarchical Autoencoder (PHA) introduced in Section 5.1.1 and based on the work described in Chapter 4 which combines trajectory patches and an LSTM network.

We used two testing procedures: K-fold cross-validation (K-fold) and Leave-One-Out

LSTMs	**A1**		**A2**		**A3**	
	f1	CE	f1	CE	f1	CE
PR_{CL}	**0.615**	**1.298**	**0.6404**	**1.287**	**0.7395**	**0.8765**
PR_{AT}	0.5263	1.697	0.5872	1.425	0.7269	0.9342
PHA(sec 5.1.1)	0.506	1.823	0.529	1.623	0.544	1.58

Table 5.2: F1 accuracy and cross-entropy error (CE) for personality recognition using the K-Fold cross-validation testing procedure. The activity types are: A1=searching activities, A2=problem-solving activities and A3=daily-routine activities.

(LOO) cross-validation. The K-fold procedure is carried-out by using 80% of the data across participants for training and 20% for testing for 5 times. The LOO procedure is carried out by using each participant as test set, training the model on all the other participants. The LOO procedure is repeated until all the participants have been used as test set. Note that the reported results are the average of all these repetitions.

In Table 5.2, we report the classification F1 score as well as the cross-entropy error (CE) of the K-Fold cross validation experiment. Note that, due to the fact that LSTMs are trained using a cross-entropy loss function, we report *CE* reflecting the difference between the predicted label $\hat{y}$ and the actual true label y distribution (Chapter 2, Eq. 2.5). Results highlight that even a general LSTM framework can successfully map the proposed nonverbal behavioral features to personality display. Furthermore, it is evident that separating behavioral patterns using a clustering technique, reduces the ambiguity of the posture sequences and improves the recognition within each activity type. The best accuracy result is obtained in the activities within the daily-routine type, where participants were asked to perform daily activities, experiencing less pressure and a low level of challenge. This set-up allowed them to create more relaxed and smooth movements, that were better captured by our autoencoder-LSTM framework.

To verify that our model does not overfit over participants' data, our second evaluation is carried-out using a LOO procedure. To overcome the imbalanced personality labels explained in Section 5.5.1, we randomly under-sample the majority classes, obtaining a dataset with ten participants per class. Therefore, for every participant p in the dataset, we classify all the corresponding posture sequences and we report the mean recall accuracy score, as well as the CE in Table 5.3. The recall accuracy was chosen in this experiment, because every sample of the test participant has a fixed personality label, making the precision accuracy always equal to 1, and therefore biasing the f1 score. The obtained results are in line with the system performance showed in the previous experiment where the proposed PR_{cl} was the most robust model. Also with this experimental procedure, the best accuracy was obtained for the daily-routine activities.

Finally, to further investigate the personality recognition performance of our system, in Fig. 5.9, we show the confusion matrices obtained in the LOO experiment procedure for all the activity types (i.e. A1= searching activities, A2= problem-solving activities and A3= daily-routine activities). Overall, the Resilient (R) and Overcontrolled classes (O) obtained the best results, showing that our LSTM framework could learn to distinguish their different configurations of personality traits. In this sense, Resilient and Overcon-

LSTMs	A1		A2		A3	
	Recall	CE	Recall	CE	Recall	CE
PR_{CL}	**0.5148**	**1.455**	**0.5333**	**1.403**	**0.6116**	**1.3897**
PR_{AT}	0.4745	1.852	0.5	1.4918	0.5134	1.5109
PHA(sec 5.1.1)	0.4533	1.972	0.49	1.9265	0.5122	1.5153

Table 5.3: Mean Recall accuracy and cross-entropy error (CE) for personality recognition using the LOO testing procedure in the three activity types. The activity types are: A1= searching activities, A2= problem-solving activities and A3= daily-routine activities.

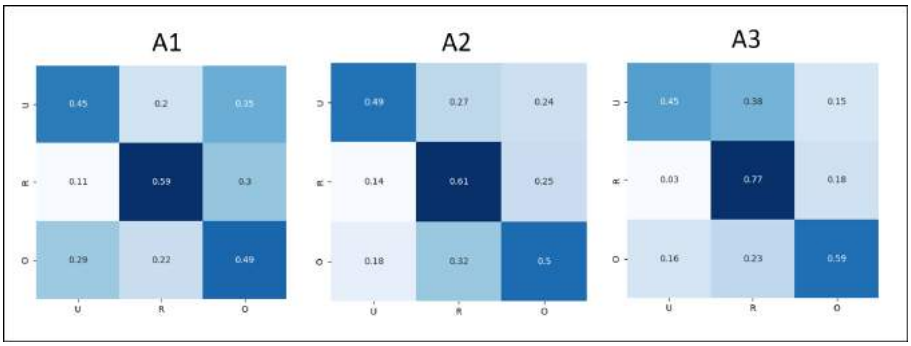


Figure 5.9: Confusion matrices obtained in the LOO experiment procedure for the three activity types. Labels are: U= undercontrolled, R= resilient, O= overcontrolled.

trolled classes contain significant differences in the Extraversion and Neuroticism traits, suggesting that these traits are the most relevant to our experiment.

5.5.4. PERSONALITY VISUALIZATION

To understand which discriminative patterns our LSTM network learns for each personality type, we display in Fig. 5.10 the part of sequences that obtained the highest confidence during the recognition phase. In particular, given our sequences of length $T = [t_1, t_2, ..t_8]$, for visualization purposes, we select the first, the middle, and the last frame, namely t_1, t_4, and t_8.

We can observe that the Resilient personality (second row), has a more relaxed and expressive posture than the Undercontrolled (first row) and Overcontrolled (third row) personalities. Given its traits configuration (i.e. low Neuroticism and high Extraversion), individual falling in the Resilient personality group may show more relaxed and talkative attributes than the other groups.

On the other hand, the Undercontrolled (first row) as well as the Overcontrolled (third row) personalities, present sequences with stiffer postures. More specifically, these postures indicate that the individuals were busy searching for objects. In a data driven way, the model have learnt to associate these types of behaviors to the Undercontrolled and Overcontrolled personality types. Hence, we can assume that these types of personalities showed searching and inspecting behaviors more often than the participants

with Resilient personality. These nonverbal behaviors may be interpreted as being focused/stressed to complete the tasks, with no interests in social contact with the experimenter, supporting its traits configuration (i.e. high Neuroticism and low Extraversion).

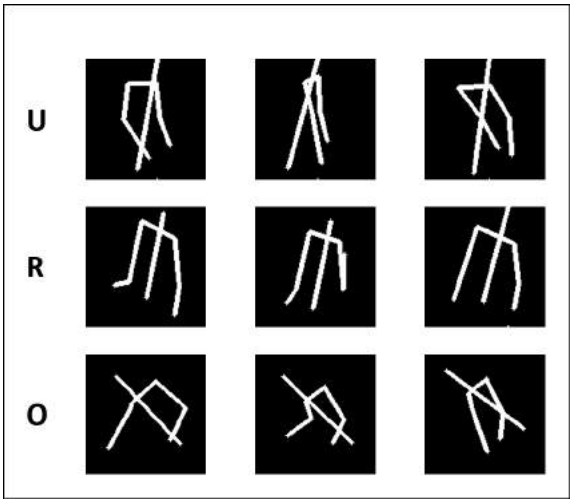


Figure 5.10: Visualization of the sequences that obtained the highest recognition confidence for each personality type. The labels are: u:undercontrolled, r:resilient, o:overcontrolled.

5.6. CONCLUSIONS

In this chapter, the extension of the multimodal dataset from Chapter 3 was introduced, and, by collecting personality traits information of the participants, we proposed a novel dataset for behavior understanding and personality recognition. Forty-six participants performed six tasks belonging to three daily activity types: searching, problem-solving and daily-routine activities. To the best of the authors' knowledge, this is the first dataset that provides depth data, skeleton tracking information for individual behavior analysis and personality labels. In the first section of this chapter, we performed an exploratory experiment using an adaptation of the Hierarchical Autoencoder introduced in Chapter 4. The results indicated that a link exists between personality attributes and participants' trajectory cues. As the motion information was not enough to fully capture the relation between behavioral patterns and personality, in this chapter, we proposed a novel framework based on body representations and LSTMs, in order to take advantage of the temporal nature of the data. Body posture information was extracted from a novel binary image feature and encoded using a deep Autoencoder. The encoded output was fed to the LSTM networks to take advantage of the temporal nature of the data. Furthermore, in this chapter, we did not consider the personality traits as independent dimensions, but we studied the distributions of the traits within an individual. In particular, following a data driven strategy, we apply a clustering technique on the participants personality traits scores, to find a higher convergence of traits (called personality types).

The effectiveness of the proposed framework and the validity of the dataset were

demonstrated by two personality recognition experiments, providing interesting insights regarding the relation between nonverbal behavioral cues and personality attributes. In the next chapter, human motion analysis is fused with surrounding information for studying the relation between human behaviors, contexts, and personality attributes.

REFERENCES

[1] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[2] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[3] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3488–3496, 2015.

[4] A. Bera, T. Randhavane, and D. Manocha, "Aggressive, tense, or shy? identifying personality traits from crowd videos," in *Proc. of the Twenty-Sixth Int. Joint Conf. on Artificial Intelligence, IJCAI-17*, pp. 112–118, 2017.

[5] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proc. of the 1st ACM Int. Workshop on Multimodal Pervasive Video Analysis*, pp. 37–42, ACM, 2010.

[6] O. Celiktutan, E. Sariyanidi, and H. Gunes, "Computational analysis of affect, personality, and engagement in human–robot interactions," in *Computer Vision for Assistive Healthcare*, pp. 283–318, Elsevier, 2018.

[7] H. Gunes, C. Shan, S. Chen, and Y. Tian, "Bodily expression for automatic affect recognition," *Emotion recognition: A pattern analysis approach*, pp. 343–377, 2015.

[8] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *European Conf. on Computer Vision*, pp. 400–418, Springer, 2016.

[9] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Systems with Applications*, p. 114602, 2021.

[10] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.," *Psychological bulletin*, vol. 111, no. 2, p. 256, 1992.

[11] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1707–1720, 2016.

[12] P. M. Cole, "Children's spontaneous control of facial expression," *Child development*, pp. 1309–1321, 1986.

[13] R. W. Picard, "Affective computing for hci," in *HCI (1)*, pp. 829–833, Citeseer, 1999.

[14] H. Mousavi, M. Nabi, H. K. Galoogahi, A. Perina, and V. Murino, "Abnormality detection with improved histogram of oriented tracklets," in *Int. Conf. on Image Analysis and Processing*, pp. 722–732, Springer, 2015.

[15] C. P. Marc'Aurelio Ranzato, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," *Advances in neural information processing systems*, pp. 1137–1144, 2007.

[16] F. Gibellini, S. Higler, J. Lucas, M. Luli, M. Stallmann, D. Dotti, and S. Asteriadis, "Towards approximating personality cues through simple daily activities," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 192–204, Springer, 2020.

[17] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 961–971, 2016.

[18] O. P. John, E. Donahue, and R. Kentle, "'the "big five," *Factor Taxonomy: Dimensions of Personality in the Natural Language and in Questionnaires."In Handbook of Personality: Theory and Research, ed. Lawrence A. Pervin and Oliver P. John*, pp. 66–100, 1990.

[19] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proc. of the 17th Int. Conf. on Pattern Recognition (ICPR 2004).*, vol. 2, pp. 28–31, IEEE, 2004.

[20] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pp. 579–583, IEEE, 2015.

[21] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 4570–4579, IEEE, 2017.

[22] M. Koppensteiner, "Motion cues that make an impression: Predicting perceived personality by minimal motion information," *Journal of experimental social psychology*, vol. 49, no. 6, pp. 1137–1143, 2013.

[23] F. Rahbar, S. M. Anzalone, G. Varni, E. Zibetti, S. Ivaldi, and M. Chetouani, "Predicting extraversion from non-verbal features during a face-to-face human-robot interaction," in *Int. Conf. on Social Robotics*, pp. 543–553, Springer, 2015.

[24] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–18, 2019.

[25] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *arXiv preprint arXiv:1702.05552*, 2017.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] F. A. Sava and R. I. Popa, "Personality types based on the big five model. a cluster analysis over the romanian population.," *Cognitie, Creier, Comportament/Cognition, Brain, Behavior*, vol. 15, no. 3, 2011.

[28] I. B. Myers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer, *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*, vol. 3. Consulting Psychologists Press Palo Alto, CA, 1998.

[29] H. J. Eysenck, "Biological basis of personality," *Nature*, vol. 199, no. 4898, pp. 1031–1034, 1963.

[30] J. H. Block and J. Block, "The role of ego-control and ego-resiliency in the organization of behavior," in *Development of cognition, affect, and social relations: The Minnesota Symposia on child psychology*, vol. 13, pp. 39–101, 1980.

[31] B. Rammstedt, "The 10-item big five inventory," *European Journal of Psychological Assessment*, vol. 23, no. 3, pp. 193–201, 2007.

6

Being the center of attention: A PERSON-CONTEXT CNN FRAMEWORK FOR PERSONALITY RECOGNITION

This chapter is based on the following publication:

- D. Dotti, M. Popa, and S. Asteriadis, "Being the center of attention: A person-context cnn framework for personality recognition", in *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 3, pp. 1–20, 2020.

Personality computing [1] is the field in which theories coming from the areas of machine learning and psychology merge to create computational models for personality understanding. In the previous chapter, we presented a study which aimed to map human movements to personality attributes in a novel dataset based on 6 problem-solving tasks. Results showed that there exists a link between how participants moved, and reacted to the tasks, and their personality scores. However, as we only started to scratch the tip of the "personality computing iceberg", several challenges connecting human behaviors and personality still remain unanswered. First of all, given the uncertain delineations of human behaviors in different situations, it is extremely challenging to build

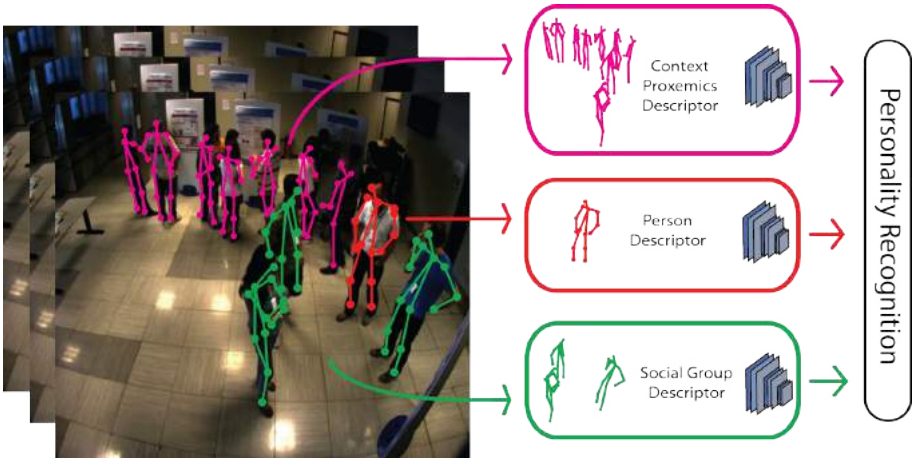


Figure 6.1: High level description of the proposed model. Individual motion descriptors as well as two Context descriptors are learned in a novel CNN framework for personality recognition. Individual motion descriptors (red color) indicate the engagement level of every person in the scene, social group descriptors (green color) indicate the engagement level of individuals in conversational groups, and finally, the context proxemics descriptors (purple color) indicate the global attitude of each individual with respect to the others.

generic systems, which would work in diverse scenarios. As a consequence, recent personality computing systems have mostly focused on scenario-specific problems (e.g. job screening interviews [2], activities of daily living [3], and work meetings [4]). Secondly, as humans are by nature social beings, personality displays have been well explored during social interaction. However, few efforts have been made towards a generic system that maps personality attributes in both social as well as nonsocial situations.

Hence, in this chapter, we focus on building a novel architecture for personality recognition in different scenarios. In computer vision, contextual information has been shown to improve several challenging tasks such as action recognition [5] and social scene understanding [6]. Building on these findings, we aim at understanding the mutual relation between behaviors that comes intrinsically from individuals (e.g. motion) and information that comes from the context (e.g. social/nonsocial situations).

Imagine a social scenario in which different interpersonal styles influence the group interaction, for example, individuals that strive to be *the center of attention* will try to actively capture the attention of the rest of the group, resulting in the group acting more passively [7]. By combining information at the individual level with information at the context level, a robust semantic understanding of the situation is perceived. Even if individuals are in a nonsocial scenario (e.g. when an individual is alone at home), it has been shown that people engage with contextual objects as they would engage with other humans (i.e. Anthropomorphism) [8]. Therefore, by examining the interaction between human behaviors and the surrounding scene, important information about human personality can be extracted.

In the field of smart monitoring systems, behavioral insights can be used in different situations depending on the final users. Institutions like the police can use advanced behavioral insights for assessing crimes or fights in public spaces [9]. Private users can

benefit from advanced behavioral features to make their systems highly personalized. In this chapter, we focus on the latter objective, aiming to advance one of the most important systems design principle: "know the user" [10], using personality computing. There are several ways in which users' information can be gathered through an interactive system; the standard way is the "active interaction", where the user is actively engaged with the system, for instance by pushing buttons or performing some actions. In the last years, with the advance of the technological power as well as sensors precision, humans started to use smart devices (e.g. mobile phone, wearables) in most of their daily activity. Hence, through passive measurements and observations of these activities, we can learn highly personalized behavioral patterns.

In this direction, we believe that automatically recognizing users' personality through "passive interactions" can improve the user's experience as well as can increase the system acceptance [11]. For example, the user does not have to actively declare his/her current affective state, mood, or personality to the system, but the system aims to understand and subsequently make actions adapted to the user's personality [12–14]. Therefore, by making automatic personality recognition systems more accurate and robust to different contexts, we are improving the system capacity to interact and adapt more like a "human" (i.e. perceiving more subtle human characteristics).

In the last few years, deep neural network (DNN) architectures achieved reliable results in the field of Personality Computing, using video [3], audio [15], as well as multimodal data [16, 17]. However, despite the growing effort, many challenges still remain to be tackled. First of all, datasets providing personality labels are still limited, resulting in ad-hoc methods for specific contexts [3, 18], and, secondly, findings about personality are often too restricted to a research field (i.e. either psychology or computer vision). Thus, computational frameworks providing both quantitative as well as qualitative results are crucial to enable future interdisciplinary collaborations.

In this chapter, we propose a novel architecture based on Convolutional Neural Networks (CNNs) [19] for personality recognition, by studying the interaction between Person and Context information in a general manner, both in a social and a nonsocial context. In Fig. 6.1, we show the high-level description of the proposed system. Besides the Person motion Descriptor (red color), computed for every individual in the scene, context information is extracted at different levels of granularity. On the one hand, we compute the Social Group Motion Descriptor (green color) on individuals engaging in close social interactions, where, by social interactions we imply the mutual interaction between two or more people [20]. Social groups are usually the result of these social interactions, where individuals stand near each other to discuss about some topics. With this descriptor, we aim at encoding the motion interactions within the same social group. Moreover, in most situations, a social scene is associated with several groups of people interacting with each other. Therefore, in order to encode the group-to-group interaction, we compute the Context Proxemics Descriptor (purple color), where we consider the interpersonal distance among all the individuals in the scene. This information is useful to understand the relation between individuals in the scene, regardless whether or not they interact in social groups. Additionally, in this chapter, we show that the proposed Context Proxemics descriptor is general enough to be applied also in a nonso-

cial scenario. As shown by several researches [21, 22], personal distances as well as personal spatial zones preferred by humans when interacting with an inanimate object (e.g. robot) would be comparable to those preferred when humans interact socially with each other. Hence, when individuals are in a nonsocial scenario, distances between individuals and the main objects in the scene are captured.

Finally, the contributions that will be presented in this chapter are as follows:
First, we propose a novel, end-to-end, multi-stream CNN framework to analyze individual, as well as context information, for personality recognition using video data. This method is robust to different types of data (e.g. RGB or depth images), different types of camera settings (e.g. fps and resolution), as well as different scenarios (e.g. social as well as nonsocial situations).
Second, in addition to individual motion descriptors [23], that are transferred from the activity recognition field, novel CNN descriptors are proposed to encode the surrounding context in different scenarios. Transforming different sources of information (e.g. Person-Context) into CNN descriptors with the same backbone structure has the great advantage of facilitating the discovery of common latent representations. Additionally, a novel pooling strategy is added, to encode the social group interaction.
Third, to demonstrate the generalizability of our algorithm, experiments are carried out on two public datasets. Results show that our Person-Context model outperforms state-of-the-art methods. Additionally, to demonstrate the robustness of the learned personality patterns, we evaluate our framework using two different sets of personality classes as training labels (personality traits [24], as well as personality types [25]).
Fourth, by visualizing the CNN activation map for both high and low scores of each personality trait, we pose the following questions: 1) *Is the relation between Person-Context dynamics reflected in the traits?* and 2) *Do the dynamics correspond to the trait attributes defined by psychologists?* Our qualitative results provide new insights into the interaction between context and individuals' behavioral cues for behavior and personality understanding.

6.1. PERSON-CONTEXT FRAMEWORK

In this chapter, we jointly model the nonverbal behavioral cues from single individuals, with surrounding context information, for personality recognition using video data. Theoretical models such as the Laban Movement Analysis (LMA) [26] conceptualize the relation between human motion and the surrounding space. Although this model was proposed initially to describe dance movements, authors in [27] and [28] adopted certain concepts for Human to Human and Human-Robot Interactions. Moreover, while psychological studies have demonstrated the tight relation between contextual information and personality patterns [29], works considering both sources of information are quite limited in the computer vision community. Therefore, inspired by these interdisciplinary research themes, we propose an automatic personality recognition system using a person-context deep model.

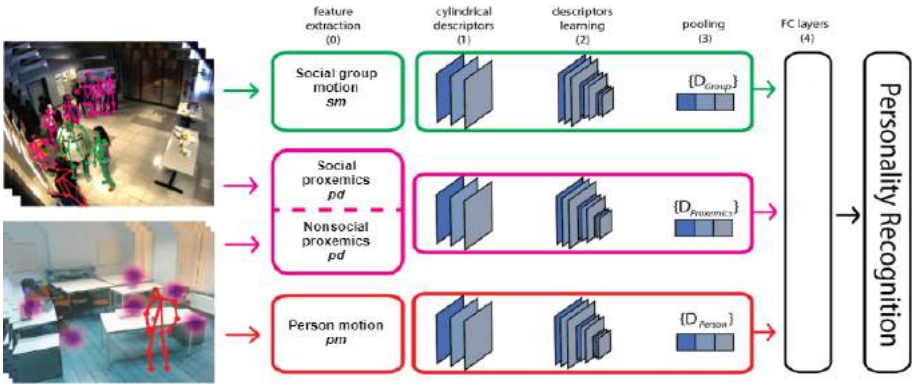


Figure 6.2: Framework architecture. The Person-Context streams (pm, sm, pd) are extracted from video sequences (0) and processed separately (1, 2, 3). The output descriptors are fused in FC layers (4) and finally, a softmax layer is employed for personality recognition.

6.1.1. ARCHITECTURE

The proposed architecture is shown in Fig. 6.2. Given a set of frame sequences $t_1, t_2, .., t_T$, we first run a pose estimation algorithm [30], to detect the individuals in the scene (Fig. 6.2). As the algorithm proposed in [30] does not track the detected skeletons, a frame by frame tracking algorithm is added using the OpenCv library [31]. The goal of the frame by frame tracker is to identify which skeleton in one frame corresponds to the same skeleton in the next frame. Given the detected skeletons $p_n(n \in [1, .., N])$ where N is the total number of subjects in a social sc^s or nonsocial scene sc^{ns}, we aim to map Person as well as Context information to personality labels. One disadvantage of jointly modeling different sources of data is that their intrinsic structure is too diverse, making the learning process more complex. Differently, one key aspect of the proposed method is to use the same backbone structure (i.e. cylindrical coordinates descriptors, Fig. 6.2 (1)), to describe the various types of information. In this way, different feature modalities are mapped onto the same feature space, allowing the creation of the same network structure (Fig. 6.2 (2) and Fig. 6.2 (3)). Note, the extracted components are general enough to deal with nonsocial scenes sc^{ns}, where, the Social Group Interaction (Fig. 6.2 (b)) will be set to 0, and the Context Proxemics Descriptor (Fig. 6.2 (sc^{ns})) will encode interactions with semantic objects in the scene. Finally, Fully-Connected Layers (FC) are adopted to fuse the features extracted by the distinct streams, and a Softmax layer is added for the personality recognition task. As follows, detailed explanation of every component is provided: the Person descriptor is described in Section 6.1.2, then, novel Context CNN descriptors are presented in Section 6.1.3 and Section 6.1.4. The training procedure is introduced in Section 6.2, and, finally, quantitative as well as qualitative results are reported in Section 6.3, Section 6.4, and Section 6.5, respectively.

6.1.2. PERSON MOTION

Nonverbal behavioral cues have been studied extensively as personality patterns descriptors [3, 32]. In this section, skeleton pose information, extracted using the algorithm

proposed in [30], is employed to create the "Person" descriptor pm (Fig. 6.3), in the proposed Person-Context framework. Inspired by [23], we create temporal "skeleton clips", to describe the relative motion of each skeleton joint $j_{1,...,J}$, where $J = 17$. In particular, for every frame sequence of size t, where t is the number of frame in 1 second of data, the relative positions of all skeleton joints are computed with respect to four reference joints $Jr_{1,...,JR}$, where $JR = 4$ (i.e left shoulder, right shoulder, left hip, and right hip). Since the joints positions are declared in the Cartesian coordinates, following [23] and [33], we transform their distances in Cylindrical coordinates (Eq. 6.1). Cylindrical coordinates are composed by three terms: ρ, θ, and z. The term ρ is defined as the euclidean distance between the reference joints Jr_{JR} and the observed joints j_J, the term θ is defined as the angle between the reference joints Jr_{JR} and the observed joints j_J, and the term z is the difference considering only the vertical axis between the reference joints Jr_{JR} and the observed joints j_J. The three descriptors $\rho_{Jr,p}, \theta_{Jr,p}, z_{Jr,p}$ are of size $t \times (J-1)$ for each reference joint Jr. As Jr_R are computed on the same temporal sequences, we decided to concatenate the information on the vertical axis. Hence, the size of our final person-motion descriptors $\rho_{pm,p}, \theta_{pm,p}, z_{pm,p}$ are of size $(t \times JR) \times (J-1)$. Finally, to make our pm descriptors suitable for a CNN architecture, we convert their value between $0-255$.

$$pm = (\rho_{Jr}, \theta_{Jr}, z_{Jr}) = (\sqrt{(x_{Jr} - x_J)^2 + (y_{Jr} - y_J)^2}, \tan^{-1}\left(\frac{y_J}{x_J}\right), y_{Jr} - y_J) \tag{6.1}$$

6

6.1.3. SOCIAL GROUP MOTION

Given a social scene sc^s, the modeling of social group interactions [34] is a challenging task for behavior understanding. Since interaction dynamics are affected by the behaviors of single individuals involved in the group, the understanding of each personality is of great importance to unfold the social group evolution. For example, an extroverted person will tend to be actively involved in the group, while an agreeable person will tend to be more passive [18]. In this section, we explain how social group motion information can be encoded using the cylindrical descriptors (Eq. 6.1).

First of all, we use the code provided by [20] to detect social groups $g_{1,...,G}$ in the scene. Every social group is composed by $p_{1,...,(M(g))}$ individuals ($M(g)$ is the total number of individuals in a social group) with his/her distinctive motion dynamics. Note that, we exclude from $M(g)$ the target person (the person whose personality will be predicted).

Hence, we utilize the same approach described in Section 6.1.2 (Eq. 6.1) to extract the person motion information on each member of the group, obtaining pm_G^M. As groups frequently vary in their number of members, we face the challenge of having different number of descriptors at different time stamps. For example, the social group g_1 at frame i is formed by $M = 4$ individuals, hence, we will encode the motion of each individual in the group, computing $pm_{g_1}^{1,..,4}$. However, at frame $i+1$, an individual left the social group g_1 to join another group. Hence, at frame $i+1$, we will compute $pm_{g_1}^{1,..,3}$ descriptors. Obviously, this solution is not optimal for CNNs which expect the input to have the same dimensions at every time stamp.

To resolve this problem, our intuition is to take advantage of our image-like descriptors, and apply a pixel-wise pooling strategy (e.g. average pooling or max pooling) on pm_G^M to downsample M descriptors to 1 descriptor. With this strategy we aim at re-

solving the practical problem of having different number of descriptors at different time stamps as well as encoding the social interaction happening within the detected social groups.

We believe that learning at the same time individual behaviors as well as social group behaviors can help us to identify more clearly personality attributes. For example, if a person has a strong personality, is likely to push his/her idea to all the members of the social group. In this situation, our descriptors could describe high motion from the individual with a strong personality and low motion from the rest of the group. Vice versa, if an individual has an agreeable personality, our descriptors could describe high motion from the individual person as well as the social group.

SOCIAL MOTION POOLING

Assume that there are $g_{1,\dots,G}$ social groups in the scene at frame i. Every social group is composed by $p_{1,\dots,(M(g))}$ individuals ($M(g)$ is the total number of individuals in a social group) with his/her distinctive motion dynamics. We consider a social group as a set of distinct individuals that interact with each other, and, therefore, affect each other's motion dynamics. Firstly, given a social group g_1, we start by encoding the behavior of every individual using Eq. 6.1 to obtain $pm^M_{g_1}$, each of size $(t \times JR) \times (J-1)$. Secondly, we aim at applying a pooling strategy on these image-like motion descriptors to encode the overall social interaction.

The pooling approach has the advantage of preserving the spatio-temporal structure of our descriptors while summarizing the social group interaction. We focus on the two most conventional pooling operations, max pooling and average pooling. Choosing the right pooling strategy is important to "downstream" the descriptors to a fixed dimension independently from the social group size, as well as encoding the group motion dynamics.

Pooling operations are performed on the three cylindrical descriptors independently ($\rho_{M(g)}, \theta_{M(g)}, z_{M(g)}$), each of size $(t \times JR) \times (J-1)$, with the goal of summarizing the joints motion (expressed as pixel value in the image-like descriptors) of the $M(g)$ individuals in the social groups. Average pooling is defined as :

$$f_{mean}(\rho_{sm,p}, \theta_{sm,p}, z_{sm,p}) = \frac{1}{M(g)} \sum_{1}^{M(g)} x_{\rho,\theta,z} \tag{6.2}$$

where, given $M(g)$ members in social groups, we compute the average of their (ρ, θ, z) pixel values. Similarly, max pooling is defined as:

$$f_{max}(\rho_{sm,p}, \theta_{sm,p}, z_{sm,p}) = \max_{M(g)} x_{\rho,\theta,z} \tag{6.3}$$

where, given $M(g)$ members in social groups, we keep the maximum (ρ, θ, z) pixel values.

Finally, the size of our social group motion descriptors $\rho_{sm,p}, \theta_{sm,p}, z_{sm,p}$ are of size $(t \times JR) \times (J-1)$ obtained by averaging/maximizing the motion values of every individual in the group.

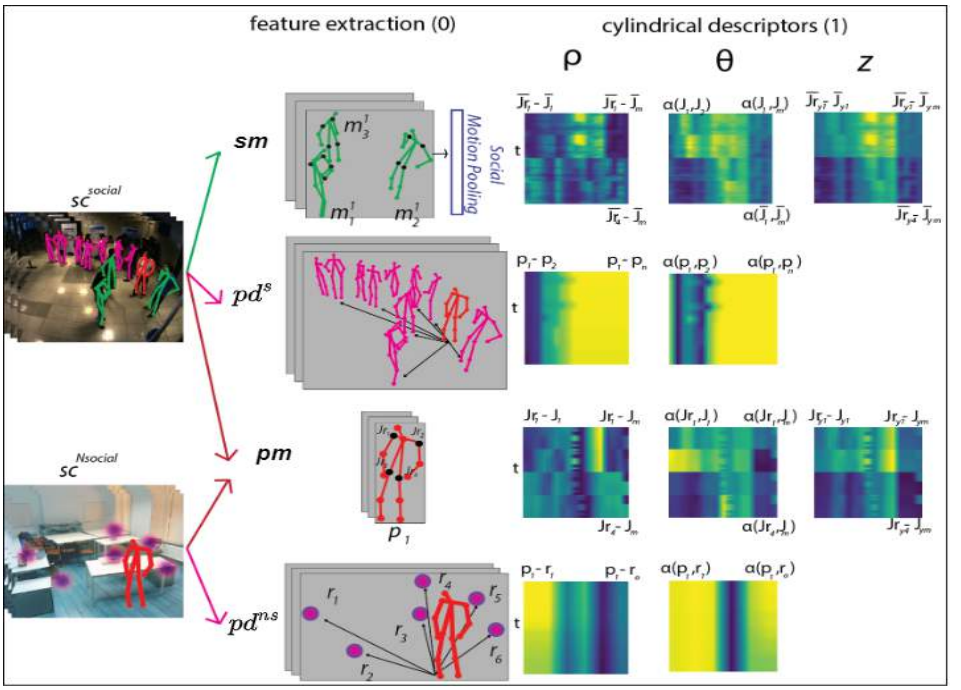


Figure 6.3: Feature extraction (step 0) and cylindrical descriptors (step 1) of the proposed framework. In step 0, the feature extraction module computes: The person motion descriptor pm, the social group motion sm, the social proxemics pd^s, and the nonsocial proxemics pd^{ns}. In step 1, these features are transformed into image-like descriptors, where they are resized and their pixel values are converted between $(0-255)$. The color blue indicates low values (low distances between skeleton joints) while the color yellow indicates high values (large distances between skeleton joints).

6.1.4. CONTEXT PROXEMICS

In the previous section, we computed descriptors encoding motion dynamics from individuals, and from social groups. In this section, we aim at computing a descriptor which encodes the interaction between individuals and their surroundings. Contextual information has been shown to yield important information about personality attributes. For example, authors in [4] estimated the level of extraversion and neuroticism of people by investigating how they used their personal space (proxemics) and their visual attention. In this chapter, we aim at encoding the interaction between the individuals and their surroundings by computing their proxemics (or interpersonal distance). In the next sections, two strategies are proposed towards encoding proxemics to different entities in different scenarios.

SOCIAL PROXEMICS

Several works highlighted the correlation between interpersonal distances (i.e. proxemics) and personality traits [4] , [20] in social scenarios. Likewise, in this section, a proxemics descriptor encoding the global position of the target person p with respect to all the other individuals in the scene N, is proposed.

In the analyzed scene sc^s, proxemics is intended as the way people use their personal space in relation to other individuals. As the above descriptors (ρ_{Jr}, θ_{Jr}) describe the distance and angle relations between body joints, we can re-use the same formulation to encode the distances between individuals.

The process is illustrated in Fig. 6.3 (pd^s step (0)), where we compute the euclidean distance (ρ descriptor) and the standing angle (θ descriptor) between the target person p^j, where and the rest of the individuals using Eq. 6.1, for a frame sequence of size t. Note that to compute the descriptors mentioned above, we consider only one skeleton joint j_2 (i.e. the neck joint).

To avoid dimensionality discrepancy due to the different number of individuals in the scene, we always consider the maximum number N of subjects for a given dataset. Note that the z descriptor is not considered, as the interpersonal distance does not occur on the vertical axis. Thus, for every individual in the scene, two descriptors of size $t \times (N-1)$ are obtained, and to be consistent with the other generated features, they are transformed into greyscale images by scaling the values between $0-255$ (Fig. 6.3 (pd^s step (1)).

This descriptor has two advantages: Firstly, the global position of every subject with respect to all the other subjects in the scene is represented, reflecting findings in the literature that correlated proxemics distance with personality displays [18]. For example, individuals who like to be engaged in group conversations (Agreeableness personality trait) will have a higher number of neighbours, while on the other hand, shy individuals will have fewer neighbours. Secondly, due to the common cylindrical coordinates structure, these descriptors can be fused with the other generated features, optimizing the learning capacity of our CNN architecture.

NONSOCIAL PROXEMICS

As humans are by nature social beings, personality displays have been well explored during social interactions. However, as reported in Chapter 5, few efforts have been made towards the understanding of personality in a nonsocial context. This problem has been shown to be important for applications in domains like Ambient Assisted Living (AAL), where it often occurs that individuals are spending most of their time at home alone. Thus, in this section, we aim to learn the relation between human behaviors and the surrounding context during nonsocial periods. We define nonsocial proxemics as the way people use their personal space in relation to objects in the scene (Fig. 6.3 (pd^{ns})).

We build on the philosophical and psychological theorem which states that people engage with objects as they engage with other humans (e.g. Anthropomorphism [8]). In this section, a novel descriptor is proposed to capture the way individuals engage with the scene. Researches from a wide array of disciplines have long noted that people tend to see nonhuman agents as human-like [8], especially in case of "lonely" situations, subjects are more likely to be subjected to anthropomorphism with nonhuman elements like objects, robots, etc [35]. Hence, our intuition is that subjects with different personality will interact differently with the objects in the scene. We describe this interaction as the distance between the target subject p and the regions containing interactive objects R. Since we aim to make a general and robust model able to adapt in any scenario, firstly, we detect the most interactive objects in the scene in an unsupervised way, and

secondly, we compute a proxemics descriptor optimized for a CNN architecture, encoding Human-Object interactions.

Human-Object interaction location discovery. We take advantage of the work presented in the previous chapter (see details in Chapter 5, Section 5.4.1 and Section 5.4.2) to build a fine-grained spatio-temporal heatmap of the scene. This heatmap will help us to detect in which areas of the scene the interaction occurs. Firstly, the video scene is divided into 3D nonoverlapping patches, where every cube is of size $h \times w \times d$, namely height, width and depth. Secondly, the heat-map histogram descriptor SP is built by counting the occurrences of the arm joint coordinates inside each cube for every time window $t_1, .., t_T$, where t contains 1 second of data.

Differently from the previous chapter, we assume that the arm motions are the most important joints to detect meaningful person-context interactions. Hence, for every frame i in t, we compute the magnitude and orientation of the arm joints $j_{1,...,6}$ with respect to the previous frame. The values are quantized in $M = 3$ bins for magnitude, and $O = 8$ bins for orientation, obtaining for every joint a final histogram of size $OM_j = O \times M$. Finally, the spatial and motion information are concatenated to form the final feature vector of size $SH_t = SP \times OM \times J$.

We employ the Gaussian Mixture Models (GMM) clustering technique [36] on the heatmap features SH_t to discover a set of clusters that defines a clear separation between interaction patterns. Seven spatio-temporal clusters were found, each of them showing different human-object interactions. However, after a visual inspection, one cluster contained spatio-temporal samples belonging to the walking behavior (i.e. no human-object interaction) and, therefore, was dropped. Finally, as we are interested in the spatial locations of these interactions, we leverage the spatial heatmap descriptor SP to select the most populated cube in each cluster.

The results are shown in Fig. 6.3 (pd^{ns}). In each of the 6 discovered spatio-temporal clusters, we keep only the location information of the most populated spatial cube. These object regions $r_{1,...,(R)}$, where $R = 6$ contain semantic information about the human-object interaction in the scene.

Nonsocial proxemics descriptors. To extract descriptors that reflect the engagement of individuals in the generated semantic object regions, we compute the distance (ρ) and the angle (θ) between subjects p_n^j and the generated regions R (Eq. 6.1) for every frame sequence of size t (Fig. 6.3 (pd^{ns} step (0)). Like in the previous section, we consider only one skeleton joint j_2 (i.e. the neck joint). Note that the z descriptor is not considered, as the computed distance does not occur on the vertical axis.

Thus, for every individual in the scene, two proxemics descriptors pd of size $t \times R$ are obtained, and to be consistent with the other generated features, they are transformed into grey-scale images by scaling the values between $0-255$ (Fig. 6.3 (pd^{ns} step (1)).

6.2. PERSON-CONTEXT INTERACTION LEARNING USING A CNN ARCHITECTURE

The integration of multiple feature representations, is a challenging task due to the inherent heterogeneity of their distributions [37]. However, CNN architectures have shown great ability in discovering common latent representations. In this chapter, a CNN frame-

work is proposed for the fusion and learning of features extracted from multiple sources: individuals' motion (Section 6.1.2), social group motion (Section 6.1.3) and context proxemics (Section 6.1.4). In order to reduce the structural differences between various distributions, a common feature representation based on cylindrical coordinates is created (Fig. 6.2(1)). This strategy allows the CNN model to use the same parameters (e.g. number of hidden units, pooling layers etc.) for all the feature representations, decreasing the duration and complexity of the training phase (Fig. 6.2(2), Fig. 6.2(3), Fig. 6.2(4)).

6.2.1. FEATURE LEARNING

In the previous sections, we introduced our Person-Context features. Every individual in the scene is described by: his/her own motion ($\rho_{pm,p}, \theta_{pm,p}, z_{pm,p}$), the motion of the social group he/she belongs to ($\rho_{sm,p}, \theta_{sm,p}, z_{sm,p}$), and his/her distance to other entities in the scene ($\rho_{sp,p}, \theta_{sp,p}$). Each descriptor is built over a frame sequence t whose size varies depending on the dataset used. For the Salsa dataset, we set $t = 15$ frames while for the Nonsocial dataset, we set $t = 30$ frames. As these descriptors are built to portray different information on different scales (e.g. motion versus proxemics distances), we resize all of them to the dimension of 68×68. Each of them is then duplicated three times to formulate a color image (i.e. $68 \times 68 \times 3$), so it can be fed into the network.

This dimensionality was chosen to preserve the mapping of the skeleton joints indexes ($J = 17$) with the rescaled descriptors. This is particularly useful, for example, to map the classification weights back to the raw joint motion values for the visualization of the CNN activations (Section 6.5).

As the proposed descriptors describe different behaviors, the VGG19 model [38] pretrained on ImageNet [39], is leveraged to extract a compact representation of each of them separately. Note that we utilize the VGG19 model only for feature extraction discarding the last 3 fully-connected layers. Early convolutional layers showed to learn more generic features, whereas, deeper layers are more influenced by the task they are trained for [23]. Since our CNN representations are very different from the images contained in ImageNet, we extract the features from the convolutional layer $conv5_1$. We justify the adoption of this layer, as authors in [23] applied it on similar image clips. Specifically, the generated cylindrical descriptors are fed as input to the VGG model. As output, from the $conv5_1$ we obtain feature representations of size $4 \times 4 \times 512$ (512 feature maps of size 4×4). Note that as our descriptors represent temporal information, we adopt the pooling strategy proposed by [23], called temporal mean pooling (TMP), ideal to embed temporal information from the input images (see details in Chapter 2 Section 2.9.1).

By using the VGG model as feature extraction step, we projected our person-context descriptors onto the same latent space. Therefore, the next step is to build a Neural Network framework to fuse this information and map them into the personality latent space. The person-context compact representations, defined as D_{pers}, D_{group}, and D_{prox} in Fig. 6.2 (3), are fused to form the input to two Fully-Connected Layers (FC), of size 1024, and a Softmax layer for the final personality recognition task (Fig. 6.2 (4)). Between the two FC layers there is a rectified linear unit (ReLU) [40] to introduce an additional non-linearity.

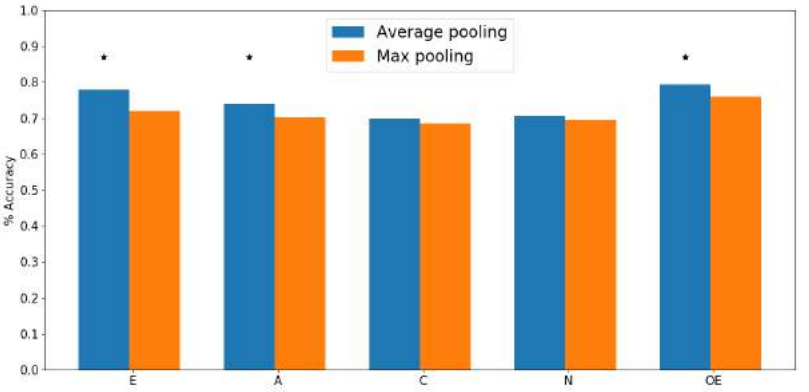


Figure 6.4: Exploratory experiment investigating the recognition accuracy using two different pooling strategies to encode the social groups' motion. The experiment is performed on the Salsa dataset poster session. The p-values less than 0.05 are summarized using one asterisk. Average pooling demonstrates better performances over all the Big-5 traits, however the results are not significant for two traits: Conscientiousness and Neuroticism.

EXPLORATIONS OF SOCIAL MOTION POOLING

An exploratory experiment is conducted to test the performance variation of the social motion pooling strategies on the Salsa dataset poster session [18]. In particular, we aim at analyzing the information carried by the social group motion descriptor in the task of personality recognition. In Fig. 6.4, accuracy results for the classification of the Big-5 personality traits, using the described pooling strategies are visualized. The results indicate that average pooling constantly reaches higher accuracy, and therefore is selected as the social group pooling strategy in the rest of the experiments. However, it is interesting to notice that the difference in performance is not statistically significant (indicated by the asterisks), when predicting two personality traits: Conscientiousness and Neuroticism.

Average motion pooling averages the motion from all the members in the social group, whereas max pooling selects the highest motion value in the group. As individuals with high Conscientiousness or Neuroticism traits are less likely to stand out or influence the social group, averaging their values with the group motion may increase their contribution in the group, and therefore, confuse the classifier decision.

FUSION

The fusion of different input cues can improve the recognition performance, as it combines different sources of information that are relevant to discover personality patterns. In our framework, fusion can be performed as early fusion (feature fusion) or late fusion (decision fusion) [41].

As feature fusion, we implemented the following methods: 1) Concatenation of the features coming from different modalities. 2) Principal Component Analysis (PCA) [42]

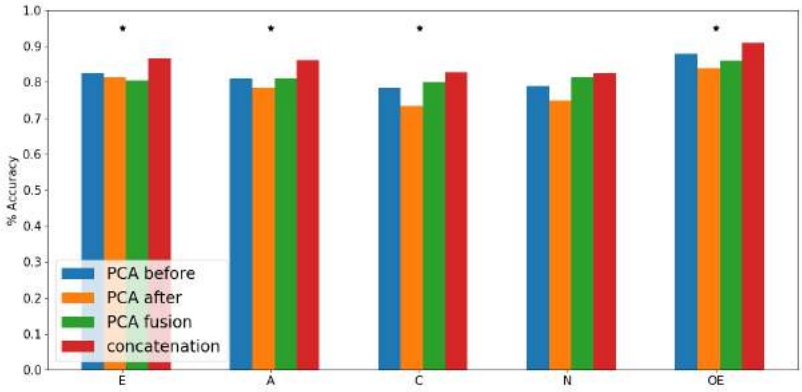


Figure 6.5: Exploratory experiment investigating the recognition accuracy using different fusion methods. The experiment is performed on the Salsa dataset poster session. The p-values less than 0.05 are summarized using one asterisk. Feature concatenation shows better results than the other methods, being significantly superior on four out of five traits.

was applied on the features extracted from different modalities, for dimensionality reduction, as well as to remove correlations between the features. Following [43], 98% of variance was kept. 3) PCA was applied on the concatenated features. As decision fusion, the FC layers as well as the softmax layer were trained on the three descriptors (e.g. D_{pers}, D_{group}, and D_{prox}) independently and the recognition decision was fused (sum-rule) [43]. Fig. 6.5 shows the personality traits recognition results using the described fusion methods on the Salsa dataset poster session. The concatenation method exhibits the best results on all the big 5 traits, being significantly higher on four out of five traits (p-value less than 0.05 is summarized using one asterisk). Therefore, this method is used for the rest of the experiments.

6.3. EXPERIMENTS

In order to examine the strength of our framework, in this section, we present personality recognition experiments on two public datasets, namely, the Salsa Dataset [18] and the Nonsocial Dataset [44] (see description in Chapter 2, Section 2.10.2). The choice of these two datasets justifies the evaluation of the system in two different scenes, where, depending on the scenario, the deep Person stream is combined with distinct Context streams (Fig. 6.3).

The experiments are organized as follows: In Section 6.3.1, the experimental setup is introduced. In Section 6.3.2, our personality labels based on the Big-5 traits called "personality types ' are described . In Section 6.3.4, we test the two sessions of the salsa dataset independently. In Section 6.3.5, we compare our work with state-of-the-art systems for personality recognition, and, in Section 6.3.6, we evaluate our framework using the Big-Five personality traits as labels. Finally, qualitative results are explained in Sec-

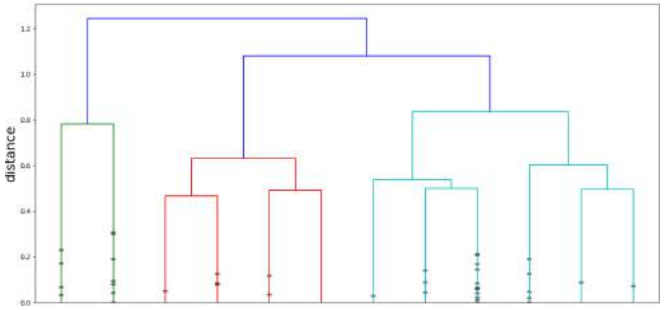


Figure 6.6: Dendrogram of the hierarchical clustering applied on the normalized personality trait scores from the used datasets. Three main clades (coloured in green, red, and azure) are found.

tion 6.4.

6.3.1. EXPERIMENTAL SETUP

Since the personality recognition experiments in Chapter 5 are the most related to the work in this chapter, the same experimental procedure is followed. In Section 6.3.2, we show how personality trait configurations are found in an unsupervised way. For the experiment in Section 6.3.4, we use the "K-fold cross validation" where $K = 10$. For the final comparison with the state-of-the-art methods, the "leave-one-out" (LOO) approach is used. Specifically, the data from a set of subjects is left out from the training procedure and used only for testing. In all our experiments we report the F1 accuracy, as it includes both the precision and recall metrics.

6.3.2. PERSONALITY TRAITS CONFIGURATIONS: PERSONALITY TYPES

As this work is a continuation of the model proposed in Chapter 5, the same experimental set-up is followed. In particular, the personality trait scores from the two datasets are normalized (between 0 and 1), and a hierarchical clustering technique [45] is applied to explore the trait score configurations.

Confirming the findings in Chapter 5, three main clusters are found. As shown in Fig. 6.6, the traits scores are grouped in three main clades (colored in green, red, and azure). Then, in order to assign a semantic meaning to the discovered clusters, we compute the average score of each personality trait in the different clusters.

In Fig. 6.7, we show the obtained results from the Salsa Dataset, which are consistent with the findings in Chapter 5 as well as the psychological theory proposed in [25]. The theory states that the Big-5 personality traits can be organized in three major types: Undercontrolled, Overcontrolled and Resilient. In Fig. 6.7, Resilient personality type (orange color) shows high Extraversion score and the lowest score in Neuroticism, the Undercontrolled type (green color) scores high in Extraversion as well as Neuroticism, and finally, the Overcontrolled type (blue color) has the lowest score in Extraversion and scores high in Neuroticism.

Finally, the three clusters $Y = [U, R, O]$ (namely, Undercontrolled personality, Resilient personality, and Overcontrolled personality) are used as labels in the personality

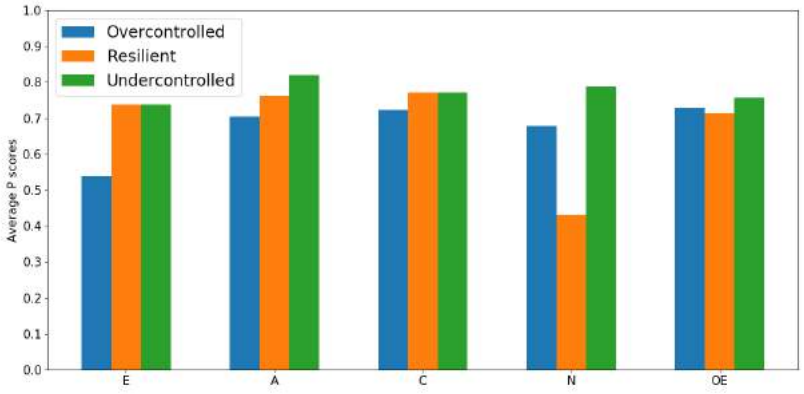


Figure 6.7: Average personality values for each trait in the discovered personality types from the Salsa Dataset. Undercontrolled personality type colored in green, Resilient personality type colored in orange, and Overcontrolled personality type colored in blue.

recognition experiments.

6.3.3. ABLATION STUDY

In order to test the contribution of every component in the proposed framework, in the next experiments, we will test different components' combinations. Specifically, for the Salsa dataset, we tested: D_{pers}, where only the Person motion stream is used. $D_{prox} + D_{group}$, where only the two context streams are used. $D_{pers} + D_{prox}$, where person motion (Section 6.1.2) is combined with social proxemics descriptors (Section 6.1.4), $D_{pers} + D_{group}$, in which the social group motion descriptors are used (Section 6.1.3), and finally, $D_{pers} + D_{prox} + D_{group}$, where all the descriptors are fused. As for the nonsocial context, we tested: D_{pers} stream alone, and $D_{pers} + D_{prox}$ encoding the person motion with the scene proxemics (Section 6.1.4).

6.3.4. SALSA DATASET SEPARATE SESSIONS

In this section, we apply our framework on the two sessions of the dataset independently. As the two sessions depict different social interactions (i.e. poster session presentation and cocktail party), our goal is to investigate whether the two sessions show different behavioral patterns. Our results using the "K-fold cross validation" testing approach are reported in Table 6.1.

For the results on the poster session part, the ablation study indicates that, although the fusion of all the context descriptors ($D_{pers} + D_{prox} + D_{group}$) obtains the highest accuracy, it does not improve significantly the result obtained using only $D_{pers} + D_{group}$. This may be explained by studying the scenario of the dataset, which displays interactions during a scientific poster session. During this type of sessions, individuals have to respect some spatial constrains, for example, if a poster presentation is already crowded,

Methods	Salsa poster-session	Salsa party-session
D_{pers}	76%	71.3%
$D_{prox} + D_{group}$	73.6%	75.83%
$D_{pers} + D_{prox}$	78.2%	75.61%
$D_{pers} + D_{group}$	79.6%	74%
$D_{pers} + D_{prox} + D_{group}$	**79.8%**	**77.25%**

Table 6.1: Personality prediction on the Salsa Dataset's sessions.

individuals cannot position themselves freely. Hence, in this situation, the proxemics descriptor does not always describe affective behaviors as explained by [20]. Additionally, note that the D_{pers} descriptor alone reaches higher accuracy than $D_{prox} + D_{group}$, supporting the explanation that proxemics features in this scenario does not show a clear link to personality patterns. On the other hand, the ablation study on the cocktail party session demonstrates the advantage of considering all available representations, where the three proposed descriptors ($D_{pers} + D_{prox} + D_{group}$) obtain the best accuracy within the rest of the combinations. As this part of the dataset depicts the participants freely interacting with each other, social behaviors are more natural and less constrained by social roles (i.e. poster presenter vs. audience). While natural behaviors may be more representative of a certain personality type, they are more complex to model, and as a consequence, the overall recognition accuracy is lower. The complexity in this scenario affects all the parts of the framework, as free interaction and group formation create occlusions and clutter, challenging the tracking and motion modeling components. Finally, we conclude that by merging two sessions together, we can overcome the technical challenges explained above. Therefore, in the next session, for each participant, the data from the poster session, and the data from the cocktail party are merged. This strategy is in line with our goal to obtain a general model able to work in different situations.

6.3.5. PERSONALITY TYPES RECOGNITION

In this section, the discovered clusters $Y = [U, R, O]$, which can be linked to the personality types, are utilized as ground-truth personality labels. The three personality types correspond to: Undercontrolled type, Overcontrolled type, and Resilient type (see description in Chapter 5).

Baseline Methods. We compare our framework against the three most related works. In particular, PR_{cl}, developed in Chapter 5, uses an AE-LSTM framework for personality recognition on the Nonsocial dataset. $Clips + MTLN$ was proposed by [23], and it uses similar skeleton descriptors as input to a CNN framework called MTLN (see details in Chapter 2 Section 2.9.1). Finally, EL-LMKL [46] was proposed for leadership recognition as well as personality traits recognition using a CNN model on optical flow features (see details in Chapter 2 Section 2.10.1). Please note that all the baselines were implemented by the authors, as to the best of their knowledge, there do not exist personality recognition results on the Salsa dataset.

Results. Personality recognition results obtained on the two parts of the Salsa dataset, as well as on the Nonsocial dataset, are shown in Table 6.2. We report the results comparing our framework with the baseline methods using a leave-one-out (LOO) approach.

Methods	Salsa LOO	Nonsocial LOO
Baseline		
PR_{cl} [3]	59.59%	55.3%
EL-LMKL [46]	61.25%	70.7%
Clips+MTLN [23]	68.48%	-
Proposed		
D_{pers}	68.03%	69.2%
$D_{prox} + D_{group}$	67.06%	-
$D_{pers} + D_{prox}$	70.2%	**72.6%***
$D_{pers} + D_{group}$	71.96%	-
$D_{pers} + D_{prox} + D_{group}$	**73%****	-

Table 6.2: Experiments on the two datasets for Personality Recognition using LOO= leave-one-out testing approach. The p-values less than 0.001 are summarized with two asterisks, p-values less than 0.01 are summarized with one asterisk.

Traits	Salsa	Nonsocial
Extraversion	62.8%	50%
Agreeableness	65.3%	55.3%
Conscientiousness	60.8%	65.1%
Neuroticism	61.9%	68.2%
Openness	62.6%	60.8%

Table 6.3: Personality Recognition using Traits scores.

Overall, the proposed method is able to reach the highest accuracy results in both datasets. Furthermore, the statistical significance of the results is computed in respect to the baselines (indicated by the number of asterisks). Looking at the ablation study results (lower half of Table 6.2), individual motion (D_{pers}) is the most informative one, indicating that in both scenarios, human motion is critical for finding behavioral patterns connected to personality. Considering only the context information ($D_{prox} + D_{group}$) is not as meaningful as human motion. Finally, by aggregating the person as well as the context information, we obtain the highest recognition accuracy demonstrating the value of the proposed framework.

6.3.6. PERSONALITY TRAITS RECOGNITION

In this section, the original Big-5 personality traits are used as ground-truth labels $Y = [E, A, C, N, OE]$ (namely, Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Open to Experience). The Big-5 personality traits [24] is the most popular model for personality recognition, hence, it is of great importance that the proposed framework is evaluated using the big five traits labels. To the best of the authors' knowledge, this is the first study presenting the recognition of personality traits on the two chosen datasets, providing new insights on personality patterns in the depicted scenarios.

In order to use the traits scores as classification labels, following the approach used by [4], the score of each trait is transformed into a binary value (HIGH or LOW), based

on the median value computed from the population. Note that every trait is evaluated independently from the scores of the other traits. As for the proposed architecture, we employ the descriptors that obtained the best results in the previous experiment. Hence, the $D_{pers}+D_{group}+D_{prox}$ framework is used for the Salsa dataset, and the $D_{pers}+D_{prox}$ framework is used for the Nonsocial dataset.

In Table 6.3, we report the personality traits recognition results following a LOO testing procedure for both datasets. For the Salsa dataset, the highest accuracy is obtained for the Agreeableness trait, followed by the rest of the traits that obtain similar accuracy. Considering the scenario in which the dataset was recorded (university environment), as well as the social contexts, behavioral attributes related to the Agreeableness trait (e.g. talkative and open for a discussion) can be easily matched with researchers' behaviors depicted in the data. The recognition scores in the rest of the traits indicates that the Salsa dataset is rich of social information. In this sense, the participants are exposed to frequent and long interactions which make them express several personality attributes. For example, as the participants are allowed to freely interact with each other, attributes like talkative and sociable belonging to the Extraversion trait, are more evident than in the other scenarios.

For the Nonsocial dataset, which contains problem solving tasks data for Activity of Daily Living (ADL) applications, the highest accuracy is obtained for the Conscientiousness and Neuroticism traits. When it comes to engagement with the scene, searching for objects, deliberate scanning strategies or curiosity are attributes related to the Conscientiousness trait, which can be matched to behaviors observed in the Nonsocial dataset. As the tasks forced the participants to find solutions to given problems in a filmed environment, attributes belonging to the Neuroticism such as stress are retrieved by our system. Moreover, given the nonsocial scenario, as expected, the Extraversion as well as the Agreeableness traits are the hardest to predict.

6.4. QUALITATIVE RESULTS

As explained in Section 6.3, all the proposed descriptors contribute to the understanding of certain behavioral patterns, bearing to an improved personality recognition performance. In this section, aiming to further investigate the learning process, we disentangle the behavior of each descriptor at test time. Specifically, we select two frame sequences t_1, t_2 belonging to a subject with high Agreeableness trait (skeleton colored in red). One sequence, (t_1) depicted in Figure 6.8 was classified correctly and the network showed high confidence that the sequence belonged to the right label, whereas the other sequence (t_2) shown in Figure 6.9 was misclassified.

We are interested in investigating the interaction between individuals, therefore, for this experiment, we focus only on the upper-body joints motion $j_{1,\ldots,10}$. The motion values of the upper-body joints are extracted from the three descriptors $\rho_{pm,p}$ (introduced in Section 6.1.2), $\rho_{sm,p}$ (introduced in Section 6.1.3), and $\rho_{pd,p}$ (introduced in Section 6.1.4) and averaged for visualization purposes. The graphs display the averaged descriptor values before being fed to the CNN learning framework (i.e. pixel values between $0-255$). On the y-axis, we show the person motion values ($\rho_{pm,p}$ red line), the average motion of the social group ($\rho_{sm,p}$ green line), and proxemics distance between the target person and the social group ($\rho_{pd,p}$ purple line) in a frame by frame sequence (x-axis).

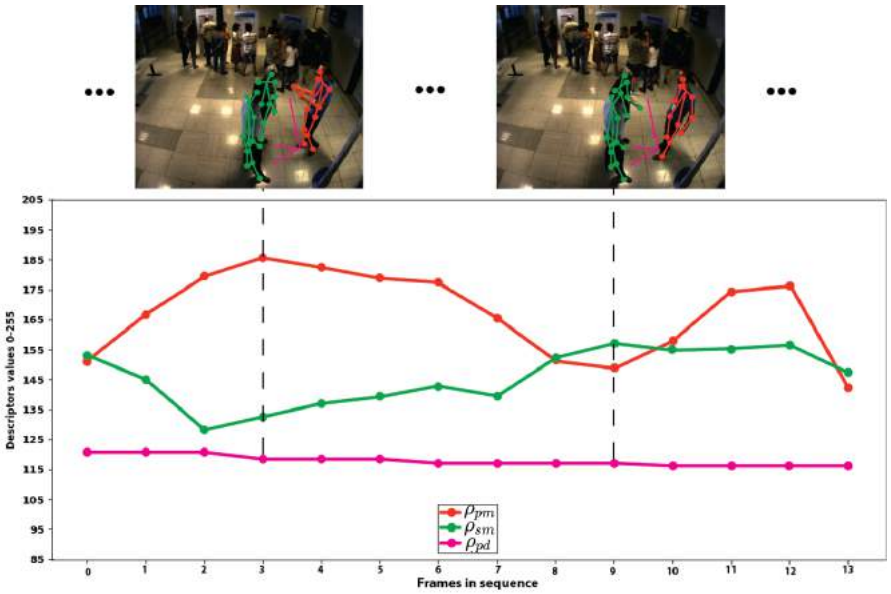


Figure 6.8: Example of correctly predicted sequence. The graph contains the descriptors values (red for the $\rho_{pm,p}$ descriptor, green for the $\rho_{sm,p}$ descriptor and purple for the $\rho_{pd,p}$ descriptor) on the y-axis, for each frame in the sequence on the x-axis. In this sequence, the model is predicting the personality class of the individual depicted with red color. An alternation of high values is noticeable between $\rho_{pm,p}$ and $\rho_{sm,p}$, indicating that the two parts are taking turns in the conversation. Correspondingly, the image at frame 3 shows high motion from the skeleton colored in red, while the image at frame 9 shows high motion from the social group.

Note that in the Salsa dataset $t = 15$ frames correspond to 1 second.

In the correctly classified sample (Fig. 6.8), we can notice an alternation of high values between $\rho_{pm,p}$ and $\rho_{sm,p}$ in time. This variation of high motion may denote that the two parts are conversing, each taking turns in a discussion. The displayed behavioral pattern, identified by our framework as belonging to the Agreeableness trait, is in line with previous literature studies, which correlate this trait with aspects like cooperation and empathy [47].

In the misclassified sample (Fig. 6.9), the conversation is more limited, and the descriptors values display low variation between $\rho_{pm,p}$ and $\rho_{sm,p}$ values. As we saw from the example above, behavioral patterns connected to the Agreeableness trait should show large and highly interactive conversational groups. This is also confirmed by literature findings in [48], which observed individuals with high Agreeableness traits participating in more groups' discussions than the rest of the individuals. The pattern observed in Fig. 6.9 violets these findings, hence, to further improve the results, our framework should be able to better model situations in which the motion is rather limited, by including additional sources of information, such as facial or audio data.

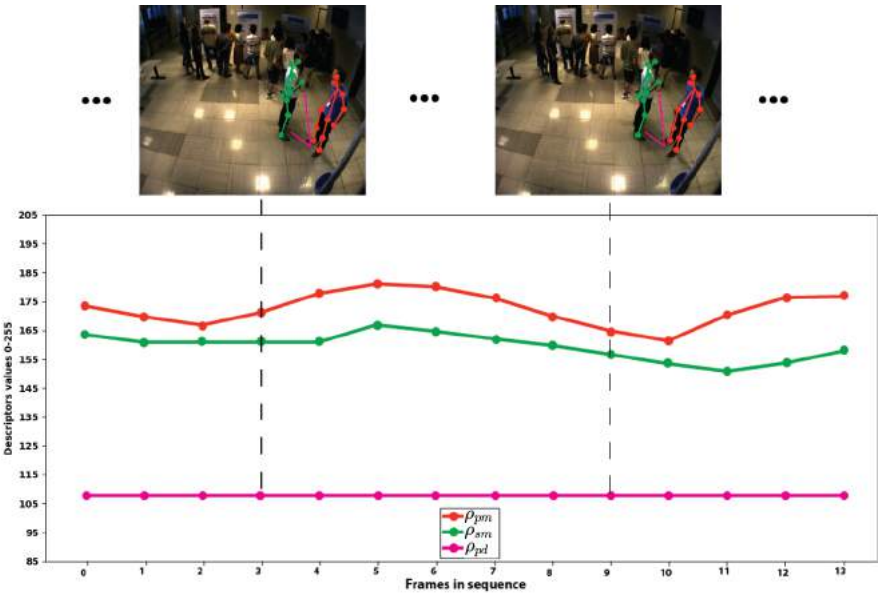


Figure 6.9: Example of misclassified sequence. The graph contains the descriptors values (red for the $\rho_{pm,p}$ descriptor, green for the $\rho_{sm,p}$ descriptor and purple for the $\rho_{pd,p}$ descriptor) on the y-axis, for each frame in the sequence on the x-axis. In this sequence, the model is predicting the personality class of the individual depicted with red color. As the interaction between $\rho_{pm,p}$ and $\rho_{sm,p}$ is limited, the descriptors values are quite low. Hence, the system did not associate this behavioral pattern to the Agreeableness trait.

6.5. DISCOVERED PERSONALITY PATTERNS

In this section, discriminative personality patterns discovered by our Person-Context CNN framework are investigated. Specifically, class activation maps [49] using the classification weights are explored in the social scenario data. We use the D_{pers}, D_{group} descriptor on the Salsa dataset, to reveal the interaction between individual and group behaviors with different personality traits. The activation of the feature maps extracted by the VGG component, corresponds to spatial information on the descriptor images [23]. We aim to discover the importance (determined by the classifier weights in the softmax layer) of both individual dynamics as well as the social group dynamics for each personality trait $Y = [E, A, C, N, OE]$, by applying the class activation maps on the Person-Context descriptors, as shown in Eq. 6.4. We define Act_c as the activation map for personality class c, where each spatial element is given by

$$Act_c(x, y) = \sum_{k=1}^{K} w_k^c f_k(x, y) \tag{6.4}$$

where w represents the weights belonging to the softmax layer, $f_k(x, y)$ is the activation of the k_{th} unit in the last convolutional layer at spatial location (x, y). The number of units K is formed of two concatenated components, one corresponding to the Person

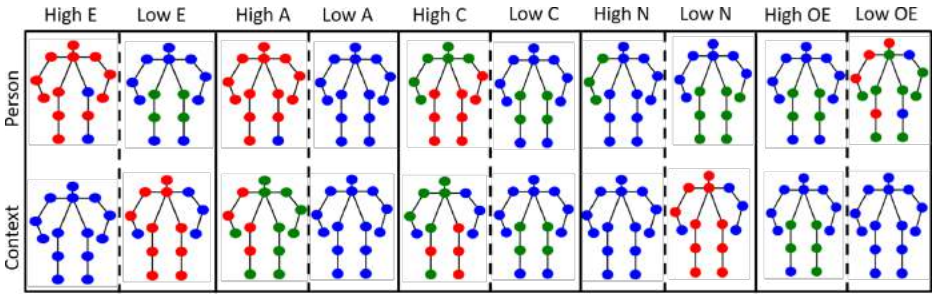


Figure 6.10: Discovered personality patterns using Person-Context descriptors, given the **High** and **Low** scores of each personality trait: **E**xtraversion, **A**greeableness, **C**onscientiousness, **N**euroticism and **O**penness to **E**xperience. Low CNN weights activation in blue, medium CNN weights activation in green, and high CNN weights activation in red.

information (D_{pers}) and one to the Context information (D_{group}). Besides, we are able to retrieve the specific joint activations due to the descriptors' dimension. In fact, as the CNN input image is of size 68×68 encoding the motion of $J = 17$ body joints, we can assume that every joint corresponds to an activation kernel of 4 pixels in our original feature.

We are interested in visualizing meaningful personality patterns for each personality trait. Specifically, we pose the following questions: 1) *How is the relation between Person-Context dynamics reflected in the traits?* and 2) *Do the dynamics correspond to the trait attributes defined by psychologists?* The discovered personality patterns for each trait are displayed in Fig. 6.10. In particular, we visualize the learned patterns from both Person and Social group CNN networks. Note that as D_{group} is obtained by averaging the body motion of all the members of the social group, in Fig 6.10, we can have a 1 to 1 comparison between the motion of each individual and the average motion of the social group.

We quantize the activation kernel of each joint in three groups: low activation (first quartile) depicted in blue color, medium activation (second quartile) depicted in green color, and finally, high activation (third quartile) is depicted in red color. When an individual has a high Extraversion trait (e.g. talkative, outgoing), the CNN weights show high activation on the features coming from the Person stream, and low activation in the features coming from the Context stream. As a consequence, if an individual has a low Extraversion trait, features coming from the Context stream are more important. Thus, to answer the first question, the two stream dynamics are learned by the CNN architecture and are mapped to personality-related behavioral patterns. In order to answer the second question, we highlight that, since individuals with a high Extraversion trait are described as talkative and outgoing, when it comes to social interaction, they tend to be the *center of attention*, with the rest of the group usually tending to become more passive [50]. On the other hand, CNN weights for individuals with high Agreeableness and Conscientiousness traits, show high activation on both Person-Context streams, demonstrating that these personality types like to be engaged with the social group, without trying to dominate the situation [51]. Medium/high CNN activation on both Person-Context streams is found also for individuals with low Neuroticism and Openness traits,

showing that having low scores in attributes like being tense/nervous, makes individuals more sociable.

6.6. CONCLUSIONS

In this chapter, we presented a novel CNN-based framework for personality recognition. Our model analyzes the scene at multiple levels of granularity. Firstly, we propose a descriptor that encodes the skeleton joints motion of each individual in the scene. Secondly, we propose a descriptor that encodes the interaction between individuals within small social groups by extracting their average skeleton joints motion. Thirdly, we propose a descriptor that encodes the interpersonal distances (proxemics) between every individual in the scene. Additionally, we demonstrate that our proxemics features can be applied also in a nonsocial scenario, encoding scene interaction information. Experiments on two personality recognition datasets demonstrate the effectiveness of our approach, showing that modeling together Person-Context information significantly improves the state-of-the-art personality recognition results. Furthermore, we presented CNN class activation maps for each personality trait, providing novel insights into nonverbal behavioral patterns linked with personality attributes defined by theories from behavioral psychology. We believe these findings are of great importance for future interdisciplinary behavioral studies, aiming to combine data-driven approaches with psychological studies.

REFERENCES

[1] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[2] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *European Conference on Computer Vision*, pp. 400–418, Springer, 2016.

[3] D. Dotti, M. Popa, and S. Asteriadis, "Behavior and personality analysis in a nonsocial context dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2354–2362, 2018.

[4] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pp. 37–42, ACM, 2010.

[5] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929–2936, IEEE, 2009.

[6] T. M. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition.," in *CVPR*, pp. 3425–3434, 2017.

[7] M. Snyder, J. A. Simpson, and S. Gangestad, "Personality and sexual relations.," *Journal of Personality and Social Psychology*, vol. 51, no. 1, p. 181, 1986.

[8] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism.," *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[9] E. Y. Fu, H. V. Leong, G. Ngai, and S. C. Chan, "Automatic fight detection in surveillance videos," *International Journal of Pervasive Computing and Communications*, vol. 13, no. 2, pp. 130–156, 2017.

[10] W. J. Hansen, "User engineering principles for interactive systems," in *Proceedings of the November 16-18, 1971, fall joint computer conference*, pp. 523–532, ACM, 1971.

[11] H. Achten, "Buildings with an attitude," in *Stouffs, R. andSariyildiz, S.(eds.), ComputationandPerformance–Proceedings of the 31st eCAADe Conference*, vol. 1, pp. 477–485, 2013.

[12] F. Mairesse and M. A. Walker, "Towards personality-based user adaptation: psychologically informed stylistic language generation," *User Modeling and User-Adapted Interaction*, vol. 20, no. 3, pp. 227–278, 2010.

[13] A. Tapus, C. Tapus, and M. J. Mataric, "Hands-off therapist robot behavior adaptation to user personality for post-stroke rehabilitation therapy," in *Proceedings 2007*

IEEE International Conference on Robotics and Automation, pp. 1547–1553, IEEE, 2007.

[14] D. Wang, B. Subagdja, Y. Kang, A. H. Tan, and D. Zhang, "Towards intelligent caring agents for aging-in-place: issues and challenges," in *Proceedings of the 2014 IEEE Symposium on Computational Intelligence for Human-Like Intelligence*, pp. 1–8, IEEE Computer Society, 2014.

[15] Y.-S. Lin and C.-C. Lee, "Using interlocutor-modulated attention blstm to predict personality traits in small group interaction," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 163–169, ACM, 2018.

[16] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2018.

[17] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. Van Gerven, and R. Van Lier, "Multimodal first impression analysis with deep residual networks," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 316–329, 2017.

[18] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1707–1720, 2016.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, pp. 568–576, 2014.

[20] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of f-formations.," in *BMVC*, vol. 2, p. 4, 2011.

[21] T. L. Chartrand, G. M. Fitzsimons, and G. J. Fitzsimons, "Automatic effects of anthropomorphized objects on behavior," *Social Cognition*, vol. 26, no. 2, pp. 198–209, 2008.

[22] M. L. Walters, K. Dautenhahn, R. Te Boekhorst, K. L. Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, and I. Werry, "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pp. 347–352, IEEE, 2005.

[23] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 4570–4579, IEEE, 2017.

[24] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[25] J. Block and J. H. Block, "The role of ego-control and ego-resiliency in the organization of behavior," in *Development of cognition, affect, and social relations*, pp. 49–112, Psychology Press, 2014.

[26] A. Hutchinson, "Labanotation," *Journal of Aesthetics and Art Criticism*, vol. 13, no. 2, pp. 276–277, 1954.

[27] K. K. Roudposhti and J. Dias, "Probabilistic human interaction understanding: Exploring relationship between human body motion and the environmental context," *Pattern Recognition Letters*, vol. 34, no. 7, pp. 820–830, 2013.

[28] K. K. Roudposhti, U. Nunes, and J. Dias, "Probabilistic social behavior analysis by exploring body motion-based patterns," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1679–1691, 2016.

[29] S. Lau and Y. Nie, "Interplay between personal goals and classroom goal structures in predicting student outcomes: A multilevel analysis of person-context interactions.," *Journal of educational Psychology*, vol. 100, no. 1, p. 15, 2008.

[30] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[31] G. Bradski, "The OpenCV Library." https://opencv.org/, 2000.

[32] M. Koppensteiner, "Motion cues that make an impression: Predicting perceived personality by minimal motion information," *Journal of experimental social psychology*, vol. 49, no. 6, pp. 1137–1143, 2013.

[33] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[34] M. Cristani, V. Murino, and A. Vinciarelli, "Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 51–58, IEEE, 2010.

[35] E. J. Vanman and A. Kappas, ""danger, will robinson!" the challenges of social robots for intergroup relations," *Social and Personality Psychology Compass*, vol. 13, no. 8, p. e12489, 2019.

[36] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proc. of the 17th Int. Conf. on Pattern Recognition (ICPR 2004).*, vol. 2, pp. 28–31, IEEE, 2004.

[37] L. Zhao, Q. Hu, and Y. Zhou, "Heterogeneous features integration via semi-supervised multi-modal deep networks," in *International Conference on Neural Information Processing*, pp. 11–19, Springer, 2015.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[41] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 399–402, ACM, 2005.

[42] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[43] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, 2016.

[44] D. Dotti, M. Popa, and S. Asteriadis, "Unsupervised discovery of normal and abnormal activity patterns in indoor and outdoor environments.," in *VISIGRAPP (5: VISAPP)*, pp. 210–217, 2017.

[45] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic acids research*, vol. 16, no. 22, pp. 10881–10890, 1988.

[46] C. Beyan, M. Shahid, and V. Murino, "Investigation of small group social interactions using deep visual activity-based nonverbal features," in *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 311–319, ACM, 2018.

[47] Y. J. Weisberg, C. G. DeYoung, and J. B. Hirsh, "Gender differences in personality across the ten aspects of the big five," *Frontiers in psychology*, vol. 2, p. 178, 2011.

[48] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 5–14, ACM, 2015.

[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

[50] Y.-E. Lu, S. Roberts, P. Lio, R. Dunbar, and J. Crowcroft, "Size matters: variation in personal network size, personality and effect on information transmission," in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 4, pp. 188–193, IEEE, 2009.

[51] C. McCarty and H. Green, "Personality and personal networks," *Sunbelt XXV International Sunbelt Social Network Conference*, pp. 16–20, 2005.

7

TEMPORAL TRIPLET MINING FOR PERSONALITY RECOGNITION

This chapter is based on the following publication:

- D. Dotti, E. Ghaleb, and S. Asteriadis, "Temporal triplet mining for personality recognition", in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 171–178.

7.1. INTRODUCTION

Extensive studies in the field of psychology showed that attitude, mood, and personality are directly connected to human behavioral patterns [1]. Since these human characteristics are often subtle, the affective computing field still faces several challenges.

In our previous studies on personality recognition (Chapter 5 and Chapter 6), we showed meaningful results on the connection between short spatio-temporal behavioral features and personality labels. Motion and context samples were extracted in separate time windows and mapped to personality labels. Yet, the time window approach splits the behaviors into fixed motion fragments risking to make the model focus on sub-actions. As human motion typically evolves in different ways, the assessment of the similarity between behavioral sequences is still a non-trivial problem. For example, even though two walking behaviors can be formed by different patterns of short motion features (e.g. slow, fast, smooth motion, or jerky motion), their overall similarity should

be high as they belong to the same semantic behavior. Another challenge faced while analysing behavioral sequences in real world scenarios is that it is very difficult to determine the beginning and the end of semantic behaviors. Behaviors may evolve in varying lengths making their interpretation and semantic comparison more challenging. In these circumstances, we aim at matching human behaviors implicitly and semantically via Deep Metric Learning (DML) [2]. Our final goal is to learn more meaningful semantic behaviors that can help the recognition of clearer personality patterns. For instance, jerky and fast motion can be associated with high-Neuroticism trait while smooth and slow motion can be associated with low-Neuroticism trait.

In this chapter, we propose a novel framework that further expands the use of short-term body information, context learning and their interaction in time, using DML [2]. DML has become popular with the advances and success of deep learning [3]. It projects embeddings produced by mapping functions ($f(x)$) such as a CNN, onto a manifold space where semantically similar samples are closer while the dissimilar ones are placed apart from each other. There have been different designs of loss functions in DML, such as contrastive [4], triplet [5] and N-pair loss [6]. In this study, the triplet loss is utilized as loss function. The triplet loss is a loss function based on triplet sets: an anchor, a similar example called "positive sample", and a dissimilar one called "negative sample". The ultimate goal of the Triplet Loss function is to construct a latent space where the anchors are closer to the positive samples than the negative ones.

Effectively measuring the similarity between two human motions is a complex problem as human poses have to be compared across a temporal set of frames. This aspect introduces several challenges such as alignment as well as pose to pose comparison. Recently, authors in [2] proposed a DML method on human motion data, showing that computing pose similarities in a latent space helps to capture the semantic relationship between motions. Hence, by adding the temporal analysis to our DML framework, we help the system to discover higher semantic movements that enhance the discovery of discriminative personality patterns, and therefore, improve the personality recognition task.

In Fig. 7.1, we show a high level description of the proposed model. The analyzed data consists of people performing activities in certain scenarios. Every person is different in the way they act and move (indicated by the empty shapes; top left), however, there exist common behavioral patterns that can be categorized into discrete personality classes (indicated by colored shapes; top left). In this chapter, first, we extract human motion as well as proxemics features in a time-window approach (top right of the figure). Second, we introduce the notion of Temporal Identification Similarity Metric Learning (TISML), which is used to train the framework and consists of two major components: The first one is an identification signal based on personality labels, while the second one is a similarity signal based on DML. The general goal of the DML approach is to construct models that bring samples with similar labels (positive examples) closer together, while pushing apart samples with different labels (negative examples). Additionally, in the training stage, our intuition is to add another constraint for the selection of positive/negative samples. We select positive samples in the temporal proximity of the anchors (within a time-window) to encourage the model to generate embeddings with temporal relation, while maintaining a high discriminative power for personality

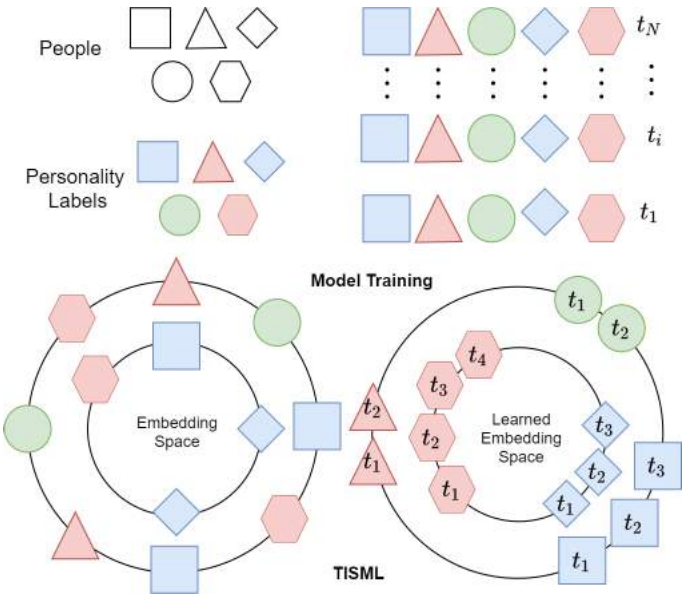


Figure 7.1: High level description of the proposed model. The general goal of our approach is to create an embedding space optimized for the personality recognition task. During training time, the model is encouraged to match similar short-term spatio-temporal descriptors using the proposed Temporal Identification Similarity Metric Learning (TISML) framework to create discriminative behavioral sequences with varied temporal relation (bottom row, right).

recognition. We assume that samples in the temporal proximity are more likely to have a semantic relation with the anchor (i.e. belonging to the same behavior), and therefore, they can carry important information for the personality recognition task.

The goal of TISML is to use the DML as well as Personality recognition signals for the assessment of meaningful personality-related behavioral patterns that yield to an improved personality recognition accuracy. The lower half of Fig. 7.1 illustrates the training process, where, before training, samples are distributed in the embedding space (bottom left). Our proposed approach employing TISML helps the model to generate temporally, as well as semantically related embeddings (bottom right). One advantage of our approach is that motion patterns are implicitly and semantically matched via DML using temporal as well as personality information. In this way, sequences of varying lengths can be matched to the same personality label without any additional constraints or alignments.

The contributions of this chapter are as follows. Firstly, we build a novel deep framework that learns temporal and discriminative motion patterns in real-world scenarios. We experimentally show that our generated embeddings perform better than state-of-the-art short-term motion samples. Secondly, using TISML, we encode the relation of temporally adjacent spatio-temporal samples, hence, without introducing any temporal constraint or alignment during training time, motion dynamics carrying similar semantic values are matched via DML. Thirdly, extensive experiments are conducted to

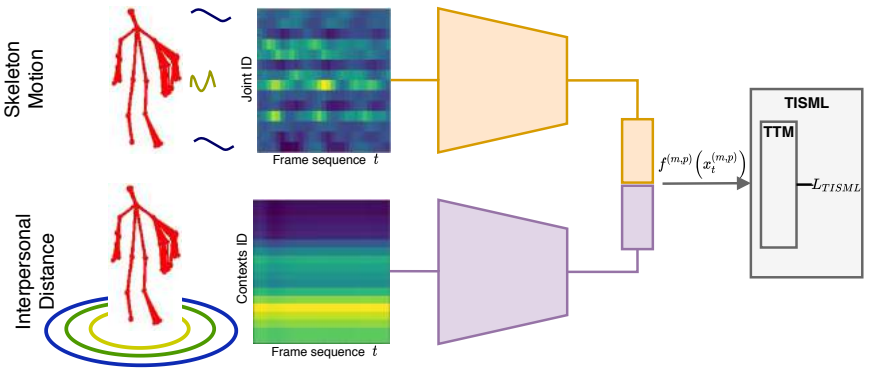


Figure 7.2: The proposed Architecture. Two descriptors representing the skeleton temporal motion as well as the spatial interaction are extracted from every frame sequences t. The descriptor images show the evolution over time (x-axis) of the reference information (y-axis). The reference information consists of the skeleton motion evolution for the person descriptor, and proxemics distances with respect to the surrounding for the context descriptor. Deep CNN models are then used to obtain a compact representation of each spatio-temporal patch. The outputs of the CNN models are concatenated and fed into the proposed learning framework TISML. Temporal Triplet Mining (TTM) is employed to select temporally related positive samples encouraging the model to learn meaningful behavioral sequences that bear a higher discriminative power. Finally, a double objective loss function L_{TISML} is adopted for personality recognition and personality retrieval.

investigate the relation between local motion features, global context features and their interactions in time using two real-world datasets.

7

7.2. THE PROPOSED FRAMEWORK

In this chapter, we propose a framework to encode local motion dynamics from the human body in combination with global interpersonal distances (proxemics) to encode personality-dependent behavioral patterns. Our work employs DML to map spatio - temporal descriptors to an optimized latent space, where, behaviors with discriminative power are learned and grouped together, whereas non-informative sequences are positioned far apart. As human behaviors are very dynamic and change according to the situation, it is very difficult to find semantic similarities between them [2]. Therefore, a novel Temporal Triplet Mining (TTM) strategy tailored for behavioral data is proposed. We argue that taking advantage of the triplet mining scheme, short-term spatio-temporal descriptors are implicitly matched, allowing the creation of an embedding space that encodes behavioral patterns of varying sizes optimized to retrieve personality-conditioned behaviors.

Fig. 7.2 shows our framework architecture. Skeleton motion as well as proxemics descriptors are extracted for every frame sequence of size T. As the two descriptors capture the motion and the spatial dynamics in a sequence, two separate CNN architectures are leveraged to obtain compact representations of the input features. The obtained representations are concatenated and fed to the Temporal Identification Similarity Metric Learning (TISML) component. TISML aims to project the concatenated motion (m) and proxemics (p) embeddings produced by the two CNNs (which serve as mapping func-

tions of the raw features) $f^{(m,p)}(x^{(m,p)}) : \mathbb{R}^{d^{(m,p)}}$ onto a shared feature space $\mathbb{R}^d$. Similar features are positioned closely and dissimilar ones are put far apart from each other based on data similarity and personality class. To do so, within TISML, a simple but effective Temporal Triplet Mining (TTM) approach is proposed to facilitate the overall learning effort (Section 7.3.3).

7.2.1. MOTION FEATURES

An increasing number of studies showed that body expressions are indicators of affective states as informative as facial expressions [7]. Moreover, systems that use solely body posture information (discarding the video data) provide a number of advantages such as privacy observance, higher flexibility and robustness to camera placement, coverage, and occlusions. In this work, skeleton information is extracted from every frame using the OpenPose library proposed by [8]. As this method does not provide a tracking function, like in Chapter 6, a frame by frame tracking algorithm is added using the OpenCv library [9]. The goal of the frame by frame tracker is to identify which skeleton in one frame corresponds to the same skeleton in the next frame. Then, local temporal information is extracted from every skeleton joint in terms of joint motion and rotation. As explained in [10], similarities between short-term motions are easier to learn in respect to long-term sequences, as they embed less noise.

For every detected skeleton joint $j_{1,\dots,J}$, where $J = 17$, in frame sequences of size T, we compute its spatial as well as rotation evolution. We form a matrix with dimensions $J \times T \times 3$, where J indicates the total number of joints positions in T frames. This matrix contains the (x,y,z) values of the 3D coordinates of the J joints. Driven by our previous findings in Chapter 6, we opt for using cylindrical coordinates ρ, θ, and z, which have been shown to provide a more invariant motion descriptor.

Finally, to leverage the learning power of CNN models, we utilize motion image clips, in which we treat cylindrical coordinate values as pixel values [10, 11]. Hence, the values are converted into $0-255$ scale using a linear transformation. Fig. 7.2 top stream shows the motion descriptor construction, where, given a motion sequence t of size T, frame by frame motion values (x-axis) of all the detected joints (y-axis) are organized in a motion image. In this example, the highest motion values (yellow color) correspond to the skeleton arms.

7.2.2. PROXEMICS DISTANCES

One of the goals of this work is to build a system able to recognize user personality in different situations (i.e. in private homes or during social events), hence, in addition to local skeleton motion image, we build proxemics distance images that can be applied in both social as well as nonsocial scenarios.

Social Proxemics

Previous studies on social proxemics [12, 13] highlighted that interpersonal distance is an effective tool to understand individuals feelings and attitudes towards others. For example, an individual that stands distant from everyone the entire time may feel uncomfortable in a given situation, while an individual that stands close to others may feel comfortable for engaging in social interactions. Therefore, in the analyzed social sce-

nario, we utilize the euclidean distance between the subjects in the scene to define their interpersonal distance. Let $s^1, .., s^S$ define all the subjects in a given dataset, and let t define a frame sequence of size T. For every subject i at frame n (s_n^i), the joint $j = 1$, which corresponds to the body torso, is empirically chosen as reference point to compute the distance on. The distance between subject s_n^i, and the rest of the subjects $s_n^{1,...,S-1}$ is computed as $dist_n^s = d(s_n^i - s_n^{1,...,S-1})$.

By combining the interpersonal distances between subjects within the T frames, we obtain a matrix of size $(S-1) \times T$, where S is the total amount of subjects in the dataset, except s^i. Please note that, to overcome the problem of finding different amounts of subjects in the scene at different times, we set S equal to the total amount of subjects in the dataset $\mathbb{D}$. In the situation when not all the subjects are present in the scene, the maximum distance value is assigned ($\max_{dist_n^s}$).

Nonsocial Proxemics

For the nonsocial scenario, we use the intuition proposed in Chapter 6 and published in [14], in which proxemics is intended as "the way people use their personal space in relation to objects". There exist several studies describing how people engage with objects, and how this engagement correlates with human-to-human interactions (called Anthropomorphism) [15, 16]. For example, authors in [16] found that subjects' personality influences personal distances as well as personal spatial zones preferred even when the interaction occurs with robots. Authors in [15] also state that people "match" the personality attributed to the nonhuman entity such as pets or inanimate objects. Therefore, an interesting idea is to extract how people move and interact with their surroundings (i.e. proxemics towards objects instead of people). For example, we hypothesize that an Overcontrolled personality that has a high level of the Conscientiousness trait is more meticulous in the searching of objects than a Resilient personality. With the proposed spatio-temporal descriptors, we are able to map the searching patterns to personality behaviors and, therefore, improve the personality recognition task.

We utilize the most important objects $O = 6$ in the scene found in an unsupervised way in Chapter 6. We aim to extract the proxemics feature computing the distance between the subjects S and the objects $1, \ldots, O$, where $O = 6$, in time. We select one skeleton joint j_2 as the reference position of the human body, the neck joint was empirically chosen as it was the most robust to noise. The euclidean distance between the target subject s^i at frame n (s_n^i) and the objects $1, \ldots, O$ was computed for every frame sequence of size T, obtaining a final matrix $O \times T$.

Finally, the values are transformed into cylindrical coordinates, and converted between 0 and 255 using a linear transformation to be suitable for a CNN architecture. Note that in this work we keep only the ρ coordinate values, discarding ϕ, and z. Fig. 7.2, bottom stream, shows the motion descriptor construction process, where, the interpersonal distance between the given subject s^i and the context entities (O in the nonsocial scenario and S in the social scenario) are depicted on the y-axis and temporal information is depicted on the x-axis.

7.3. TEMPORAL IDENTIFICATION SIMILARITY METRIC LEARNING (TISML)

7.3.1. DEFINITIONS

In this study, for every subject s, motion and feature descriptors ($x^{(m,p)}$) are created to embed sequences of T frames. As a result, a dataset $\mathbb{D}$ contains a total of N descriptors associated with a personality label y, and can be defined as follows:

$$\mathbb{D} = \{(x_1^{1...T}, y_1), (x_2^{1...T}, y_2), ..., (x_S^{1...T}, y_S)\} \tag{7.1}$$

For simplicity, we refer to $x^{(m,p)}$ as x_s, which includes both motion and proxemics embeddings.

7.3.2. FORMULATION

TISML optimizes the mapping function $f(x_s)$ to generate embeddings correlated with a personality class. In our work, the personality recognition task is carried out using two loss functions. The first function is a similarity measure based on a DML loss which positions semantically related embeddings closer to each other (decreasing the intra-class variations), and positions the semantically unrelated embeddings far apart (increasing the inter-class variations) [17]. Specifically, we apply a DML based on the triplet loss strategy.

Triplet loss uses triplet sets: $\{f(x_s), f(x_s+), f(x_s-)\}$, where $f(x_s)$ is an anchor (baseline), $f(x_s+)$ is a positive (similar) sample to $f(x_s)$, and $f(x_s-)$ is a negative sample (i.e. different label) to $f(x_s)$. As shown in eq. (7.2), the optimization procedure aims to minimize the distance between the anchor (baseline) input to a positive sample while maximizing the distance from the anchor to the negative sample within a margin [5].

$$L_{DML}(f(x_s), f(x_s+), f(x_s-)) = \|f(x_s) - f(x_s+)\|_2^2 - \|f(x_s) - f(x_s-)\|_2^2 + \text{margin} \tag{7.2}$$

The second function in our work is an identification signal, which classifies a given embedding into one of the given personality type labels (e.g. $Y = [U, R, O]$). The identification signal is achieved by computing the softmax loss, i.e. a softmax activation plus a cross-entropy loss to predict the probability distribution over the personality labels [18] defined as:

$$L_{Ident}(f(x_s), y) = -\sum_{i=1}^{Y} -y_i \log \hat{y}_i \tag{7.3}$$

where $f(x_s)$ refers to the mapping functions that produced the motion and proxemics embeddings, and y is the target class. y_i is the target personality distribution, where $y_i = 0$ for all i except $y_i = 1$ for the target personality class i. $\hat{y}_i$ is the predicted probability value for personality i. Finally, the optimization of the network is achieved through the joint loss and is formulated as follows:

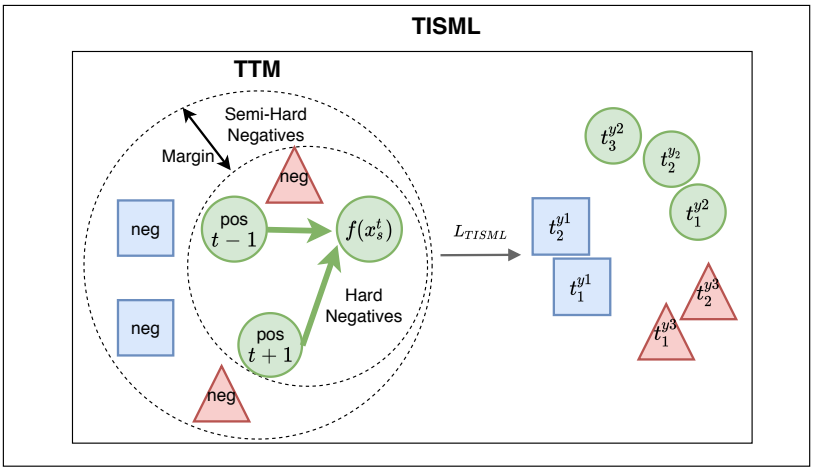


Figure 7.3: Temporal Triplet learning (TTM) strategy in our TISML framework. For a given anchor $f(x_s)$ at time t, the positive samples are selected if they are within a time-window tw. In this example, $tw = 3$ is centered to the anchor temporal position t, therefore, positive samples are selected at $t-1$ and $t+1$. The triplet loss is computed on the hard-negative as well as semi-hard negative samples.

$$L_{TISML} = L_{DML} + \lambda L_{Ident} \tag{7.4}$$

The goal of this formulation is to minimize the total loss L_{TISML} by combining the individual losses L_{DML} and L_{Ident}. λ is the weight used to trade off the class-wise triplet loss and the softmax loss in the total loss.

7.3.3. TEMPORAL TRIPLET MINING (TTM)

A prominent problem when using the triplet mining strategy is that the possible number of triplet sets could be extremely large, and training the DML can be challenging and prohibitively expensive. As a result, one of the main challenges in the triplet loss based DML is the slow-convergence during the training process. Without a careful and smart strategy to select the triplets, DML could only learn to map correctly easy sequences with little discriminative power. Therefore, in our TISML framework, we adopt a semi-hard triplet-sets mining strategy to guide the training process during the selection of the triplet sets. Moreover, as temporally adjacent short-term descriptors are likely to belong to the same semantic behavior, we propose the novel Temporal Triplet Mining (TTM) strategy.

The training process is displayed at Fig. 7.3 (TTM). For a given anchor $f(x_s)$ at time t, we restrict the selection of its positive samples to the temporal vicinity (i.e. within a temporal window tw). For example, if we set $tw = 3$ centered to the anchor temporal position t, the positive samples will be selected at $t-1$ and $t+1$. Regarding the negative samples, they are randomly chosen from other personality classes and could be from any time-window. Clearly, the choice of tw is critical to obtain the best optimization performance and its impact is discussed in the experiments section (Section 7.5.3).

The optimization process of L_{DML} (Eq. 7.2) is based on the online DML where the selection of triplet sets is based on mini-batches in each iteration during the training phase [6, 19]. Specifically, at every batch, we compute the loss expressed in Eq. 7.2 on all the triplets that satisfy the constraint expressed in Eq. 7.5, in which both the hard-negatives as well as the semi-hard negatives are considered. Hard negatives are defined as samples that are closer to the anchor than the positive samples, i.e. $d(f(x_s^t), f(x_{s-}^t)) < d(f(x_s^t), f(x_{s+}^t))$. Semi-hard negatives are defined as samples that are not closer to the anchor than the positive samples, but which still have positive loss due to the margin, i.e. $d(f(x_s^t), f(x_{s+}^t) < d(f(x_s^t), f(x_{s-}^t)) < d(f(x_s^t), f(x_{s+}^t) + margin$ (Fig. 7.2 (TTM)). A crucial step is to not take into account the easy negatives (i.e. $d(f(x_s^t), f(x_{s-}^t)) > d(f(x_s^t), f(x_{s+}^t)) + margin$) which would give a small loss, and therefore, yielding little information to the learning procedure.

$$d(f(x_s^t), f(x_{s-}^t)) < d(f(x_s^t), f(x_{s+}^t)) + margin \tag{7.5}$$

Since the proposed TTM minimizes the distance between samples in the temporal vicinity, adjacent short-term semantically related descriptors are aggregated forming an informative series of sequences with different lengths. One advantage of this approach is that unlike approaches like Dynamic-Time-Warping (DTW), we do not need any explicit time synchronization or alignment to find similarities between sequences of different lengths [2].

7.4. IMPLEMENTATION DETAILS

We use the Keras [20] and Tensorflow frameworks [21] for all computations in this work. As the datasets used are recorded using different frame rates, we experimentally set the frame sequence duration to $T = 180$ and $T = 90$ frames for the Nonsocial dataset and the Salsa dataset respectively (note that due to different frame rate, each sample sequence contains 6 seconds of data). For our image descriptors, we resize the final images to a 32×32 image to be a suitable input for the CNN architecture. Since our descriptor is not a real image, this dimensionality has the advantage of not being computationally expensive while still preserving the discriminative information. Given the motion as well as proxemics images as input, the VGG19 architecture [22], pre-trained on ImageNet [23], is adopted for feature extraction and learning. Although CNN models demonstrated to learn discriminative and generic features applicable in novel domains [10], early convolutional layers learn more low-level generic features, while higher-order convolutional layers learn more task-specific features. We use the Conv3 layer as output of our VGG model and, as our features describe skeleton temporal evolutions, we apply the pooling strategy proposed by [10], called Temporal Mean Pooling (TMP). This strategy applies the pooling only over the temporal, or row dimension of the feature maps (see description in Chapter 2, Section 2.9.1). The output descriptors are concatenated, and fed to two Fully-Connected Layers with batch normalization. Finally, our embeddings dimension of size 128 are used as input to our TISML.

Training. The proposed framework is trained on an Nvidia TITAN V GPU for 80 epochs. The batch size is obtained from two parameters (Fig. 7.3): the number of randomly selected anchors a and the size of the temporal-window tw. In order to obtain

batches containing a balanced amount of data for each personality label, we set $a = 15$ (i.e. 5 anchors for each label), while tw is set to be $tw = 5$ (more details on this parameter selection in Section 7.5.4). We use Adam optimizer with a $7e-6$ learning rate and the margin in the triplet loss (Eq. 7.2) is set to 0.1.

7.5. EXPERIMENTS

To evaluate the proposed study, we present personality recognition experiments on two public datasets recorded in different scenarios.

7.5.1. DATASETS AND LABELS

Following the experimental procedure of Chapter 6, all our experiments are carried out on the Salsa dataset and on the Nonsocial dataset (see description in Chapter 2, Section 2.10.2).

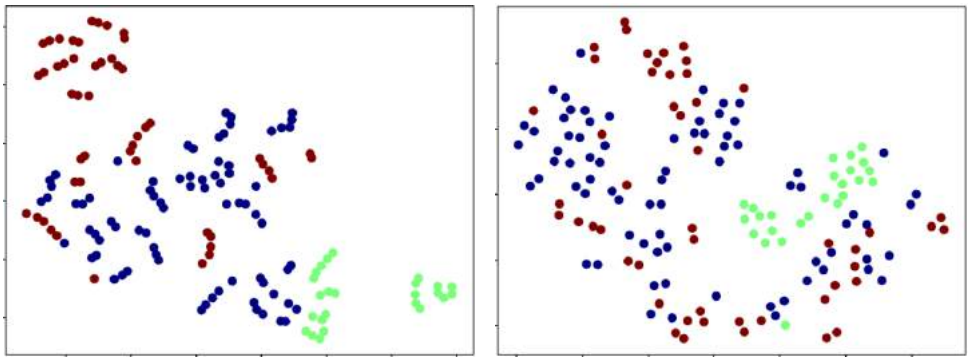
Similarly to our experiments in Chapter 6, our TISML model is tested using two types of personality labels: the three personality types [24] $Y = [U, R, O]$ (namely, Undercontrolled, Resilient, and Overcontrolled), and the Big-5 personality traits [25] $Y = [E, A, C, N, OE]$ (namely, Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness). For the personality types labels, we use the ones provided by [14], in which, the Big Five personality traits are projected onto three semantically higher categories called personality types [24]. The three personality types have the advantage of representing the commonly used one-dimensional independent traits (e.g. high/low Extraversion independent from high/low Neuroticism) as multidimensional dependent factors. For example, the Resilient personality type is represented by high Extraversion and Openness traits, and low Neuroticism (as discussed in chapter 5, Section 5.5.1). This multidimensional representation of personality behaviors was shown to be more similar to human judgments of behavioral characteristics [26]. For the Big-5 personality traits, we use the publicly available labels provided in both datasets.

7.5.2. EVALUATION PROTOCOL

As the TISML framework contains two objective functions (i.e. L_{DML} and L_{Ident}), following the experimental setup described in [18], we use the prediction of the softmax layer when comparing to the state-of-the-art results on the personality labels. On the other hand, the discriminative power of the generated embeddings is evaluated separately, as to the authors' knowledge this is the first work that employs a DML strategy for personality recognition. In all of our experiments (e.g. Table 7.1 and Table 7.2), we follow a leave-subjects-out based evaluation, in which a set of 6 subjects for the Nonsocial dataset, and a set of 3 subjects for the Salsa dataset, are left out from the training procedure and used only for testing. All results are reported in terms of F1 score. We compare our performance against various state-of-the-art results for both datasets. Please note that, from Chapter 6, only results using the same experimental protocol are reported for comparison.

Table 7.1: F1 score on the personality recognition task using different triplet mining strategies.

Triplet mining strategy	Salsa	Nonsocial
Random triplet mining RTM	72.0%	71.6%
Triplet mining through TTM	75.6%	74.9%



(a) Temporal Triplet Mining (TTM) Embeddings.

(b) Random Triplet Mining (RTM) Embeddings.

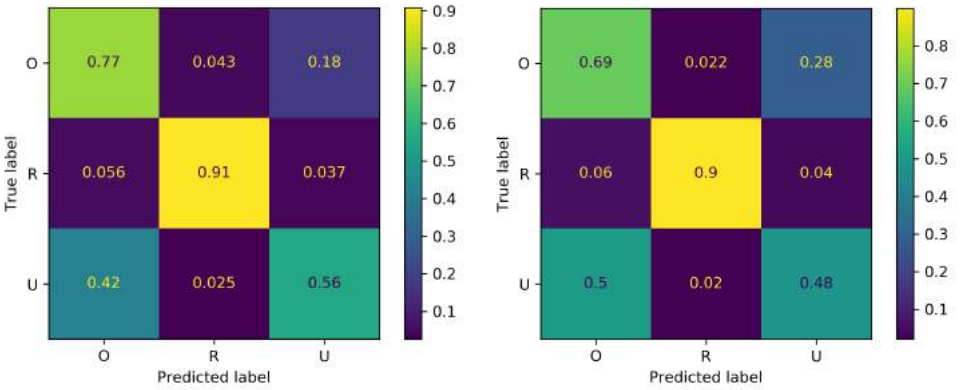
Figure 7.4: TTM helps in creating a more explicit separation between the personality classes. The Resilient personality is depicted using red, the Undercontrolled personality is depicted using blue, and the Overcontrolled personality is depicted using green. Moreover, short-term spatio-temporal descriptors are temporally aligned via TISML creating higher semantical sequences that are easier to map to personality labels.

7.5.3. TTM VERSUS RANDOM TRIPLETS MINING (RTM)

In this section, we investigate the proposed TTM selection strategy (Section 7.3.3) compared to the triplet loss standard random selection (RTM). In RTM, we select random positive samples from other subjects with the same personality. Formally, given an anchor $f(x_s^t)$ with a personality label y_s, positive samples $f(x_{s+}^t)$ are chosen given the following constraints: $y_{s+} = y_s$ and $s+ \neq s$.

The results are reported in Table 7.1. From this table, we can conclude that selecting random samples from different subjects with the same personality helps the model to learn similarities invariant to the identity of a subject. However, as the samples embed only short sequences (90 or 180 frames depending on the dataset), the discrimination between the personality classes becomes harder. On the other hand, selecting triplets with temporal constraints forces the model to learn similarities over samples further away in time, and therefore, learning more comprehensive behaviors which results in a stronger personality recognition performance.

In Fig. 7.4, we provide a visual example of the embeddings generated through the different triplet mining strategies. In particular, Fig. 7.4(a) depicts the short-term spatio-temporal descriptors of 5 batches from the Nonsocial dataset [13] encoded via the proposed TTM, while Fig. 7.4(b) shows the short-term spatio-temporal descriptors of the same 5 batches encoded using RTM. It is easy to notice that the separation between the $Y = [U, R, O]$ personality classes (where, the Resilient personality is depicted using red,



(a) Confusion matrix after TTM.

(b) Confusion matrix after RTM.

Figure 7.5: Confusion matrices on the personality prediction using two different triplets selection (TTM and RTM). Overall, selecting temporally related triplets yields a higher accuracy compared to randomly selected triplets.

the Undercontrolled personality is depicted using blue, and the Overcontrolled personality is depicted using green) is more explicit in Fig. 7.4(a), confirming the results of Table 7.1. Moreover, TTM embeddings that were in the temporal vicinity in the input space seem to be organized in sequences also in the latent space. As our latent space exploits the spatio-temporal similarity between samples, we can assume that the created subclusters are likely to belong to the same semantic behavior 7.4(a). On the other hand, the embeddings generated by RTM do not present any visible structure 7.4(b). As the short-term sequences are temporally aligned during the TISML learning, more discriminative behavioral patterns are determined to enhance the overall understanding of personality displays. Finally, analyzing the confusion matrix in Fig. 7.5, we argue that the proposed TTM yields an improvement in the class distinctions over RTM, showing that temporal information helps in the discrimination of personality patterns.

7.5.4. IMPACT OF TIME-WINDOW SELECTION

As explained in Section 7.3.3, the time-window parameter tw controls the selection of positive samples in the temporal vicinity of the anchor. This parameter is crucial to capture the affective behaviors of the analyzed subjects. Positive samples that are temporally too far away from the anchor risk to carry little similarity, and therefore, deceive the final goal of aggregating semantically related descriptors. On the other hand, positive samples that are too temporally close to the anchor risk to be "too similar", and therefore, yield an insignificant contribution to the learning objective.

In Fig. 7.6, we show the impact of the time-window selection. The Nonsocial dataset [13] contains data of subjects performing problem-solving activities in an indoor environment. As the subjects are moving to complete the given tasks, the behavioral data contains several active and fast interactions, hence, selecting positive samples in a large time-window range, (e.g. $tw = 11$), is not beneficial (blue line). As a matter of fact, fast

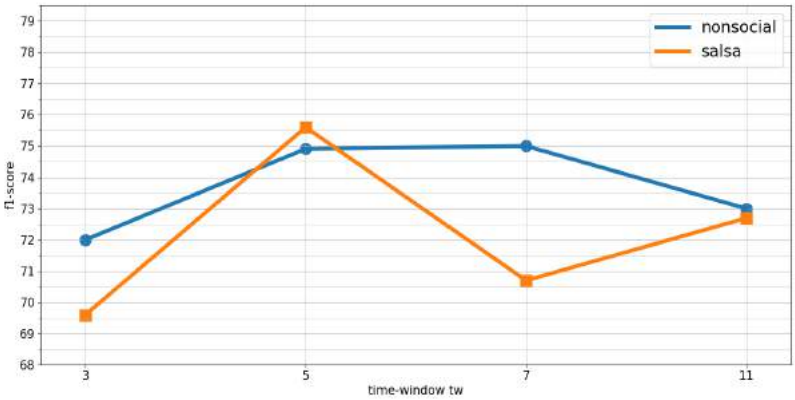


Figure 7.6: Time-window parameter selection. Selecting positive samples in a large time-window range, (e.g. $tw = 7, 11$), is not beneficial on both datasets.

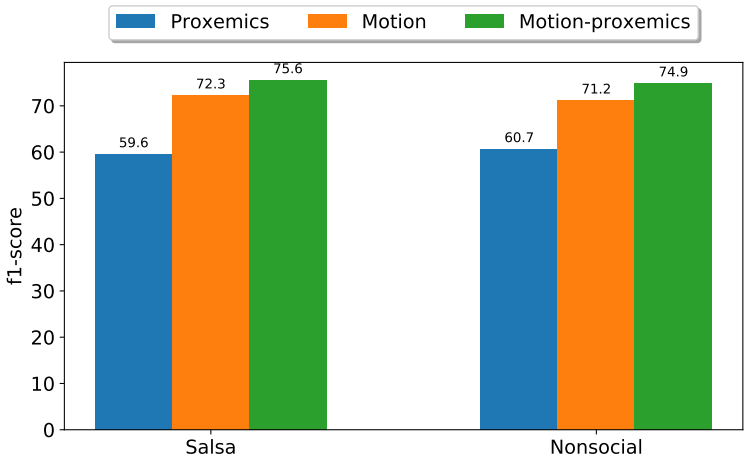


Figure 7.7: Ablation study to evaluate the contribution of the input features. The framework is trained solely on motion, proxemics, and on both motion and proxemics features.

interactions have a short duration, and therefore, highly informative samples have to be selected from a shorter time-window (e.g. $tw = 5$). On the other hand, the Salsa dataset [27] contains interactions from a poster session and a cocktail party in a university environment. As subjects are engaged in social interactions, movements are slower and the impact of longer time-windows is less visible. Given the results, $tw = 5$ is selected for the rest of the evaluation in both datasets.

7.5.5. ABLATION STUDY

An ablation study was conducted to verify the contribution of the chosen input descriptors. In this experiment, the framework is trained and evaluated using a single input descriptor per time (i.e. skeleton motion or proxemics features), or using the combination of the two cues as displayed in Fig. 7.2 on the personality types labels.

Results are reported in Fig. 7.7. Observing the results obtained using the features independently, the skeleton motion input confirms the findings in the literature and in the previous chapters (Chapter 5, Section 5.5.3, and Chapter 6, Section 6.3.5) suggesting that bodily expressivity is a strongly informative descriptor for personality understanding .

With regards to the descriptor corresponding to proxemics, although it does not fully capture affective body expressions, it describes the interaction between subjects, and it still constitutes an informative cue for personality analysis (see more detailed discussion in Section 6.5, Chapter 6). According to Fig. 7.7, the combination of the two cues yields the best results, confirming our initial hypothesis and, thus, it will be used in the rest of our experiments.

7.5.6. COMPARISON WITH BASELINE TECHNIQUES

We compare our performance against various state-of-the-art results for both the Salsa and the Nonsocial dataset. For a fair comparison, Table 7.2 is organized according to the input features.

The first part of the table indicates the performance of methods that use solely skeleton motion features. In particular, PR_{cl}, proposed in Chapter 5 uses an Autoencoder-LSTM framework to learn skeleton motion dynamics. $Clips+MTLN$, proposed by [10], uses skeleton motion descriptors (called clips) that are similar to our motion descriptors. Clips are fed into a CNN framework called MTLN, which processes all frames of the generated clips in parallel to incorporate spatial structural information for action recognition. EL-LMKL [28] was proposed for leadership recognition as well as personality trait recognition using optical-flow motion information. As El-LMKL uses optical flow-based features, it cannot be applied to the Nonsocial dataset where the image data is not available.

In the second part of Table 7.2, we report the performance of baselines methods that use motion-proxemics features. In particular, Person-Context CNN, proposed in Chapter 6, maps short-term motion-context descriptors to personality labels using a multi-stream CNN framework. Motion features are extracted for every individual in the scene, and context features are extracted in terms of proxemics distances as well as social groups interactions.

Additionally, to evaluate the effect of each term in our objective function L_{TISML} (equation 7.4), we also train the model with individual loss functions, L_{DML} and L_{Ident}, separately. The results of this evaluation are indicated as "TISML L_{DML}" and "TISML L_{Ident}". Note that to evaluate the output of "TISML L_{DML}" we use the conventional K-Nearest Neighbor (*KNN*) classifier with Euclidean distance on the generated embeddings ($\hat{x}_s^t$). We set $K = 5$, please see Section 7.5.8 for the investigation of different KNN values.

Results show that the proposed TISML framework achieves higher results in all the tested feature settings. Specifically, we achieve higher results compared to the state-of-

Table 7.2: F1 score on the personality recognition task using different features of TISML compared to baselines and other approaches.

Feature Type	Method	Salsa	Nonsocial
Motion	PR_{cl} [13]	59.6%	55.3%
	EL-LMKL [28]	61.2%	-
	Clips+MTLN [10]	68.5%	70.7%
	TISML Skeleton motion (ours)	**72.3%**	**71.2%**
Motion-Prox	Person-Context CNN [14]	73.0%	72.6%
	TISML L_{DML} (ours)	68.2	67.7
	TISML L_{Ident} (ours)	73.2	73.0
	TISML (ours)	**75.6%**	**74.9%**

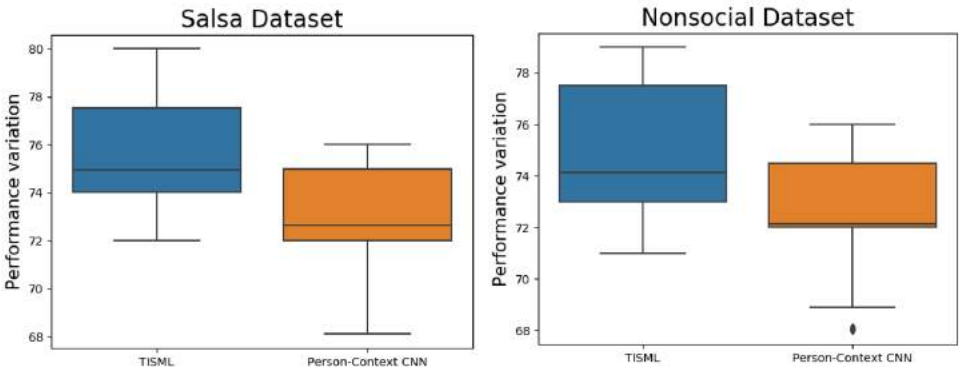


Figure 7.8: The F1 score distributions of TISML and Person-Context CNN [14] methods obtained using the different experimental folds in the personality recognition experiment.

the-art models that use only motion by 3.8% for the Salsa dataset and by 0.5% on the Nonsocial dataset. When using skeleton motion and proxemics, we improve the personality recognition state-of-the-art results by 2.6% on the Salsa dataset, and by 2.3% on the Nonsocial dataset.

Furthermore, the TISML trained using a double objective loss reaches higher performance results than when trained using individual signals. This shows that using a double term is beneficial to create more informative embeddings leading to better recognition performance on both datasets.

Finally, in Fig. 7.8, we investigate the performance range/variation obtained in the experimental folds, where each fold contains a set of different subjects used as test data.

The F1 score distributions of the two best methods (i.e. TISML and Person-Context CNN [14]) in the considered datasets are displayed in box plots. Results indicate that the TISML F1 scores on the experimental folds are significantly higher (p-values less than 0.05) than the Person-Context CNN [14] scores.

7.5.7. PERSONALITY TRAITS RECOGNITION

In order to use the traits scores as classification labels, following the approach used in [29], the score of each trait is transformed into a binary value (HIGH or LOW), based on the median value computed from the population. Note that every trait is evaluated independently from the scores of the other traits.

Figure 7.9 and Figure 7.10 show the prediction results of the proposed method, compared to the Person-Context method introduced in Chapter 6. The Person-Context method maps short-term motion-context descriptors to personality labels using a multi-stream CNN framework. As the two frameworks use similar features (i.e. human motion and context information) the results display the same trend for all the traits. Overall, TISML reaches better accuracy in all the traits proving the benefit of learning behaviors similarity via DML.

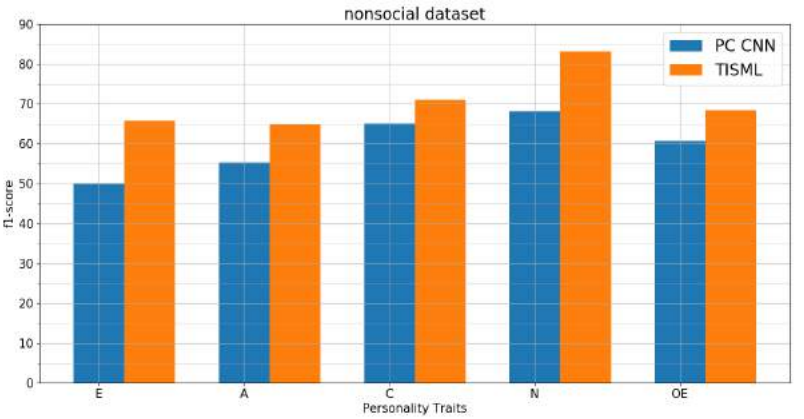


Figure 7.9: Personality Traits prediction using Person-Context CNN (Chapter 6) and the proposed method TISML on the Nonsocial dataset.

In the Nonsocial dataset, the best prediction accuracy is obtained on the Neuroticism trait, whereas the lowest result is obtained on the Extraversion trait. As this dataset contains behavioral data of participants involved in problem-solving tasks, high-low Neuroticism can be a factor influencing the way they behave. On the other hand, as these tasks are performed individually, the Extraversion trait can be hard to express. In the Salsa dataset, the best prediction accuracy is obtained on the Agreeableness trait, whereas the lowest result is obtained on the Conscientiousness trait. As this dataset contains behavioral data of participants involved in social events in an university, high-low Agreebleness trait can be a factor influencing the participants' way of meeting and dis-

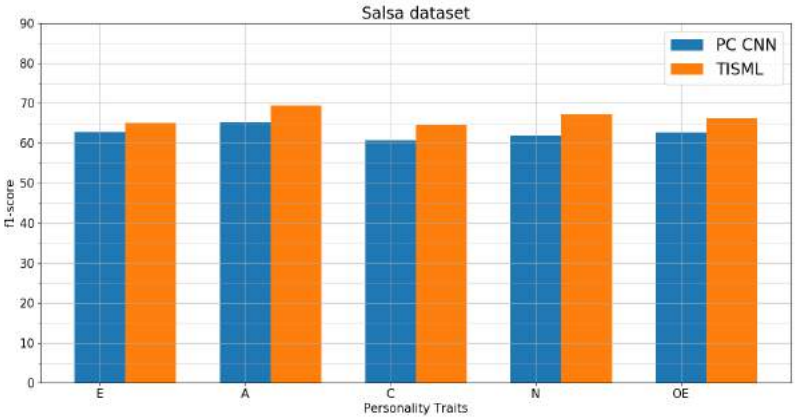


Figure 7.10: Personality Traits prediction using PC CNN [14] and the proposed method TISML on the Salsa dataset.

cussing. This result is aligned with the results presented in [27], where the authors found a significant correlation between the Agreeableness trait and group conversations.

7.5.8. KNN CLASSIFIER INVESTIGATION

In the previous subsection (Section 7.5.6), we showed that the results obtained using the combination of the two losses yielded better results than using independent loss signals. In this section, we perform an experiment to evaluate the discriminative power of the embedding space created using our DML strategy.

As TISML belongs to the DML domain, we aim to investigate whether embeddings belonging to the same personality class are placed close to each other, while embeddings belonging to different personality classes are placed further away. We can evaluate this by using the conventional K-Nearest Neighbor (*KNN*) classifier with Euclidean distance. By varying the number of nearest neighbors K, we can evaluate the structure of the embedding space and the distances between samples with respect to their personality labels.

Fig. 7.11 shows the results of several K for the personality recognition task on the analyzed datasets. Good performance is obtained when K is set to higher values, in a range between $[25, 100]$. For example, at $k = 50$, we obtain 73.5% and 73.3% f1-score, for both, Salsa and Nonsocial datasets, respectively. These results show that the embedding space has been optimized to separate well the samples according to the personality labels.

7.6. CONCLUSIONS

In this chapter, we proposed a framework for automatic personality recognition that is able to embed different behavioral dynamics evoked by diverse real world scenarios. Specifically, motion features were designed to encode local motion dynamics from the human body, and interpersonal distance (proxemics) features were designed to encode

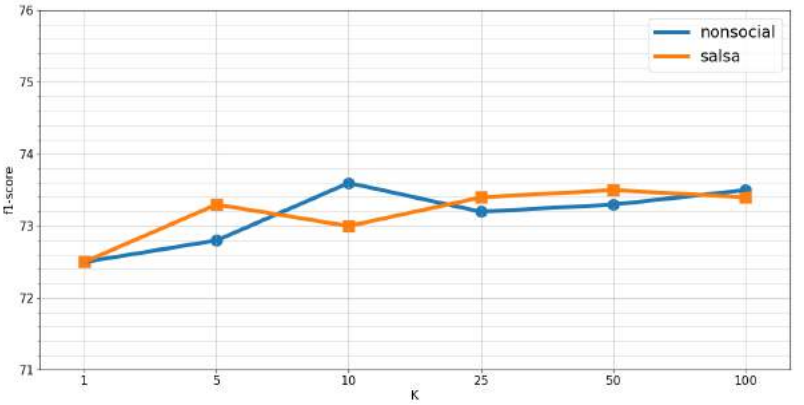


Figure 7.11: Evaluation of the discriminative value of the generated embeddings using a knn classification.

global dynamics in the scene. By using a Convolutional Neural Network (CNN) architecture which utilized a triplet loss Deep Metric Learning (DML), we learned temporal, as well as discriminative spatio-temporal streams of embeddings to represent patterns of personality behaviors. The learning task was accomplished using the novel TISML component. In TISML, a Temporal Triplet Mining (TTM) strategy was employed to leverage the similarity between temporally adjacent short-term descriptors as they are likely to belong to the same semantic behavior and, thus, have higher chances to lead to robust modeling of personality labels. Finally, a double term objective function was used for personality recognition and personality retrieval tasks. Experiments showed that our framework generated embeddings that exploit the semantic similarity of samples in the temporal vicinity. In other words, our latent space found subclusters of temporally related samples that are also semantically related (i.e. they are likely to belong to the same activity). Empirical experiments showed that TISML discovered more meaningful behavioral patterns that improve the state-of-the-art results. Moreover, as these sequences contain a higher semantic value, they are easier to compare with respect to short-term spatio-temporal descriptors, facilitating the discovery of critical behavioral patterns linked to personality descriptions.

REFERENCES

[1] K. Loewenthal and C. A. Lewis, *An introduction to psychological tests and scales*. Psychology press, 2018.

[2] H. Coskun, D. Joseph Tan, S. Conjeti, N. Navab, and F. Tombari, "Human motion analysis with deep metric learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 667–683, 2018.

[3] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, "Deep metric learning to rank," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1861–1870, 2019.

[4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*, pp. 850–865, Springer, 2016.

[5] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[6] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, pp. 1857–1865, 2016.

[7] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2012.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

[9] G. Bradski, "The OpenCV Library." https://opencv.org/, 2000.

[10] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3288–3297, 2017.

[11] A. Aristidou, D. Cohen-Or, J. K. Hodgins, and A. Shamir, "Self-similarity analysis for motion capture cleaning," in *Computer Graphics Forum*, vol. 37, pp. 297–309, Wiley Online Library, 2018.

[12] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino, "Towards computational proxemics: Inferring social relations from interpersonal distances," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 290–297, IEEE, 2011.

[13] D. Dotti, M. Popa, and S. Asteriadis, "Behavior and personality analysis in a nonsocial context dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2354–2362, 2018.

[14] D. Dotti, M. Popa, and S. Asteriadis, "Being the center of attention: A person-context cnn framework for personality recognition," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 3, pp. 1–20, 2020.

[15] T. L. Chartrand, G. M. Fitzsimons, and G. J. Fitzsimons, "Automatic effects of anthropomorphized objects on behavior," *Social Cognition*, vol. 26, no. 2, pp. 198–209, 2008.

[16] M. L. Walters, K. Dautenhahn, R. Te Boekhorst, K. L. Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, and I. Werry, "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pp. 347–352, IEEE, 2005.

[17] A. Aristidou, D. Cohen-Or, J. K. Hodgins, Y. Chrysanthou, and A. Shamir, "Deep motifs and motion signatures," in *SIGGRAPH Asia 2018 Technical Papers*, p. 187, ACM, 2018.

[18] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, pp. 1988–1996, 2014.

[19] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–769, 2016.

[20] F. Chollet *et al.*, "Keras." https://keras.io/, 2015.

[21] M. Abadi, A. Agarwal, P. Barham, *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems." http://tensorflow.org/, 2015. Software available from tensorflow.org.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[24] J. H. Block and J. Block, "The role of ego-control and ego-resiliency in the organization of behavior," in *Development of cognition, affect, and social relations: The Minnesota Symposia on child psychology*, vol. 13, pp. 39–101, 1980.

[25] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[26] J. C. S. J. Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. Van Gerven, R. Van Lier, *et al.*, "First impressions: A survey on vision-based apparent personality trait analysis," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[27] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "Salsa: A novel dataset for multimodal group behavior analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1707–1720, 2016.

[28] C. Beyan, M. Shahid, and V. Murino, "Investigation of small group social interactions using deep visual activity-based nonverbal features," in *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 311–319, ACM, 2018.

[29] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pp. 37–42, ACM, 2010.

8

CONCLUSIONS AND FUTURE RESEARCH

This thesis investigated how automatic human behavior understanding using sensory data can support real-world applications in the fields of AAL, surveillance and HCI. As nonverbal communication (e.g. body postures, facial expressions as well as gestures) conveys rich information about behaviors, in this thesis, we proposed several novel methods to extract, learn, and visualize meaningful patterns of human behaviors. Specifically, we focused on human body posture and motion analysis making use of RGB and depth information from video data.

We are in the midst of a wave of technological innovations that are revolutionizing several aspects of our life. This advancement creates infinite opportunities to investigate the use of smart applications to help tasks that until now required manual and tedious human interventions. In this direction, this thesis investigated how the understanding of human motion can be used to automatize certain tasks in the fields of AAL, surveillance and HCI.

This chapter provides the conclusions of this thesis. In Section 8.1 we answer the research questions posed in the introduction, while recommendations for future research are given in Section 8.2.

8.1. ANSWERS TO THE RESEARCH QUESTIONS

Five research questions were formulated in Chapter 1 concerning different aspects of human behavior understanding. These inquires guided my research throughout the PhD study, and, in the following subsections, we explain the implications of our findings.

8.1.1. RESEARCH QUESTION 1: HOW CAN MOTION TRAJECTORIES BE LEVERAGED FOR THE DISCOVERY OF NORMAL AS WELL AS ABNORMAL BEHAVIORAL PATTERNS?

In Chapter 3, we proposed a novel system for detecting normal and abnormal human behaviours from trajectory data. Analyzing motion trajectory data has several advantages. As the majority of surveillance settings must cover wide areas, the captured information can pose several challenges, including low resolution videos or even total absence of video signals (e.g. radar signals can replace video cameras in some fields) [1]. In these scenarios, it is difficult to compute complex behavioral features like gestures, body poses or appearances. Hence, the most valuable information is the positions of objects recorded along the tracks (i.e. trajectories). In Chapter 3, we investigated various temporal features in combination with spatial information to create spatio-temporal insights, describing objects motions in the scene. Furthermore, we obtained an improved and more efficient feature representation by applying an Autoencoder Neural Network on top of the trajectory features, which, through experiments, showed to be useful for representing the underlying motion distribution.

Another characteristic that is common in surveillance systems is the abundance of unlabeled as well as noisy data. Data is automatically recorded every day, and it usually includes only 1% of meaningful events. The process of manually discovering these rare events can be expensive and time-consuming. Therefore, smart surveillance systems leveraging unsupervised methods for motion analysis can drive important advancements in the surveillance field, helping to reduce the cost and optimizing human efforts. In this direction, in Chapter 3, we proposed an unsupervised approach for obtaining data annotations by performing clustering on the extracted features. Our result was a map of the environment organized in spatio-temporal motion clusters. With this map, we aimed to simplify the labeling process as the final users of the system will have to label k activity patterns instead of all the individual trajectories. Furthermore, this process can be very useful when the system needs to be deployed in different environments and the labelling task has to be fast and generalized. One important aspect of our work lies in the flexibility and generalization ability of the proposed system. Hence, we performed experiments on both indoor and outdoor surveillance data.

Smart-camera systems have been mostly applied in public environments (e.g. train stations, public squares, and roads). Due to important privacy concerns about video data in private homes, few indoor smart systems have been commercialised. However, with an increasingly growing population in Europe, and the willingness of the elderly to live independently in their home as much as possible, smart monitoring systems are called for a great opportunity. In this direction, in the second part of Chapter 3, we investigated our spatio-temporal model for the detection of abnormal behaviors (connected to various diseases including Dementia and Alzheimer's). As there exist few datasets concerning the detection of abnormal trajectories in indoor private spaces, we proposed a novel dataset where we recorded 19 participants performing 6 tasks that simulate home activities (e.g. searching for keys, making tea). This dataset was also used in the context of a European H2020 program (Grant Agreement $N°$ 690090, ICT4Life project [1]), for in-

[1] http://www.ict4life.eu/

vestigating and training the model on abnormal trajectories that could happen in real world AAL scenarios. The applicability of our proposed framework could enhance the prevention mechanisms and contribute to an increased feeling of security supporting the older population suffering from Dementia and Alzheimer's diseases.

8.1.2. RESEARCH QUESTION 2: HOW CAN MOTION TRAJECTORIES BE LEVERAGED FOR REAL-TIME SURVEILLANCE APPLICATIONS?

Real-time analysis of motion intervals is a critical step for the understanding of objects mobility. We are living in exciting times for the automation of several tasks that involve objects' navigation. Self driving objects (i.e. cars, robots, trucks) are being tested in real-world scenarios, collecting data to train their models on the uncertainties of real-life situations. In this thesis, we explored real-time prediction of future objects' trajectories in outdoor as well as indoor scenarios.

The degree of complexity that concerns objects' trajectories in public spaces is very high. Especially us humans, we can make a great range of movements in a split of a second to reach a desired destination. Human motion is affected by internal needs, for example, going to the supermarket to do grocery shopping, as well as by external environmental factors, for example, obstacles along the way. In Chapter 4, we proposed a hierarchical framework for modeling motion trajectories in real-time. The hierarchical architecture was designed to capture short spatio-temporal trajectory patches in the lower levels. Short motion patches are highly variant, containing short deviations and turns caused by moving obstacles all around. Then, in the higher levels, we combined motion patches using a grid structure where we smoothed short deviations of individual patches and learned longer and more meaningful spatio-temporal motion trajectories. Finally, we modeled temporal motion transitions using Bayesian probability, by inferring the future trajectory step given the current motion information.

The effectiveness of our approach was demonstrated in both indoor and outdoor public spaces for short-term as well as long-term path prediction. Furthermore, the presented approach was also suitable for abnormality detection, computed using the likelihood of the conditional probabilities. As every object has different characteristics and rules, we also tested the model on the ability to discriminate between different classes of moving objects (i.e pedestrians vs. cars) using trajectory data. We believe it is extremely important for our society to improve and automate surveillance systems to help the operators to prioritize meaningful events. In this direction, our proposed model was able to predict the next trajectory step as well as detect when a trajectory is moving in an unexpected way. These types of automatic systems are important to direct the attention of the operators occupied in managing the surveillance systems, often consisting of dozens of cameras.

8.1.3. RESEARCH QUESTION 3: POSTURE SEQUENCE MODELLING AND AFFECTIVE COMPUTING: WHAT CAN WE AUTOMATICALLY LEARN ABOUT PERSONALITY USING BODY POSTURES?

"Human groups are characterized by a constant flow of verbal and nonverbal communication; people are talking, listening, smiling, gesturing, touching, and laughing." [2].

Unlike verbal communication, which is based on languages that have a well studied semantic and syntactic structures, nonverbal communication is lacking a generic underlying architecture [3]. Human body postures and gestures are two of the most important forms of nonverbal communication. They include movements of hands, head and other parts of the body that allow individuals to communicate a variety of feelings, thoughts and emotions [4].

In this thesis, we investigated the use of deep learning frameworks for skeleton-based posture extraction and understanding. Specifically, in Chapter 5, we introduced a novel approach to learn upper body posture representations using an Autoencoder Neural Network. As the temporal evolution of human behaviors plays a critical role when it comes to decoding the semantic meaning of behaviors, we proposed the use of LSTM networks for learning behavior dynamics. Additionally, by adding a classification layer on top of the LSTM output, the mapping between behaviors and personality was performed.

Skeleton-based behavior understanding has received great attention due to its wide range of applications in smart home environments, HCI, and surveillance. One of the great advantages of using skeleton information is that it records only the position of the joints in time. As no image data and no personal data is saved, the privacy of the users is not invaded. Therefore, we extended the "Multimodal dataset for abnormal behavior understanding" by recording and analyzing the full body movements of 46 participants performing 6 tasks in an indoor environment.

In Chapter 5, we demonstrated that this data can provide useful insights on the study of human motion trajectories as well as for detecting affective human behaviors. While observing participants performing problem-solving tasks under pressure, we noticed that our participants expressed behavioral and interaction patterns that were worth to explore more in-depth. In this direction, personality psychology provides several models that aim at categorizing how we, as individuals, tend to behave in ways that are broadly consistent over time. When mapping computational models and features to personality theories, we highlighted the need of interdisciplinary research, between psychology, computer vision and affective computing, for enhancing the ability to understand and automatically recognize human personality at a deeper level.

8.1.4. RESEARCH QUESTION 4: ARE CONTEXTUAL CUES INFORMATIVE PREDICTORS IN ADDITION TO POSTURE FOR PERSONALITY RECOGNITION?

As shown by several psychological researches, humans tend to show consistent behavioral patterns independently from the situation or the geographical location [5]. However, due to external factors like contextual rules and social environments, it is extremely challenging to build general systems able to detect these patterns in diverse scenarios. In this thesis, we investigated the use of contextual information, merged together with individual motion features, for a more robust affective computing and behavior understanding system.

In computer vision, contextual information has been shown to improve several challenging tasks such as action recognition [6] and social scene understanding [7]. Building on these findings, in Chapter 6, we proposed a framework which mapped the mutual

relation between individual behaviors and contextual information to personality labels.

In Chapter 6, we presented a novel CNN-based framework for personality recognition. Our model analyzed the scene at multiple levels of granularity: firstly, we encoded spatio-temporal descriptors for each individual in the scene. Secondly, we extracted spatio-temporal descriptors from social groups, and thirdly, we encoded the global proxemics of every individual in the scene. Additionally, we demonstrated that the proposed proxemics features can be applied also in a nonsocial scenario, encoding scene interaction information. Experimental results demonstrated the effectiveness of our approach, showing that modeling together Person-Context information significantly improves the individual features on personality recognition tasks.

The Analysis of Person-Context interactions provides useful information for several real-world applications including social role understanding and event prediction [8]. One of the main challenges of merging these two sources of information is modeling the variation of social groups that, in most cases, are varying in size and context. We believe that the understating of personality patterns, as an additional source of information, can help in the construction of deeper social group models, and vice-versa. For example, imagine a social scenario in which different interpersonal styles influence the group interaction. Individuals that strive to be in the center of the conversation will try to actively capture the attention of the rest of the group, resulting in the group acting more passively [9]. By combining information at the individual level with information at the context level, a robust semantic understanding of the situation is perceived. Even if individuals are in a nonsocial scenario (e.g. when an individual is alone at home), it has been shown that people engage with contextual objects as they would engage with other humans (Anthropomorphism) [10]. Therefore, by examining the interaction between human behaviors and the surrounding scene, important information about human personality can be extracted.

During the past decades, several personality models have been proposed and widely studied. In this thesis, we focused on two popular models among psychologists: the Big-Five personality traits [11], and the three personality types [12]. The Big-Five traits model splits the behaviors in five traits, i.e. Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. These traits are often considered and evaluated independently, without addressing their connections. In contrast, numerous studies confirmed the theory proposed in [12], according to which personality traits can be organized in three major types: Resilient, Undercontrolled and Overcontrolled (see Chapter 5 and Chapter 6). In order to avoid being biased towards one personality model, in this thesis, we tested our frameworks using labels coming from both models. Although the Big-5 model is more popular among affective computing specialists, the model in [12] can capture the interplay among individual personality traits. It is our belief that attempting to establish unique combinations of personality traits may be more useful for real-world applications such as automatic job-screening or HCI. These combinations can be more similar to how individuals describe others, for example, if we describe “George” as being friendly, friendliness should have a high score on both the Agreeableness and Extraversion scales. On the contrary, an individual with high Agreeableness but low Extraversion will tend to be perceived as docile or conformist [13].

8.1.5. RESEARCH QUESTION 5: DOES MODELLING THE TEMPORAL NATURE OF HUMAN BEHAVIORS IMPROVE LATENT REPRESENTATIONS AND CONSEQUENTLY PERSONALITY RECOGNITION?

Comparing and categorizing behavioral patterns for personality understanding is a challenging task due to the multiple representations the same behavior can acquire. For example, the same behavior can be performed in different way due to the context (i.e. social versus nonsocial), or due to individuals' internal drives (e.g. personality, mood, emotion, goals, cognitive state). A simple action like walking, can be expressed in several different ways depending on the attitude/personality (i.e. nervous versus calm personality), or depending on the scenario (i.e. crowded versus empty scenario). In this thesis, we aimed at expanding the use of body motion as well as context information to learn their dynamic interaction in time.

In Chapter 7, we proposed a novel model that leverages the similarities between temporally adjacent short-term descriptors as they are likely to belong to the same semantic behavior. This learning process is carried out using a novel strategy, in the field of Deep Metric Learning (DML), called Temporal Triplet Mining (TTM). Effectively measuring the similarity between two human motions is a non-trivial problem as human poses have to be compared across a temporal set of frames. This introduces several challenges such as alignment as well as pose to pose comparison. Therefore, our intuition was to select temporally related positive examples to encourage the DML model to generate embeddings with temporal relation while maintaining a high discriminative power for personality recognition. We found that learning the temporal similarity allowed the model to assemble longer sequences that contain higher semantic value than the input features. Moreover, as we did not add any constraint on the temporal relations, the model automatically learned sequences of varied temporal durations.

Real-world applications often involve great amounts of noisy as well as sparse data, that requires a lot of resources for manual annotations (e.g. actions or events). For example, behavior understanding designed as a supervised task, involves an important effort in segmenting the actions in time (i.e. finding a start and an end), as well as defining the semantic label of given behaviors. On the contrary, in Chapter 7, we adopted a self-supervised learning strategy [14], where the greatest amount of supervision is given by the *data*, and not by human annotators. In this way, we observed that, by selecting temporally related samples in our DML strategy, we encouraged our framework to generate embeddings with temporal relation without any human labels.

Effectively measuring the similarity between human behaviors is a critical task for several applications in the fields of Surveillance, HCI, as well as Healthcare. Moreover, by adding the affective component into the model, we aimed to be on a crossroad between standard computer vision algorithms and social psychological studies. This effort can lead, in the long-term, to socially intelligent surveillance and monitoring technologies [15], and the role of this research contributes to placing the foundations towards this goal.

8.2. FUTURE RESEARCH

Behavior analysis. Nowadays, humans are transferring a large amount of their activities (i.e. work, social, hobbies) on their computers and on the internet. Currently, computational systems are becoming a ubiquitous presence in our life. As humans are spending more and more time with these technologies, building intelligent, adaptable and individualized systems is more necessary than ever. By detecting and adapting to the users' personalities and emotions, there is a great potential for introducing novel user experiences.

As more and more data regarding digital activities is available, it is possible to extract targeted and more detailed features for user behavior understanding. For example, there is an abundance of data relating to online shopping behaviors, data about sport activities, social interactions and so on. However, even though the activity channels are changing, we must not forget that the center of our study has to be the *user*. Only by using and integrating the knowledge acquired from studying humans' behaviors from multiple perspectives, can we enhance technology-based user experience.

Psychologists have long studied human behaviors and personality, and throughout the years different theories have been proposed to explain and understand them [16]. In this thesis, we mainly worked with the Big Five Traits model [11], and the three personality types model [12]. The main criticism against the use of personality trait models is that they are purely descriptive and do not correspond to actual characteristics of individuals [16]. Another debate on personality theory is about self-impressions versus the personality attributes that other individuals perceive. In this thesis, we based our studies on self-reported questionnaires [17], as this annotation approach was the most commonly used in the affective computing literature.

Future research must aim at integrating multidisciplinary inputs to the behavior and personality understanding problems. All the behavioral cues that may show the true personality of individuals have to be analyzed from multiple sides, and only in this way, technology could achieve a better and complete understanding of human behaviors.

Behavior analysis in Healthcare. The experience gained in an ambitious H2020 European project such as ICT4Life is very relevant for the advancement of smart technologies in healthcare. Machine Learning researchers, medical doctors, and professionals gathered and investigated together practical solutions to improve the life of seniors affected by Dementia and Alzheimer's diseases. The involvement of smart technologies in supporting disabled people requires the study of several critical challenges. Privacy, accuracy of the sensors, reliability of the technology and collaborations between different professionals were the key messages learnt from this experience.

Future research in this domain must aim at keeping always the disabled users at the center of the design. Users should feel comfortable to have the technologies installed in their home, and the technologies should be designed to support them. Another great challenge is the collaboration between professionals from different fields. Multidisciplinary projects like ICT4Life require full-time involvement of all the parts to design platforms able to support and improve life conditions.

Computer Vision and Artificial Intelligence. The increasing amounts of video and image data and the advances in Deep Neural Networks have made computer vision one of the most important research areas in artificial intelligence. As we saw in this thesis,

video surveillance is one of the areas where computer vision can have a great impact. Tackling challenges like multi-camera and multi-person tracking as well as person re-identification could enable the fully automated use of intelligent video surveillance systems in public spaces, like train stations and airports. Additionally, in this thesis, we covered behavior understanding using trajectory as well as human posture data. Future works should focus on the fusion of multiple data sources such as face, audio, biometrics signals to allow the algorithms to capture meaningful and complete behavioral patterns. In this regard, strategies like Deep Metric learning showed to have great abilities into fusing different data sources to create a common embedding space optimized for behavior and personality recognition. Finally, to make more scalable and adaptive computer vision solutions, future works should focus on reducing human labeling components, i.e., targeting more the unsupervised as well as semi-supervised learning strategies. As the cost associated with the labeling process is often large, in several domains, fully labeled datasets become infeasible, whereas, the acquisition of unlabeled data is relatively easier. In such situations, unsupervised as well as semi-supervised learning can be of great practical value for the expansion of computer vision applications in real-world scenarios.

REFERENCES

[1] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models," *International journal of computer vision*, vol. 95, no. 3, pp. 287–312, 2011.

[2] G. Håkansson and J. Westander, *Communication in humans and other animals*, vol. 4. John Benjamins Publishing, 2013.

[3] L. C. G. F. d. Santos, *Laban Movement Analysis: A Bayesian Computational Approach to Hierarchical Motion Analysis and Learning.* PhD thesis, 2014.

[4] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, pp. 1–1, 2018.

[5] D. P. Schmitt, J. Allik, R. R. McCrae, and V. Benet-Martínez, "The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations," *Journal of cross-cultural psychology*, vol. 38, no. 2, pp. 173–212, 2007.

[6] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929–2936, IEEE, 2009.

[7] T. M. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition.," in *CVPR*, pp. 3425–3434, 2017.

[8] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3048–3056, 2017.

[9] M. Snyder, J. A. Simpson, and S. Gangestad, "Personality and sexual relations.," *Journal of Personality and Social Psychology*, vol. 51, no. 1, p. 181, 1986.

[10] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism.," *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[11] B. Rammstedt, "The 10-item big five inventory," *European Journal of Psychological Assessment*, vol. 23, no. 3, pp. 193–201, 2007.

[12] J. H. Block and J. Block, "The role of ego-control and ego-resiliency in the organization of behavior," in *Development of cognition, affect, and social relations: The Minnesota Symposia on child psychology*, vol. 13, pp. 39–101, 1980.

[13] F. A. Sava and R. I. Popa, "Personality types based on the big five model. a cluster analysis over the romanian population.," *Cognitie, Creier, Comportament/Cognition, Brain, Behavior*, vol. 15, no. 3, 2011.

[14] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8052–8060, 2018.

[15] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino, "Towards computational proxemics: Inferring social relations from interpersonal distances," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 290–297, IEEE, 2011.

[16] J. Junior, C. Jacques, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. van Gerven, *et al.*, "First impressions: A survey on computer vision-based apparent personality trait analysis," *arXiv preprint arXiv:1804.08046*, 2018.

[17] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

IMPACT PARAGRAPH

In this addendum, the scientific as well as social impact of the presented work is discussed. The paragraph below addresses the drafted four questions given in the "Regulations for obtaining the doctoral degree Maastricht University" [1].

(Research) What is the main objective of the research described in the thesis and what are the most important results and conclusions?

This thesis investigated how automatic human behavior understanding from video data can support real-world applications in the fields of Ambient Assisted Living (AAL), Surveillance and Affective Computing (AC). Unlike language, which has well studied semantic and syntactic structure, human behaviors are lacking a generic underlying architecture [2]. As a matter of fact, human behaviors are generated in different forms and levels of complexity. In this thesis, we aim to study human behaviors in different environments by interpreting how body motion and body postures evolve over time. Specifically, in Chapter 3 and Chapter 4, we build two novel frameworks that extract and interpret spatio-temporal motion features using trajectory data. We demonstrate that by learning the spatial as well as the temporal distribution of the trajectory points, we can detect abnormal behaviors and predict in near real-time what will happen next. These models showed to be beneficial in surveillance applications, such as detecting abnormal events in public environments like train stations or public squares, and in AAL applications, such as detecting dangerous events in elderly affected by dementia and Alzheimer's disease. In Chapter 5, Chapter 6, and Chapter 7, we build novel frameworks to encode body postures and their interaction with context information over time. Our main objective in these chapters is to learn stable behavioral patterns defined as *Personality patterns*. Our research highlights that integrating computational models with psychological theories, such as the Big-5 personality traits, can help the interpretation of human behaviors in social as well as nonsocial environments.

(Relevance) What is the (potential) contribution of the results from this research to science, and, if applicable, to social sectors and social challenges?

We are in the midst of a wave of technological innovations that are revolutionizing several sectors of our society. However, the integration and the automation of new technologies in our society remains a fundamental challenge to be tackled. In this thesis, we study the automatic understanding of human behaviors and its applications in the fields of Ambient Assisted Living, Surveillance and Affective Computing.

As the average population age in Europe, and in the world in general, is steadily increasing, new challenges surrounding the economic burden of having more old people than young are emerging. In this context, smart and automatized healthcare applications made with low-cost sensors could reduce the economic impact while improving the living conditions of the old population [3]. In this thesis, we investigate the detection of abnormal events such as confusion and repetitive behaviors that impact elderly affected by dementia and Alzheimer's disease. In Chapter 3, we propose a dataset which

encourages the analysis of human behaviors in unrestricted settings for the discovery of abnormal patterns from spontaneous activities. Datasets containing abnormal agitation and confusion behaviors from video data are rare to find as well as difficult to collect. Hence, with the help of medical professionals, in this dataset, we designed certain activities which might provoke trajectory data similar to the ones created by people in confusion states. The dataset is used to design and train a machine learning model that distinguishes between normal and abnormal activities. Finally, by applying a transfer learning strategy, we test the model directly in hospitals with four patients affected by Alzheimer's disease. This work is part of the ICT4Life European project, which aimed to implement a platform integrating a series of innovative services, targeting aging people with cognitive impairments.

Affects and emotions are fundamental human experiences that have a great impact on humans' lives, choices, well-being, and so forth. Consequently, the ability to automatically recognize and interpret affective attributes and emotional patterns can have a huge impact on several sectors in our society. Specifically, in this thesis, we study how human body postures, interactions, and behaviors can be automatically mapped to personality labels. As human beings, we are able to interpret affective states of other individuals from little information. Therefore, the obtained results are of great importance for future interdisciplinary behavioral studies, aiming to combine data-driven approaches with psychological studies to enhance the understanding of human behavior from machines' perspective.

(Target group) To whom are the research results interesting and/or relevant? And why?

The primary target group of the presented studies is the researchers in Computer Vision and the field of AI in general. Five novel frameworks have been introduced, from Chapter 3 to Chapter 7. Quantitative as well as qualitative experiments are carried out to investigate the efficacy of our methods on public datasets against state-of-the-art computer vision algorithms.

The secondary target group is the potential users of the AI applications presented in this thesis and are described below.

8

For the users in the healthcare field, we presented the ICT4Life platform (Chapter 3), which aimed at providing innovative ICT services targeting the aging population. The ICT4Life platform includes AI algorithms to monitor, detect, and prevent dangerous events supporting the old generation to live independently for as long as possible. Nevertheless, the aging process brings several difficulties that affect not only the old generation, but also the elder's family and healthcare professionals. In this context, the ICT4Life platform focuses on the integration of different information (sensory information as well as medical files) to provide high level insights on the patient's health condition. These insights are accessible to all the users of the platform, i.e. patient, carers, medical staff.

For the users in the surveillance field, long and unconstrained security footage is hard to monitor due to limited human resources. The ability to hold attention and to react in case of rarely suspicious events is demanding and prone to human error [4]. Thus, the presented work in Chapter 3 and Chapter 4 which differentiate important events versus unimportant events is critical to optimize the limited attention of surveillants, and, at the same time, to alleviate the costs of surveillance systems.

For the users in the affective computing field, we presented three novel frameworks

(Chapter 5, Chapter 6, and Chapter 7) which link body postures and context interactions to personality attributes. The detection and the recognition of human affective cues can be applied in several applications for industries as well as for education. For example, companies can understand their customers better and provide more satisfactory services by pulling emotional strings that are difficult to reach with the standard marketing strategies. Schools and universities can personalize the educational contents by looking and understanding the involvement of the students.

(Activity) In what way can these target groups be involved in and informed about the research results, so that the knowledge gained can be used in the future?

In this thesis, we presented five works that have been published to peer-reviewed international conferences as well as high-impact journals. The work published in peer-reviewed conferences has been displayed in presentation form (Chapter 3 and Chapter 7) as well as poster form (Chapter 5). Furthermore, the work in Chapter 3 was part of the ICT4Life European project [2], and it has been presented during international consortium meetings as well as revisions to the European commission.

REFERENCES

[1] UM, "Regulations for obtaining the doctoral degree Maastricht University," *Regulation Governing the Attainment of Doctoral Degrees*, 2018.

[2] L. C. G. F. d. Santos, *Laban Movement Analysis: A Bayesian Computational Approach to Hierarchical Motion Analysis and Learning.* PhD thesis, 2014.

[3] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: a survey," *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015.

[4] A. Hampapur, L. Brown, J. Connell, S. Pankanti, A. Senior, and Y. Tian, "Smart surveillance: applications, technologies and implications," in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, vol. 2, pp. 1133–1138, IEEE, 2003.

[2] https://cordis.europa.eu/project/id/690090

ACKNOWLEDGEMENTS

Choosing to start a PhD is a very important decision. Working as a PhD implies committing to one specific topic for many years, where it is in your hands to bring novel solutions to unanswered scientific questions. I have always believed that doing a PhD is like running your own business, if you do not prioritize it and work hard by investing extra time to make advancements, no one will do it for you. Fortunately, when I arrived to Maastricht, I immediately felt connected to the city, and most importantly, I was intrigued by the PhD topic as well as its real-world applications in fields like healthcare, surveillance, and affective computing.

Throughout my PhD at Maastricht University I worked with several remarkable researchers that inspired me every day.

I would first like to thank my primary promotor Stelios Asteriadis whose expertise in research as well as in the European project was very valuable for my academic growth. Thank you for being a present supervisor, I enjoyed our bi-weekly meetings which helped me in shaping the research questions and methodology. Your guidance and support made me achieve great objectives like presenting our work at international conferences and in front of the European commission. I will be always grateful for all these experiences which shaped me as a person and researcher.

I would like to thank my co-promotor Mirela Popa for her precious collaboration in the research and in the European project. Thank you for your patient support in both the design and the data analysis of our experiments. Your insights were always useful both in dealing with the technical as well as theoretical problems I faced during my PhD. I always felt that our collaboration was an added value that made us achieve great objectives. I would also like to thank my office-mates, which with the years also became my friends, Enrique Hortal, Esam Galeb, and Christos Athanasiadis. Thanks guys for the wonderful time we had in, and out of the work environment. Coming to the office everyday was a pleasure as we created a nice atmosphere, where we supported each other and laughed together in times of struggle. I will also miss our historical and political discussions during breaks. Furthermore, thank you to all the DKE staff and PhDs who made the work environment very enjoyable.

On the personal side, I would like to first thank my wonderful partner Nofar Ben Itzhak, without her, I would not have reached this important milestone. We always supported each other, brainstormed, and practiced presentations before important events. Thank you for always being there to support me and motivate me.

Grazie anche alla mia famiglia, anche se lontani, vi ho sempre sentiti vicini. Grazie di avermi lasciato libero di prendere le mie decisioni, sostenendomi sempre nel migliore dei modi.

CURRICULUM VITÆ

Dario Dotti was born on July 19, 1989 in Brescia, Italy. In 2008, he started to study for his bachelor's in Communication Technology at the University of Trento, Italy. During this period, he followed courses in Computer Science, Human-Computer Interaction, and Human Cognition which sparked his interest in the multidisciplinary field of Artificial Intelligence. In 2011, he participated in the Erasmus+ Program and spent 6 months at Radboud University, Nijmegen, The Netherlands, where he attended courses of the bachelor's program in Artificial Intelligence. In 2012, he started a master's at the University of Trento in Artificial Intelligence, Language, and Multimodal Interaction. During the master's, he specialized in advanced signal processing and data analysis in research areas such as Natural Language Processing and Computer Vision. After his thesis in the area of Computer Vision and a short experience as a Research Assistant at Inria Sophia Antipolis-Méditerranée Research Centre, he was accepted as a PhD researcher at the Department of Data Science and Knowledge Engineering, Maastricht University, The Netherlands. The research carried out from 2016 to 2019 resulted in several publications in international conferences as well as scientific journals in the area of Computer Vision and Machine Learning. Furthermore, his PhD was funded by the European Union's Horizon 2020 Research and Innovation Program (ICT4Life), which gave him the opportunity to actively collaborate with other universities as well as companies with the final goal of making smart systems for the aging population. Besides performing scientific tasks, he was involved as a teaching assistant in the course of Human-Computer Interaction as well as in supervising students with their thesis.

List of Publications

References

[1] D. Dotti, E. Ghaleb, and S. Asteriadis, "Temporal triplet mining for personality recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 171–178.

[2] D. Dotti, M. Popa, and S. Asteriadis, "Being the center of attention: A person-context cnn framework for personality recognition," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 3, pp. 1–20, 2020.

[3] D. Dotti, M. Popa, and S. Asteriadis, "A hierarchical autoencoder learning model for path prediction and abnormality detection," *Pattern Recognition Letters*, vol. 130, pp. 216–224, 2020.

[4] F. Gibellini, S. Higler, J. Lucas, M. Luli, M. Stallmann, D. Dotti, and S. Asteriadis, "Towards approximating personality cues through simple daily activities," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 192–204, Springer, 2020.

[5] D. Dotti, M. Popa, and S. Asteriadis, "Behavior and personality analysis in a nonsocial context dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2354–2362, 2018.

[6] F. Alvarez, M. Popa, V. Solachidis, G. Hernandez-Penaloza, A. Belmonte-Hernandez, S. Asteriadis, N. Vretos, M. Quintana, T. Theodoridis, D. Dotti, *et al.*, "Behavior analysis through multimodal sensing for care of parkinson's and alzheimer's patients," *IEEE Multimedia*, vol. 25, no. 1, pp. 14–25, 2018.

[7] F. Alvarez, M. Popa, N. Vretos, A. Belmonte-Hernández, S. Asteriadis, V. Solachidis, T. Mariscal, D. Dotti, and P. Daras, "Multimodal monitoring of parkinson's and alzheimer's patients using the ict4life platform," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE.

[8] D. Dotti, M. Popa, and S. Asteriadis, "Unsupervised discovery of normal and abnormal activity patterns in indoor and outdoor environments.," in *VISIGRAPP (5: VISAPP)*, pp. 210–217, 2017.

SIKS Dissertation Series

2011

1. Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
2. Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
3. Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
4. Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
5. Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
6. Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
7. Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
8. Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
9. Tim de Jong (OU), Contextualised Mobile Media for Learning
10. Bart Bogaert (UvT), Cloud Content Contention
11. Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12. Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13. Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14. Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
15. Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16. Maarten Schadd (UM), Selective Search in Games of Different Complexity
17. Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18. Mark Ponsen (UM), Strategic Decision-Making in complex games
19. Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20. Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21. Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22. Junte Zhang (UVA), System Evaluation of Archival Description and Access
23. Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24. Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25. Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26. Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27. Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

28. Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29. Faisal Kamiran (TUE), Discrimination-aware Classification
30. Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31. Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32. Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33. Tom van der Weide (UU), Arguing to Motivate Decisions
34. Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35. Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36. Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37. Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38. Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39. Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40. Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41. Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42. Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43. Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44. Boris Reuderink (UT), Robust Brain-Computer Interfaces
45. Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46. Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47. Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48. Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49. Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

1. Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
2. Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
3. Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
4. Jurriaan Souer (UU), Development of Content Management System-based Web Applications
5. Marijn Plomp (UU), Maturing Interorganisational Information Systems
6. Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
7. Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
8. Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
9. Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10. David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11. J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12. Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13. Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14. Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15. Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16. Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17. Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18. Eltjo Poort (VU), Improving Solution Architecting Practices
19. Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20. Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21. Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22. Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23. Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24. Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25. Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26. Emile de Maat (UVA), Making Sense of Legal Text
27. Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28. Nancy Pascall (UvT), Engendering Technology Empowering Women
29. Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30. Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31. Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32. Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33. Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34. Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35. Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36. Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37. Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38. Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39. Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40. Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41. Sebastian Kelle (OU), Game Design Patterns for Learning
42. Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43. (Withdrawn)
44. Anna Tordai (VU), On Combining Alignment Techniques

45. Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46. Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47. Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48. Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49. Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50. Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51. Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013

1. Viorel Milea (EUR), News Analytics for Financial Decision Support
2. Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
3. Szymon Klarman (VU), Reasoning with Contexts in Description Logics
4. Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
5. Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
6. Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
7. Giel van Lankveld (UvT), Quantifying Individual Player Differences
8. Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
9. Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10. Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11. Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12. Marian Razavian (VU), Knowledge-driven Migration to Services
13. Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14. Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15. Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16. Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17. Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18. Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19. Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20. Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21. Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22. Tom Claassen (RUN), Causal Discovery and Logic
23. Patricio de Alencar Silva (UvT), Value Activity Monitoring
24. Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25. Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26. Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning

27. Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28. Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29. Iwan de Kok (UT), Listening Heads
30. Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31. Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32. Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33. Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34. Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35. Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36. Than Lam Hoang (TUe), Pattern Mining in Data Streams
37. Dirk Börner (OUN), Ambient Learning Displays
38. Eelco den Heijer (VU), Autonomous Evolutionary Art
39. Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40. Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41. Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42. Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43. Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014

1. Nicola Barile (UU), Studies in Learning Monotone Models from Data
2. Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
3. Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
4. Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
5. Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
6. Damian Tamburri (VU), Supporting Networked Software Development
7. Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
8. Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
9. Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10. Ivan Salvador Razo Zapata (VU), Service Value Networks
11. Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12. Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13. Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14. Yangyang Shi (TUD), Language Models With Meta-information
15. Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16. Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17. Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18. Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19. Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20. Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21. Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22. Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23. Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24. Davide Ceolin (VU), Trusting Semi-structured Web Data
25. Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26. Tim Baarslag (TUD), What to Bid and When to Stop
27. Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28. Anna Chmielowiec (VU), Decentralized k-Clique Matching
29. Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30. Peter de Cock (UvT), Anticipating Criminal Behaviour
31. Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32. Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33. Tesfa Tegegne (RUN), Service Discovery in eHealth
34. Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35. Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36. Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37. Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38. Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39. Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40. Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41. Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42. Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43. Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44. Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45. Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46. Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47. Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015

1. Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
2. Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls

3. Twan van Laarhoven (RUN), Machine learning for network data
4. Howard Spoelstra (OUN), Collaborations in Open Learning Environments
5. Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
6. Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
7. Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
8. Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
9. Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10. Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11. Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12. Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13. Giuseppe Procaccianti (VU), Energy-Efficient Software
14. Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15. Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16. Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17. André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18. Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19. Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20. Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21. Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22. Zhemin Zhu (UT), Co-occurrence Rate Networks
23. Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24. Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25. Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26. Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27. Sándor Héman (CWI), Updating compressed colomn stores
28. Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29. Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30. Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31. Yakup Koç (TUD), On the robustness of Power Grids
32. Jerome Gard (UL), Corporate Venture Management in SMEs
33. Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34. Victor de Graaf (UT), Gesocial Recommender Systems
35. Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016

1. Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
2. Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

3. Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
4. Laurens Rietveld (VU), Publishing and Consuming Linked Data
5. Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
6. Michel Wilson (TUD), Robust scheduling in an uncertain environment
7. Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
8. Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
9. Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10. George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11. Anne Schuth (UVA), Search Engines that Learn from Their Users
12. Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13. Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14. Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15. Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16. Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17. Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18. Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19. Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20. Daan Odijk (UVA), Context & Semantics in News & Web Search
21. Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22. Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23. Fei Cai (UVA), Query Auto Completion in Information Retrieval
24. Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25. Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26. Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27. Wen Li (TUD), Understanding Geospatial Information on Social Media
28. Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29. Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30. Ruud Mattheij (UvT), The Eyes Have It
31. Mohammad Khelghati (UT), Deep web content monitoring
32. Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33. Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34. Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35. Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36. Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37. Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38. Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39. Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40. Christian Detweiler (TUD), Accounting for Values in Design
41. Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42. Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43. Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44. Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45. Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46. Jorge Gallego Perez (UT), Robots to Make you Happy
47. Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48. Tanja Buttler (TUD), Collecting Lessons Learned
49. Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50. Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

1. Jan-Jaap Oerlemans (UL), Investigating Cybercrime
2. Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
3. Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
4. Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
5. Mahdieh Shadi (UVA), Collaboration Behavior
6. Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
7. Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
8. Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
9. Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10. Robby van Delden (UT), (Steering) Interactive Play Behavior
11. Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12. Sander Leemans (TUE), Robust Process Mining with Guarantees
13. Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14. Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15. Peter Berck (RUN), Memory-Based Text Correction
16. Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17. Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18. Ridho Reinanda (UVA), Entity Associations for Search
19. Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20. Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21. Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22. Sara Magliacane (VU), Logics for causal inference under uncertainty
23. David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24. Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25. Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26. Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27. Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28. John Klein (VU), Architecture Practices for Complex Contexts
29. Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30. Wilma Latuny (UvT), The Power of Facial Expressions
31. Ben Ruijl (UL), Advances in computational methods for QFT calculations
32. Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33. Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34. Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35. Martine de Vos (VU), Interpreting natural science spreadsheets
36. Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37. Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38. Alex Kayal (TUD), Normative Social Applications
39. Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40. Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41. Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42. Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43. Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44. Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45. Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46. Jan Schneider (OU), Sensor-based Learning Support
47. Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48. Angel Suarez (OU), Collaborative inquiry-based learning

2018

1. Han van der Aa (VUA), Comparing and Aligning Process Representations
2. Felix Mannhardt (TUE), Multi-perspective Process Mining
3. Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling,

Model-Driven Development of Context-Aware Applications, and Behavior Prediction

4. Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
5. Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
6. Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
7. Jieting Luo (UU), A formal account of opportunism in multi-agent systems
8. Rick Smetsers (RUN), Advances in Model Learning for Software Systems
9. Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10. Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11. Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12. Xixi Lu (TUE), Using behavioral context in process mining
13. Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14. Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15. Naser Davarzani (UM), Biomarker discovery in heart failure
16. Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17. Jianpeng Zhang (TUE), On Graph Sample Clustering
18. Henriette Nakad (UL), De Notaris en Private Rechtspraak
19. Minh Duc Pham (VUA), Emergent relational schemas for RDF
20. Manxia Liu (RUN), Time and Bayesian Networks
21. Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22. Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23. Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24. Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25. Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26. Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27. Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28. Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29. Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30. Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

1. Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
2. Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
3. Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data
4. Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
5. Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
6. Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
7. Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

8. Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
9. Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10. Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11. Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12. Jacqueline Heinerman (VU), Better Together
13. Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14. Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15. Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16. Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17. Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18. Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19. Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20. Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21. Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22. Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23. Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24. Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25. Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26. Prince Singh (UT), An Integration Platform for Synchromodal Transport
27. Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28. Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29. Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30. Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31. Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32. Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33. Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34. Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35. Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36. Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37. Jian Fang (TUD), Database Acceleration on FPGAs
38. Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

1. Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
2. Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
3. Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

4. Maarten van Gompel (RUN), Context as Linguistic Bridges
5. Yulong Pei (TUE), On local and global structure mining
6. Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
7. Wim van der Vegt (OUN), Towards a software architecture for reusable game components
8. Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
9. Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10. Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11. Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models
12. Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13. Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14. Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15. Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16. Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17. Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18. Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19. Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20. Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21. Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22. Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23. Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24. Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25. Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26. Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27. Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28. Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29. Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30. Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31. Gongjin Lan (VU), Learning better – From Baby to Better
32. Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33. Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34. Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35. Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021

1. Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
2. Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
3. Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
4. Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
5. Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
6. Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
7. Armel Lefebvre (UU), Research data management for open science
8. Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
9. Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10. Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11. Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12. Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13. Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14. Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15. Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm