Image and Vision Computing 31 (2013) 137–152

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information ☆

Angeliki Metallinou *, Athanasios Katsamanis, Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA, United States

ARTICLE INFO

Article history:
Received 29 October 2011
Received in revised form 7 July 2012
Accepted 22 August 2012

Keywords:
Continuous emotion tracking
Dimensional emotional descriptions
Gaussian Mixture Model mapping
Body language
Improvised dyadic interactions

ABSTRACT

We address the problem of tracking continuous levels of a participant's activation, valence and dominance during the course of affective dyadic interactions, where participants may be speaking, listening or doing neither. To this end, we extract detailed and intuitive descriptions of each participant's body movements, posture and behavior towards his interlocutor, and speech information. We apply a Gaussian Mixture Model-based approach which computes a mapping from a set of observed audio–visual cues to an underlying emotional state. We obtain promising results for tracking trends of participants' activation and dominance values, which outperform other regression-based approaches used in the literature. Additionally, we shed light into the way expressive body language is modulated by underlying emotional states in the context of dyadic interactions.

1. Introduction

Human expressive communication is characterized by the continuous interplay of multimodal information, such as facial, vocal and bodily gestures, which may convey the participant's affect. The affective state of each participant can be seen as a continuous variable that evolves with variable intensity and clarity over the course of an interaction. It can be described by certain continuous attributes (dimensions): activation, valence and dominance. Activation describes how intense is the emotional experience, valence describes the level of pleasure related to an emotion, and takes positive and negative values for pleasant and unpleasant emotions respectively, while dominance describes the level of control of a person during an emotional experience. This approach was introduced in psychology research based on evidence that humans may perceptually use such a representation to evaluate emotional situations [1–3]. It may also be a more generic way to classify emotions, especially for emotional manifestations that may not have a clear categorical description.

This work addresses the problem of continuous tracking of activation, valence and dominance, when they are considered to be continuously valued. Our goal is to obtain a continuous description of each participant's underlying emotional state through the course of an improvised dyadic interaction. Our experimental setup is generic; participants express a wide variety of emotions that are not pre-defined but are elicited through their interaction, and have varying roles throughout the performance (speaker, listener, neither). This approach has the potential to shed light into the temporal dynamics of emotions through an interaction and highlight regions where abrupt emotional change happens. These could be viewed as regions of emotional saliency.

Our contributions could be summarized as follows:

1. We present a statistical framework to dynamically track the emotional content that is displayed over time by participants of an interaction, using bodily and vocal information.
2. We systematically examine how body language behavior is modulated by underlying emotional states in dyadic interactions.
3. We discuss the data annotation design for continuous ratings, which is a challenging problem in itself.

We apply a Gaussian Mixture Model (GMM) based methodology, originally introduced in [4], to compute an optimal statistical mapping between an underlying emotional state and an observed set of audio–visual features, both evolving through time. Extending our previous work [5], we formulate the emotion tracking problem at various time resolutions, to investigate the effect of the tracking detail on the final performance. For our experiments, we use the USC Creative IT database which contains detailed full body Motion Capture (MoCap) information in the context of expressive theatrical improvisations [6]. We extract a variety of psychology-inspired body language features describing each participant's body language and relative interaction behaviors with respect to their interlocutor. We systematically examine the relevant emotional content of each feature to select body language feature sets tailored to each emotional attribute. In addition to emotion tracking, our goal is to examine the way expressive body language is modulated in order to reflect different emotional states. This allows us to revisit qualitative psychological observations from a quantitative perspective.

☆ This paper has been recommended for acceptance by Hatice Gunes and Bjoern Schuller.

* Corresponding author. Tel.: +1 2139052659.
E-mail addresses: metallin@usc.edu (A. Metallinou), nkatsam@sipi.usc.edu (A. Katsamanis), shri@sipi.usc.edu (S. Narayanan).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2012.08.018

Finally, the data annotation design is an important part of the data preparation, since continuous tagging is a challenging task and often results in low inter-evaluator agreement. Our annotation results show that people tend to agree more on the trends rather than the absolute values of emotional attributes. This suggests that humans find it more straightforward to define emotions in relative (e.g., more activated, more dominant), rather than absolute terms (similar observations are described in [7]).

Our experimental results indicate that we are better at tracking changes in emotional attributes rather than the absolute values themselves, following a similar trend as the human annotations. Furthermore, the proposed GMM based tracking method outperforms other examined methods, in terms of correlation-based performance metrics (estimating trends of attributes). For activation trends, the tracking performance is close to human agreement, while for dominance we achieve encouraging results. Body language seems to carry rich activation and dominance related information, reflected in features such as body and hand movements, orientation and approach–avoidance behaviors.

2. Related work

The use of dimensional representations of emotions has been adopted by many researchers but typically the dimensional values are quantized into discrete levels. However, a continuous representation may allow a more generic and flexible treatment of emotions. Examples of work that avoid discretizing the emotional dimensions include [8,9] where regression approaches, such as Support Vector Regression (SVR), were used to estimate continuous dimensional attributes from speech cues of presegmented utterances.

Most of the existing literature, including works that focus on recognition of emotions as part of an emotion sequence [10,11], presegment the time dimension into units for recognition, e.g., consecutive words or utterances. Few works have avoided segmenting the temporal dimension and have addressed the problem of continuously tracking emotions across time. For example, in [12] the authors present continuous recognition of the emotional content of movies using a Hidden Markov Model (HMM) which classifies dimensional attributes into discrete levels.

A relatively small amount of literature treats both time and emotion variables as continuous. In [13] the authors describe a multimodal system to continuously track valence and activation of a speaker, using SVR and Long-Short Term memory (LSTM) regression, with LSTM being the best performing approach. Similarly, single-modality systems were proposed in [14,15] using SVR and LSTM neural networks for regression to continuously estimate valence and activation values from emotional speech. An unsupervised method for mapping the emotional content of movies in the valence–activation space was proposed in [16,17] using low-level audio and video cues. In our work, we propose a supervised, GMM-based methodology to continuously track an underlying emotional state using body language and speech information.

The use of multimodal information allows for a more complete description of the expressed emotion, therefore many works utilize both facial expressions and vocal cues [18,19], while an increasing amount of recent literature investigates body language. In [20,21] the authors use upper body language information along with facial expressions to recognize emotions, while in [13] shoulder movement cues were used along with facial and vocal cues for continuous emotion tracking. In [22] authors investigate a variety of upper body descriptions of movement and symmetry in order to extract a minimal representation of affective gestures. Works that examine affective full body language include [23] where authors advantageously use full body motion cues, alongside facial and vocal information, and [24] where authors use features describing movement quality to classify basic emotional states. In [25], authors use the setup of a body-movement-based videogame and recognize emotions such as defeat, triumph etc., using MoCap derived features. Few works have addressed body language behavior in the context of social interaction, for example the work in [26], that examines dominance and synchronization phenomena during collaborative social tasks, and [27] where measures of posture are used to examine approach–avoidance behaviors during the interaction of two seated participants.

Various body language feature sets have been proposed in the literature, ranging from lower-level features such as joint angles [25,28], to more interpretable features such as distances and angles between body parts [29,30] and this work, to higher-level posture and movement properties (contraction index, smoothness/fluidity of motion) [22,24]. An overview of various body language features in the literature can be found in [31]. In this work we extract a large set of interpretable body language features, which measure properties of a person's posture, motion, and body behavior with respect to the interlocutor. Although there seems to be no standard feature set for body language, several body language features in the literature measure similar qualities. For example, in [29] authors measure horizontal and vertical distances between a subject's hands and shoulder, while here we compute the relative positions of a person's hands with respect to his torso.

Our work lies in the intersection of many of the above areas; we address the issue of emotion tracking when both the emotion and time dimensions are continuous, using full body language features and speech information. Body language is examined in the context of affective dyadic interactions. Additionally, our setup is generic; the examined subjects are not restricted to produce specific emotions or body gestures. On the contrary, through their improvisation a wide variety of emotional states, body language gestures and interaction dynamics are elicited in a naturalistic manner.

3. Framework overview

3.1. Overview

Fig. 1 presents a summary of our work. As illustrated in the left of Fig. 1, our study relies on video, audio and MoCap data collected from two actors engaged in emotional dyadic improvisations. The center part of Fig. 1 describes the data processing, specifically the extraction of detailed body language and speech information from both participants, as well as the data annotation. Data annotation was performed by multiple human evaluators who were asked to continuously rate the perceived valence, activation and dominance levels of each participant during each interaction. The result is multiple emotional curves which are averaged to provide the ground truth for further experiments. After these steps, we have available for each participant various body language features x_{body} extracted throughout the interaction, speech features x_{speech} extracted from regions where that person is speaking, and the corresponding emotional curves y. The joint distribution $P(x,y)$ is modeled using a Gaussian Mixture Model (GMM), where x can be a visual or audiovisual feature vector and y is one of the three emotional attributes. The conditional distribution $P(y|x)$ is also a GMM. The GMM-based tracking approach consists of computing a mapping from the observed features to the underlying emotional curve by maximizing the conditional probability of the emotion given the features, e.g., $\hat{y} = argmax P(y|x)$. In the right part of Fig. 1 we present an example of the resulting emotional curve estimate.

3.2. Framework for continuous tracking of emotional states and emotional changes

Let $\mathbf{x}_t$ denote the vector of body language and speech observations at time t of an interaction recording and y_t be the underlying emotional attribute, namely activation, valence or dominance. One way to predict y_t given $\mathbf{x}_t$ would be by maximizing the corresponding conditional probability:

$$\hat{y}_t = arg\max_{y_t} P\left(y_t \middle| \mathbf{x}_t, \lambda^{(y,\mathbf{x})}\right) \quad (1)$$