Bimodal emotion recognition through audio-visual cues

SUMMARY

Emotions are a key component in human-human communications, with a highly complex socio-psychological nature. A great amount of affective information is displayed through facial expressions, gestures, speech, and other means. Recently, Multimodal Emotion Recognition (MER) has gained a notable amount of research interest. It aims to recognize the displayed affective information through techniques and methods from the fields of Affective Computing (AC) and Artificial Intelligence (AI). Furthermore, the recent technological advancements brought interactivity between people and digital devices to a completely different level, making computers and mobile phones an important part of our daily lives. Besides, AI is a rapidly improving field, offering us new mathematical methods for data representations and classification procedures. Therefore, there is an increased interest in the Human-Computer Interaction (HCI) field towards enhancing digital devices with emotion recognition abilities for obtaining a more natural HCI experience. However, HCI still lacks elements of emotional intelligence to enable a more human-centered interaction. Human-centered computation through affective computing can help in recognizing emotions, and generate proper actions to achieve richer and human-like communications in settings like HCI. Automatic systems with affective capabilities can be essential in many applications of affective computing, which range from education, autonomous driving, entertainment to health-care.

Nonetheless, the facts that emotions are multifaceted, socio-psychological, and biological concepts, make automatic emotion recognition a challenging task. This dissertation addresses various problems in multimodal emotion recognition, which is an important task towards achieving Affective Computing (AC) goals. Chapter 1 motivates the research problem and introduces its theoretical foundations. In particular, the research of this dissertation focuses on the two primary forms of emotion expressions and modulations: the face and the voice. In human-human communications, these two modalities are the most expressive and perceived channels. They are widely used in our daily communication for our social interactions. Although information obtained from non-obvious signals of emotion expression (e.g. heart beating, sweating, and respiration) could be informative as well, people rely on the apparent cues in sensing others' emotions. Besides, sensing emotions from auditory and visual channels is not invasive. As a result, this dissertation aims to predict emotions through audio-visual cues, which consequently can lead to enhanced interactions between humans and robots and machines in general. It employs and proposes progressive research towards audio-visual emotion recognition, coming from state-of-the-art techniques in the field of Artificial Intelligence, which are presented in the technical Chapter 2.

Chapter 3 introduces an extensive literature review of Affective Computing (AC) and Multimodal Emotion Recognition (MER). Earlier research in emotion recognition targeted, either individual modalities (such as facial expressions and acoustic-prosodic cues) or global multimodal emotion recognition. On the other hand, the research in this

dissertation focuses on multimodal recognition and exploits temporal interactions between audio-visual channels. It aims to capture modalities' strengths for emotion recognition to utilize their complementary and supplementary information. It adopts recent advances in Affective Computing (AC) such as Deep Neural Networks (DNNs), Deep Metric Learning (DML), end-to-end learning, and the attention mechanism for Audio-Video Emotion Recognition (AVER). Also, over the course of this research, the literature was lacking in-depth analyses regarding automatically extracted, dynamic interactions between audio and video signals in emotionally rich contexts. This dissertation presents studies that investigate the temporal relationships of both modalities and exploits their strength for emotion recognition. Besides, it employs state-of-the-art methods, such as Deep Metric Learning (DML) to perform similarity learning for multimodal emotion recognition.

In this dissertation, four research questions and objectives are introduced to address the joint modeling of audio-visual cues for Multimodal Emotion Recognition (MER). The objective of Chapter 4 is related to data modeling and producing robust multimodal representations for emotion recognition. In two studies, this chapter addresses the first research question: *How to extract and fuse robust features and which is their contribution to automatic emotion recognition?*. The first one deals with emotion recognition in video clips, where audio and visual cues are the primary information for emotion perception. It presents a hierarchical framework for multimodal emotion recognition. The proposed research employs Fisher Vectors (FVs) representations to aggregate frame-level features in a video sample. This encoding is applied on different types of audio and visual features (e.g. Dense Scale-Invariant Feature Transformation (SIFT), geometric, Convolutional Neural Network (CNN), audio), enabling mapping them into a common space, where feature level fusion is performed. It then uses a strategy of employing information gain principles, for selecting the best combination of features to be fused. Finally, a decision-level fusion approach on top of the best features is applied to optimize modalities' weights for each emotional state using a genetic search algorithm. The experimental results show that the two fusion schemes on the employed modalities and their features improve the accuracy of emotion prediction compared to unimodal emotion recognition. The second part of the chapter studies the correlation between students' self-reported affective states, according to the Theory of Flow (ToF), and their interactions with learning materials. This study designs a framework to track contextual information and interaction features during learning activities. It utilizes a standard tracking tool, the xAPI framework, for learning analytics. The conducted evaluations highlighted the potential usage of interaction parameters with learning materials as a useful channel for measuring affective states.

Chapter 5 focuses on the objective of efficient fusion for audio-visual representations. It addresses the second research question: *what is the impact of multimodal learning on emotion recognition?*. It introduces a modality-specific Multimodal Emotion Recognition Metric Learning (MERML). This method is applied to improve the latent-space of audio-visual data representations. It successfully exploits the complementary information of audio and video modalities for emotion recognition. As a result of this approach, audio-visual representations are well structured in the newly learned subspace, and their capacity for optimized emotion recognition is maximized. The conducted

quantitative and qualitative evaluations of the method demonstrated the contribution of the method to increased classification accuracy. Chapter 6, benefits from the findings of the MERML framework, and builds an end-to-end Deep Metric Learning (DML) with triplet loss for audio-visual temporal emotion recognition. In addition, it aims at the objective of exploiting the temporal dynamics of emotion display and perception, and also at answering the third research question: *What is the role of temporal dynamics in audio-visual cues, in automated emotion recognition?*. In the study of Chapter 6, inspired by the gating paradigm, we investigate how introducing multimodal cues with increasing durations impacts the recognition rates of positive and negative emotions. The procedure employs Long-Short Term Memory (LSTM)s between time windows for incremental perception to mimic the gating paradigm. The proposed framework embeds audio-visual cues overtime, taking advantage of the temporal display of emotions. It also checks the contribution of audio, visual, and audio-visual fusion in emotion recognition. The framework showed efficiency in modeling the temporal context of multimodal emotion recognition. Besides, within the introduced framework, algorithms to tackle the challenges of triplet sets' mining and the convergence of Deep Metric Learning (DML) are proposed. The developed approach and the associated techniques, such as Multi Window Triplet Sets Mining (MWTSM), contributed significantly to the stability and the performance of the framework for Multimodal Emotion Recognition (MER). In addition, the evaluations proved the benefits of the incremental perception of both audio and visual cues in the recognition rates overtime. Additionally, the temporal differences of the recognition speed for positive and negative emotions differ, where positive emotions are recognized faster than the negative ones.

Chapter 7 targets the research objective of attending to informative time segments in temporal audio and video signals. It addresses the fourth research question: *How can we capture the contributions of the temporal dynamics of affect display using attention-mechanisms?*. The study of this chapter employs attention mechanisms on audio-visual embeddings over time windows to capture their temporal properties for emotion recognition. The evaluation of the proposed method, namely Multimodal Attention mechanism for Temporal Emotion Recognition (MATER), highlights the importance of weighing the time windows in audio-visual cues. The presented method offers interpretability and explainability of the attention mechanisms for temporal and multimodal fusion. Furthermore, MATER presents extensive studies and meta-analysis findings, linking the outputs of our proposition to research from psychology. For example, it gives more insights with regards to the multimodal interaction and presentation of audio-visual cues. It examines how the attention mechanisms helps in joint modeling of multimodal cues and subsequently enhances their performance. Moreover, this study shows that the contribution of the video modality in multimodal fusion is greater than the one of the audio modality. Finally, it applies noise injection into the time windows embeddings, during the evaluation or/and the training phases, to demonstrate the robustness of the method and its ability to adapt to challenging conditions.

Finally, Chapter 8 concludes the conducted research, highlights its findings, and points out some directions for future work in affective computing and multimodal emotion recognition.