TRACKING CHANGES IN CONTINUOUS EMOTION STATES USING BODY LANGUAGE AND PROSODIC CUES

Angeliki Metallinou[1]*, Athanassios Katsamanis*[1]*, Yun Wang*[2] *and Shrikanth Narayanan*[1]

[1]University of Southern California, Los Angeles, CA, [2] Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

Human expressive interactions are characterized by an ongoing unfolding of verbal and nonverbal cues. Such cues convey the interlocutor's emotional state which is continuous and of variable intensity and clarity over time. In this paper, we examine the emotional content of body language cues describing a participant's posture, relative position and approach/withdraw behaviors during improvised affective interactions, and show that they reflect changes in the participant's activation and dominance levels. Furthermore, we describe a framework for tracking changes in emotional states during an interaction using a statistical mapping between the observed audiovisual cues and the underlying user state. Our approach shows promising results for tracking changes in activation and dominance.

Index Terms— emotion recognition, emotion tracking, body language, gaussian mixture models, features to emotion mapping

1. INTRODUCTION

Human expressive communication is characterized by a continuous flow of interacting multimodal cues, including body gestures and speech. Understanding the role of body language during emotional expression might be of use both for emotion recognition applications and for the design of emotional virtual agents. Furthermore, when designing emotion classification systems it is important to consider the continuous nature of audiovisual cues and of the expressed emotions. Along these lines, the focus of this paper is two-fold; first to analyze the emotional content of body language gestures and second to use this rich information flow to continuously track the changes in emotion states.

Body language cues such as body posture, orientation, proxemics and approach/avoidance behaviors have been long studied in behavioral reseach and have been shown to convey emotional information [1]. However, few engineering works focus on the analysis of body gestures and their application for emotion recognition, [2], perhaps due to the lack of appropriate behavioral databases. Here, we use a large multimodal and multispeaker database of dyadic improvised interactions, containing detailed body language information that allows us to analyze the emotional content of various body gestures [3].

Furthermore, despite the variety of intensity and clarity of affective displays during an interaction, little work has been done in tracking continuous emotional expressions through the course of time. In [4] continuous emotions are recognized from presegmented utterances using speech cues, while in [5] a neural network architecture is used to continuously recognize continuous emotions of a speaker, using speech and linguistic information. In this work, we compute a statistical mapping between the underlying continuous emotional states and the observed behavioral features [6]. This enables us to develop a framework for continuous tracking of the unfolding emotional changes using multimodal cues extracted from long unsegmented recordings, where the participant may be speaking, listening or engaged in neither.

Our analysis suggests that body language features that describe relative positions, orientation and approach/avoidance behaviors of the participants towards their interlocutor in an interaction, reflect their level of activation and dominance but are less informative about their valence (positive vs negative). Prosodic features are also highly informative of speaker activation. Our emotion tracking experiments are moderately successful in recognizing changes (increase and decrease) in the participants' activation and dominance through the course of their interaction. These results can be seen as a first step towards dynamically tracking emotional changes and detecting emotionally salient regions in complex affective interactions.

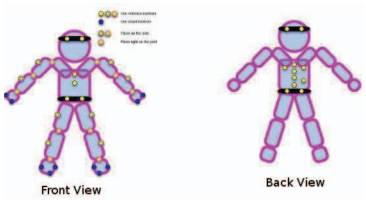
2. DATABASE AND ANNOTATION PROCESS

In this work we use the USC CreativeIT database, which is a novel, multimodal and multidisciplinary database[3]. The CreativeIT database consists of theatrical improvisations of pairs of actors. Actors make use of expressive body language in order to produce a great variety of affective displays and intentions. The goals of this database include studying affective communication and interaction of actors and humans. The actors wear Motion Capture (MoCap) markers on their body and close talking microphones and they are recorded by MoCap and high definition cameras. In this work, we use the detailed MoCap information and the speech cues.

Our data annotation process was designed to preserve the continuous flow of body language and dialog during the improvisations. Therefore, instead of segmenting the recordings into sentences or arbitrary chunks we annotated the whole recording using continuous emotional attributes; valence (positive vs negative), activation (excited vs calm) and dominance (dominant vs submissive). We use the Feeltrace software [7], which enables a human annotator to continuously indicate the attribute value, ranging from -1 to 1, while watching the video of a recording. Each actor in each recording is annotated by 3 or 4 different annotators. Currently, three sessions of the database are fully annotated, which correspond to 20 audiovisual recordings of pairs of actors, from 7 different actors (2 male, 5 female) in total. Recordings range from 2 to 8 minutes. Here we track the emotional state of each actor and therefore our working dataset contains 40 instances.

Defining inter-evaluator agreement for the continuous annotations is less straighforward than for discrete labels. In Fig. 2(a) (section 4.2) we present a segment of the activation annotations of an actor provided by 4 annotators (normalized to unit standard deviation). Although the annotations differ, they are positively correlated.

This work was supported in part by funds from NSF



(a) Marker Positions

feat1	distance of A to B	feat11	relative velocity of A's right hand towards A
feat2	angle of A's face towards B	feat12	relative velocity of A's right hand towards B
feat3	angle of A's body towards B	feat13	relative velocity of A's left hand towards A
feat4	absolute velocity of A	feat14	relative velocity of A's left hand towards B
feat5	relative velocity of A to B	feat15	min distance of A's right hand to B's hand
feat6	absolute velocity of A's right arm	feat16	min distance of A's left hand to B's hand
feat7	absolute velocity of A's left arm	feat17	min distance of A's right hand to B's head
feat8	angle of A's body leaning towards B	feat18	min distance of A's left hand to B's head
feat9	distance of A's hands	feat19	min distance of A's right hand to B's torso
feat10	angle of A's head (looking up, down etc)	feat20	min distance of A's left hand to B's torso

(b) Extracted features

Fig. 1. The positions of the Motion Capture markers and a description of the extracted body language features

We define agreement when the linear correlation between two annotations is greater than a threshold. For each actor annotation, we take the union of all annotator pairs with linear correlations greater than 0.4; this annotation subset is used to compute the ground-truth for the corresponding annotation. If no annotators are selected then we exclude the instance from our analysis. This results in selecting 35, 33 and 30 instances for activation, valence and dominance respectively and the median of the evaluator correlations is 0.55, 0.48, 0.47 respectively (the correlations for activation and dominance are presented in Figs. 2(f) and (i) in section 4.2).

3. FEATURE EXTRACTION AND ANALYSIS

3.1. Body Language and Prosodic Feature Extraction

The CreativeIT database contains detailed body marker information, obtained from MoCap (Fig. 1(a)). This provides knowledge of the body positions of both speakers during the improvisation and enables us to extract detailed descriptions of each actor's body language. Our features are motivated from the psychology literature which indicates that body language behaviors, such as looking at the other or turning away, approaching, touching etc, carry emotional information [1]. The extracted body language features are summarized in Fig. 1(b) and roughly describe the posture of the actor as well as his position and approach behaviors with respect to the other actor. A key point is that when extracting features from an actor we use information from both actors; therefore each actor's features convey information about the dyadic interaction flow. The features are geometrical and are computed in a straightforward manner by defining local coordinate systems for each actor and by computing euclidean distances and relative positions and velocities.

We also extract standard spectral envelope and prosodic features; 18 Mel Filterbank features, pitch and energy, from the microphone signal of each actor. Speech features are extracted only from regions where the actor is speaking, using available segmentations indicating the start and end of each actor's speech.

3.2. Statistical Analysis of Body Language Features

It is suggested from behavioral research that human undestanding of emotions is relative rather than absolute and depends on the emotional context. This is supported by our annotation data where annotators are more likely to agree in the increase, decrease or stability of emotional attributes, rather than in the absolute values of those attributes. Here we analyze the body language behaviors that are associated with an emotional change of an actor, as perceived by annotators.

For each of our data instances we average all the annotations of an attribute and we approximate the mean annotation with a piecewise linear curve (with linear regions of 2.5sec). Each region is denoted as increase, decrease or stability of the corresponding emotional attribute, according to the slope of the fitted line. We collect these regions from our data and for each region we extract body language features and compute higher level descriptors of body behavior. For example, from the absolute velocity of an actor we compute the percentage of time that the actor is walking versus standing. For each region of increase, decrease or stability we compute the percentage of time where each of the body language behaviors of Table 1 occurs. Statistically significant differences between the mean value of the percentages in different regions indicate that actors resort to certain body gestures more or less often according to whether they display decrease, increase or stability of their emotional attributes (two sided t-test, p=0.05).

Table 1. *Statistical significance tests on body language features for activation (p=0.05). Features are extracted on regions where activation either increases (inc), decreases (dec) or stays the same (st).*

	dec vs. st	inc vs. st	inc vs. dec
Face Position towards other (feat 2)			
towards	-	-	-
sideways	-	-	-
opposite	-	-	-
Walking vs Standing (feat 4)			
walking	-	$m^A_{inc} > m^A_{st}$	$m^A_{inc} > m^A_{dec}$
Body Leaning Towards or Away from Other (feat 8)			
leaning towards	-	$m^A_{inc} > m^A_{st}$	-
leaning away	-	-	-
Approaching or Avoiding Other (feats 5,11,12,13,14)			
body moving towards	-	$m^A_{inc} > m^A_{st}$	$m^A_{inc} > m^A_{dec}$
body moving away	$m^A_{dec} > m^A_{st}$	-	$m^A_{dec} > m^A_{inc}$
right hand moving towards	-	$m^A_{inc} > m^A_{st}$	$m^A_{inc} > m^A_{dec}$
right hand moving away	-	-	-
left hand moving towards	-	$m^A_{inc} > m^A_{st}$	$m^A_{inc} > m^A_{dec}$
left hand moving away	-	-	-
Hands touching each other (feat 9)			
hands touching	$m^A_{dec} > m^A_{st}$	-	$m^A_{dec} > m^A_{inc}$
Hands touching Other (feats 15,16,17,18,19,20)			
right hand touching Other	$m^A_{dec} > m^A_{st}$	$m^A_{inc} > m^A_{st}$	
left hand touching Other	-	-	-

Table 1 shows the results of our statistical analysis for the activation attribute. The inequalities of the mean proportions for each attribute are presented only when the means are significantly different at the 0.05 level. For example, looking at "walking" from Table 1, we conclude that actors tend to walk more when increasing their activation, as compared to decreasing or keeping it stable. In general, for an actor that becomes increasingly active, our analysis suggests that he/she is walking more, and is approaching the other actor more ($m^A_{inc} > m^A_{dec}, m^A_{st}$ for "walking", "body moving towards" and "hands moving towards"). On the contrary, in the

case of decreasing activation, the actor tends to withdraw more and hold hands together ($m^A_{dec} > m^A_{st}, m^A_{inc}$ for "body moving away" and "hands touching"). The results for valence and dominance are not presented here for lack of space. Our observations are that, while differences in dominance seem to be meaningfully reflected in approach-withdrawal behavioral cues, face position, and walking, valence changes do not seem to be well captured by our features.

4. TRACKING CHANGES IN EMOTIONS

4.1. Joint Audiovisual-Emotional Gaussian Mixture Model

The continuous nature of our data motivates the use of a method that enables continuous tracking of changes in emotional states. For this purpose, we apply a Gaussian Mixture Model (GMM)-based mapping, which was originally introduced for the problem of speech to articulatory movement inversion [6], i.e., tracking continuous hidden vocal tract properties given continuous speech observations. Here the objective is formulated as finding an optimal statistical mapping in the Maximum Likelihood sense (MLE) between the observed body language and speech features and the hidden emotional attributes (activation, valence and dominance).

Let's denote $\mathbf{x_t}$ as the hidden attributes and $\mathbf{y_t}$ as the observed features at time t. We train a GMM to model the joint probability distribution of $\mathbf{x_t}$ and $\mathbf{y_t}$, i.e., $P(\mathbf{x_t}, \mathbf{y_t}|\lambda^{(\mathbf{x},\mathbf{y})}) = \sum_{m=1}^{M} a_m N(\mathbf{x_t}, \mathbf{y_t}; \boldsymbol{\mu_m}^{(x,y)}, \Sigma_m^{(x,y)})$ where $a_m, \boldsymbol{\mu_m}^{(x,y)}$ and $\Sigma_m^{(x,y)}$ are the corresponding component weights, means and covariance matrices. The conditional distribution of the emotion x_t given the observation y_t is also represented as a GMM, i.e.,

$$P(\mathbf{x_t}|\mathbf{y_t}, \lambda^{(\mathbf{x},\mathbf{y})}) = \sum_{m=1}^{M} P(m|\mathbf{y_t}, \lambda^{(\mathbf{x},\mathbf{y})}) P(\mathbf{x_t}|\mathbf{y_t}, m, \lambda^{(\mathbf{x},\mathbf{y})}). \quad (1)$$

where $P(\mathbf{x_t}|\mathbf{y_t}, m, \lambda^{(\mathbf{x},\mathbf{y})})$ is the m-th component distribution and $P(m|\mathbf{y_t}, \lambda^{(\mathbf{x},\mathbf{y})}$ is the corresponding so-called occupancy probability. For each test recording, we can then estimate the hidden emotional state given the observed features by maximizing the likelihood of this conditional model:

$$\hat{\mathbf{x}}_\mathbf{t} = arg \max_{\mathbf{x_t}} P(\mathbf{x_t}|\mathbf{y_t}, \lambda^{(\mathbf{x},\mathbf{y})}). \quad (2)$$

To also incorporate dynamic information, we augment the vectors $\mathbf{x_t}$, $\mathbf{y_t}$ with corresponding derivative estimates [6]. By using information from neighboring frames, this approach basically allows us to exploit the continuous nature of our observations and hidden attributes and can provide us with relatively smooth emotional state trajectory estimates.

4.2. Experimental Setup and Results

For our experiments, we randomly split our dataset into 6 disjoint sets of similar size. The mapping of the recordings belonging in a set is computed from a GMM trained on the recordings of the remaining 5 sets. The underlying emotional state x is assumed to be the mean of the annotator curves. All the features y and the emotional states x of an actor's recording are normalized to have zero mean and unit standard deviation. The implementation of this method when using only body language features as our observed features y is as follows. We split our recordings in segments of 5sec length (300 MoCap frames) and overlap 2.5sec and we use these segments to train a joint visual-emotional GMM. We refer to the MoCap based features as visual, basically because these features can be visually perceived. During the testing stage we apply the MLE-based mapping in each 5sec segment to compute the underlying emotional state. The overlapping output curves belonging to a single recording are lowpassed and merged into a single curve, using the overlapp-add method for neighboring overlapping segments. This results in smooth curves that approximate the underlying emotional state.

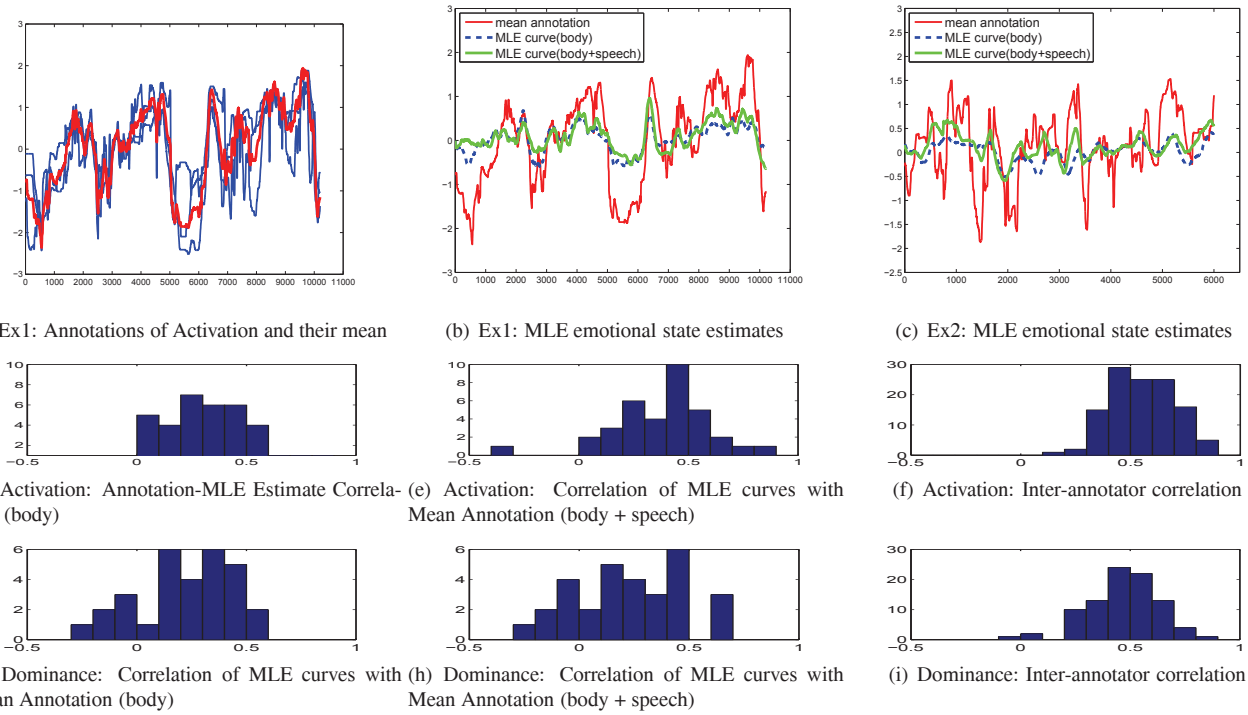
In the multimodal case, where we incorporate prosodic features we need to take into account that for large portions of the recording the actor may not speak. Therefore there are no speech features in these portions, while body language features are extracted throughout the recording. We use the train segments containing both body language and speech to train an audio-visual-emotional GMM. During the testing stage, if our current segment contains speech we use the audio-visual-emotional GMM for the MLE-mapping else we use the visual-emotional GMM. We then augment the features **y** and the emotional states **x** with their first and second derivatives and we train audio and audio-visual GMMs with full covariance matrices. All the presented curves are produced by the MLE based estimation described in Sec. 4.1.

In Fig. 2(a) we present an example of 4 annotations of activation in a recording (blue curves) along with their mean activation (red), which is the ground truth. In Fig. 2(b) we present the MLE curve computed from body language features (blue) and from body language and speech (green) for the example of Fig. 2(a). The (linear) spearman correlation of the mean annotation with the MLE curve is 0.80 when visual cues are used and increases to 0.81 when audiovisual cues are used. The MLE curves generally seem to follow the trends of the mean annotation, although they may be far from the respective amplitude values. A second example is shown in Fig. 2(c) where the mean annotation of activation of an actor recording (red) is presented along with the MLE-mapping curves computed from visual cues (blue) and audiovisual cues (green). Here, incorporating speech information significantly improves the correlation between the mean and the MLE mapping from 0.40 to 0.54.

For each actor's recording, we compute the spearman correlation coefficient between the mean annotation and the MLE curve. These correlations measure how well the computed curves capture the changes of the underlying emotional attributes. We also compute the correlations of the multiple annotators for each actor recording and present their histogram as a performance upper bound. In Figs. 2(d),(e) we present the correlation histograms for activation when visual (median 0.31) and audiovisual (median 0.42) cues are used respectively. In Fig.2(f) we show the inter-evaluator correlations for activation (median 0.55). The MLE curves are positively correlated with the underlying emotional state, and for audiovisual cues, the correlations on average increase significantly and are comparable with the inter-evaluator correlations. In that sense, we could argue that we are able, to a large extent, to detect activation changes. Similar histograms are presented for dominance in Figs. 2(g),(h) and (i) (respective median values are 0.26, 0.23 and 0.47). Here, the correlation coefficients are lower and incorporating speech information does not improve our tracking performance. However, as indicated by the large number of positive correlations in the histograms in Figs. 2(g) and (h), we can detect some of the changes in dominance. For valence, the MLE mapping curves fail to track the changes and the respective median correlations are close to zero.

5. DISCUSSION

A key point is the observability of the underlying emotional states through our extracted features, which directly influences our track-



(a) Ex1: Annotations of Activation and their mean

(b) Ex1: MLE emotional state estimates

(c) Ex2: MLE emotional state estimates

(d) Activation: Annotation-MLE Estimate Correlation (body)

(e) Activation: Correlation of MLE curves with Mean Annotation (body + speech)

(f) Activation: Inter-annotator correlation

(g) Dominance: Correlation of MLE curves with Mean Annotation (body)

(h) Dominance: Correlation of MLE curves with Mean Annotation (body + speech)

(i) Dominance: Inter-annotator correlation

Fig. 2. Tracking Continuous Activation and Dominance: MLE Curves and Correlation histograms. Results for two examples (Ex1, Ex2) are shown in the first row. For the correlation estimates in the other two rows, spearman correlation has been used.

ing performance. The results of Sec. 4.2 suggest that the extracted body language features reflect activation and dominance changes but fail to capture valence. Introduction of prosodic cues is very beneficial for activation, a finding that agrees with existing emotion recognition literature, e.g., [5], but does not increase dominance and valence performance. Further, the MLE curves are much better at tracking changes in emotional states rather than the exact amplitute values of the underlying emotional ground truth. This could be attributed to our extracted features (they could be more informative of relative rather than absolute change) and to the general difficulty of quantifying emotional states in absolute terms.

Furthermore, the statistical analysis of section 3.2 sheds light on the body gestures that have emotional connotations. For example, increase in activation is often displayed by more walking and more approach behaviors towards the interlocutor, compared to stability or decrease. The poor performance of our body language features for detecting valence is also reflected in the statistical analysis. Examining the emotional content of individual body-language gestures gives us important intuitive guidelines and, along with the tracking curves, could help us identify emotionally salient regions of an interaction. Understanding the perceptual effect of combinations of features is the next step, since such audiovisual feature combinations seem to constitute primarily an emotional experience. Finally, understanding the link between improvised acting and natural emotional expression could help us extrapolate our analysis to natural human behavior.

6. CONCLUSIONS AND FUTURE WORK

In this work, we present an analysis of the emotional content of body language cues, describing posture, relative orientation and approach/avoidance behaviors of an interlocutor in improvised affective interactions. We also describe a framework for continuously monitoring changes in the underlying affective state of the interlocutor through the course of an interaction, when speaking, listening or engaged in neither of the two actions. Our immediate work includes completing the annotation of the remaining sessions of the CreativeIT database and increasing the number of annotators per recording, which would provide us with more data and more robust ground truth. Future goals include extracting body features tailored to each emotional attribute and working towards automatic detection of emotionally salient parts of an interaction.

7. REFERENCES

[1] J.A. Harrigan, R. Rosenthal, and K.R. Scherer, *The new handbook of Methods in Nonverbal Behavior Research*, Oxford Univ. Press, 2005.

[2] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," *Affect and Emotion in Human-Computer Interaction*, vol. 4868, pp. 92–103, 2008.

[3] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *LREC Workshop on Multimodal Corpora, Malta*, 2010.

[4] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[5] F. Eyben, M. Woellmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *J Multimodal User Interfaces*, vol. 3, pp. 7–19, 2010.

[6] Tomoki Toda, Alan W. Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, pp. 215–227, 2008.

[7] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroeder, "Feeltrace: An instrument for recording perceived emotion in real time," 2000.