Chapter 57
Depth Maps and Deep Learning for Facial Analysis

Paulo C. Brito
Universidade Aberta, Portugal

Elizabeth S. Carvalho
Universidade Aberta, Portugal

ABSTRACT

Gathering and examining progressively multi-modular sensor information of human faces is a critical issue in PC vision, with applications in examinations, entertainment, and security. However, due to the exigent nature of the problem, there is a lack of affordable and easy-to-use systems, with real-time, annotations capability, 3D analysis, replay capability and with a frame speed capable of detecting facial patterns in working behavior environments. In the context of an ongoing effort to develop tools to support the monitoring and evaluation of the human affective state in working environments, the authors investigate the applicability of a facial analysis approach to map and evaluate human facial patterns. The challenge is to interpret this multi-modal sensor data to classify it with deep learning algorithms and fulfill the following requirements: annotations capability, 3D analysis, and replay capability. In addition, the authors want to be able to continuously enhance the output result of the system with a training process in order to improve and evaluate different patterns of the human face.

INTRODUCTION

Facial expressions are nowadays used for determining emotions, yet they additionally can be utilized to demonstrate mental states. (Craig, DMello, Witherspoon, & Graesser, 2008) looked at facial expressions while students worked with an online tutoring system and identified that frustration was associated with activity in the inner and outer brow raiser and dimple; confusion was associated with brow lowered, lip tightened and lip corner puller. Moreover, preliminary results by (Dinges et al., 2005) suggested that high and low stressing situations could be discriminated based on the facial activity in the mouth and eyebrow regions. In addition, stress can be inferred using multimodal sensor data as stated in a survey (Sharma & Gedeon, 2012).

DOI: 10.4018/978-1-7998-0414-7.ch057

Vast amounts of data from different sensors can be used to try to deduce human facial affective states. Interpreting this data in a meaningful way, is challenging because of the volume of data to analyze and classify and the method to fuse the data from different types.

In studies conducted by Fan, Peng, and Zhong, (2010), Aissaoui, Martinet, and Djeraba, (2014) and (Tjahyaningtijas, Puspitasari, Yamasari, Anifah, & Buditjahyanto, 2018) several techniques that can produce 3D facial data from a range of sensors allowing the 3D reconstruction of faces in real time. This data can be used later on to evaluate patterns in a facial dataset.

In this study, the authors apply a few software engineering procedures, for example, visual investigation, depth learning and facial following, to a facial dataset with a specific aim of deciphering multimodal sensor information. The authors exchange between information examination, to find out structures in the information, and representations, to pick up bits of knowledge. This exploratory investigation was done to find out patterns between facial expressions and emotional states that can be estimated with sensors.

This work has as main challenge the interpretation of multi-modal sensor data to classify it with deep learning algorithms and fulfill the follow requirements: annotations capability, 3D analysis and replay capability. This is still a field that has to be much more explored. The main advantage that it can introduce is the capability of detecting automatically people's facial expression totally based on their micro-expressions. This can help to support a better evaluation of human behavior or mind state in areas such as forensic or psychology, to name a few for instance.

BACKGROUND

Affective computing (Lin, Pan, Wang, Lv & Sun, 2010) is the advancement and investigation of frameworks and gadgets that can distinguish, comprehend, process, and recreate human emotions. The inspiration for the study and research of this area is the ability to simulate empathy. The system should infer the emotional state of humans and adapt its actions to them.

The way people participate in an activity has been studied from several perspectives in HCI and psychology. The term "engagement" involves attentional and emotional involvement with a task. Engagement is also not stable, but fluctuates throughout an interaction experience.

Identifying emotional data starts with uninvolved sensors, which catch information about the person's physical state or behavior, without translating the input. The information accumulated is practically equivalent to the signs that people use to see feelings in others. For instance, a camcorder may catch outward appearances, body stance, and signals, while a receiver may catch discourse. Other sensors detect emotional cues by directly measuring physiological data, such as skin temperature and galvanic resistance (Garay-Vitoria, Cearreta, López, & Fajardo, 2006).

Recognizing emotional information requires the extraction of substantial patterns from the gathered data. This may be obtained using machine learning techniques, (Deshmukh, Paygude, & Jagtap, 2017) that process different modalities, such as facial expression detection, natural language processing, or speech recognition, and produce labels (i.e. 'awe') to identify what emotion is being expressed.

The vast majority of present systems are data-dependent. This creates one of the biggest challenges in detecting emotions based on speech, as it implicates choosing an appropriate database to train the classifier. Most of the currently possessed data was obtained from actors and is thus a representation of archetypal emotions. Those so-called acted databases are usually based on the Basic Emotions theory (by Paul Ekman), which assumes the existence of six basic emotions (anger, fear, disgust, surprise, joy,

Related Content