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

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.