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

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.

Keywords: continuous emotion tracking, dimensional emotional descriptions, Gaussian Mixture Model mapping, body language, improvised 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, 2, 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.

Email addresses: metallin@usc.edu (Angeliki Metallinou), nkatsam@sipi.usc.edu (Athanasios Katsamanis), shri@sipi.usc.edu (Shrikanth Narayanan)

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.

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 movement, 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., us-

ing 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

Figure 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., $\hat{y} = arg\max P(y|x)$. 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:

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

assuming a specific model $\lambda^{(y,\mathbf{x})}$ for two concurrent instantiations of $\mathbf{x}$ and y. However, given the continuous nature of the involved variables, it would be beneficial to incorporate dynamic information in this estimation. This can be achieved by also jointly modeling the first and second temporal derivatives of y_t and $\mathbf{x}_t$, denoted here as Δy_t, $\Delta^2 y_t$ and $\Delta\mathbf{x}_t$, $\Delta^2\mathbf{x}_t$ respectively. By replacing y_t with $\mathbf{Y_t} = [y_t, \Delta y_t, \Delta^2 y_t]^T$ and x_t with $\mathbf{X_t} = [\mathbf{x}_t^T, \Delta\mathbf{x}_t^T, \Delta^2\mathbf{x}_t^T]^T$, the optimal estimate $\hat{y} = [y_1, \ldots, y_t, \ldots y_T]$ of the emotional flow for the course of the interaction can be found as:

$$\hat{y} = \arg\max_{y} P(\mathbf{Y}|\mathbf{X}, \lambda^{(Y,X)}), \quad (2)$$

where $\mathbf{X} = [\mathbf{X}_1^T, \mathbf{X}_2^T, \ldots, \mathbf{X}_t^T, \ldots, \mathbf{X}_T^T]^T$ is the sequence of the dynamic information-augmented features and $\mathbf{Y} = [\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_t^T, \ldots, \mathbf{Y}_T^T]^T$ the corresponding emotional attribute and its derivatives for the entire interaction. Following the paradigm that was originally introduced for voice conversion [32], we consider the model $\lambda^{(Y,X)}$ of the joint probability of $(\mathbf{Y_t}, \mathbf{X_t})$ to be a Gaussian Mixture Model (GMM):

$$P(\mathbf{Y_t}, \mathbf{X_t}|\lambda^{(Y,X)}) = \sum_{m=1}^{M} a_m \mathcal{N}([\mathbf{Y_t}^T, \mathbf{X_t}^T]^T; \mu_m^{(Y,X)}, \Sigma_m^{(Y,X)}) \quad (3)$$

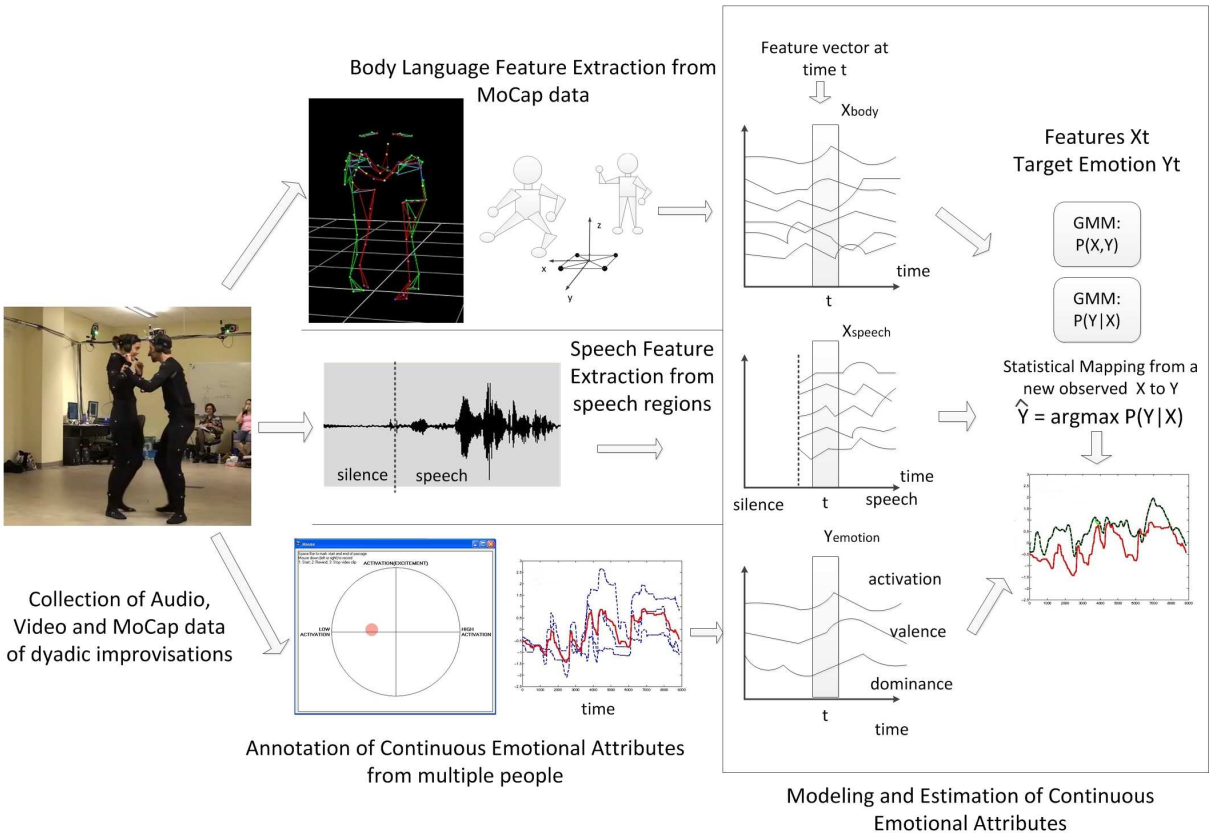


Figure 1: An overview of the work presented in this paper. From left to right we depict the data collection setting, the audio-visual feature extraction and data annotation processes, as well as the GMM-based statistical mapping approach that we follow for estimating the emotional curves.

with a_m, $\mu_m^{(Y,X)}$ and $\Sigma_m^{(Y,X)}$ being each component's weight, mean and covariance respectively:

$$\mu_m^{(Y,X)} = \begin{bmatrix} \mu_m^{(Y)} \\ \mu_m^{(X)} \end{bmatrix}, \Sigma_m^{(Y,X)} = \begin{bmatrix} \Sigma_m^{(YY)} & \Sigma_m^{(YX)} \\ \Sigma_m^{(XY)} & \Sigma_m^{(XX)} \end{bmatrix}. \quad (4)$$

The conditional probability in (2) can be written as [32]:

$$\begin{aligned} P(\mathbf{Y}|\mathbf{X}, \lambda^{(Y,X)}) &= \\ &\sum_{over\ all\ \mathbf{m}} P(\mathbf{m}|\mathbf{X}, \lambda^{(Y,X)}) P(\mathbf{Y}|\mathbf{X}, \mathbf{m}, \lambda^{(Y,X)}) \\ &\approx \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\mathbf{X}_t, \lambda^{(Y,X)}) P(\mathbf{Y}_t|\mathbf{X}_t, m, \lambda^{(Y,X)}) \quad (5) \end{aligned}$$

where $\mathbf{m} = [m_1, \ldots, m_t, \ldots, m_T]$ is a sequence of mixture components and:

$$P(m|\mathbf{X}_t, \lambda^{(Y,X)}) = \frac{a_m \mathcal{N}(\mathbf{X}_t; \mu_m^{(X)}, \Sigma_m^{(XX)})}{\sum_{i=1}^{M} a_i \mathcal{N}(\mathbf{X}_t; \mu_i^{(X)}, \Sigma_i^{(XX)})} \quad (6)$$

$$P(\mathbf{Y}_t|\mathbf{X}_t, m, \lambda^{(Y,X)}) = \mathcal{N}(\mathbf{Y}_t; \mathcal{E}_{m,t}^{(Y)}, D_m^{(Y)}). \quad (7)$$

and:

$$\mathcal{E}_{m,t}^{(Y)} = \mu_m^{(Y)} + \Sigma_m^{(YX)} \Sigma^{(XX)-1} (\mathbf{X_t} - \mu_m^{(X)}), \quad (8)$$

$$D_m^{(Y)} = \Sigma_m^{(YY)} - \Sigma_m^{(YX)} \Sigma_m^{(XX)-1} \Sigma_m^{(XY)}. \quad (9)$$

Estimation of the underlying emotional flow $\hat{y}$ for the entire utterance can finally be achieved based on (2) via Expectation Maximization as described in detail in [32, 4]. The initial estimate is just the Minimum Mean Squared Error (MMSE) estimate based on the conditional probability distribution (5) without using dynamic information. Due to the use of dynamic information in the estimations, the final estimate at each time instant ends up being affected by the entire sequence of observations. It has been shown that in the case of a single Gaussian Model the incorporation of derivatives in an analogous scenario corresponds to fixed-lag Kalman smoothing [33]. The lag depends on the window length

$2L-1$ over which the derivatives are approximated (second derivatives are computed by applying (10) to the first derivatives):

$$\Delta y_t = \frac{\sum_{\theta=-L}^{\theta=L} \theta(y_{t+\theta} - y_{t-\theta})}{2\sum_{\theta=-L}^{\theta=L} \theta^2}. \quad (10)$$

This scheme has been successfully applied for voice conversion [32], lip movement - speech synchronization [34] and acoustic to articulatory speech inversion [4]. Speech inversion refers to the problem of recovering the underlying articulation during speech production from just the observed speech acoustics. In a similar way, herein, we are trying to recover the underlying emotional state as it is represented by activation, valence and dominance from the observed body language and speech observations.

4. Database and Annotation process

4.1. Database Description

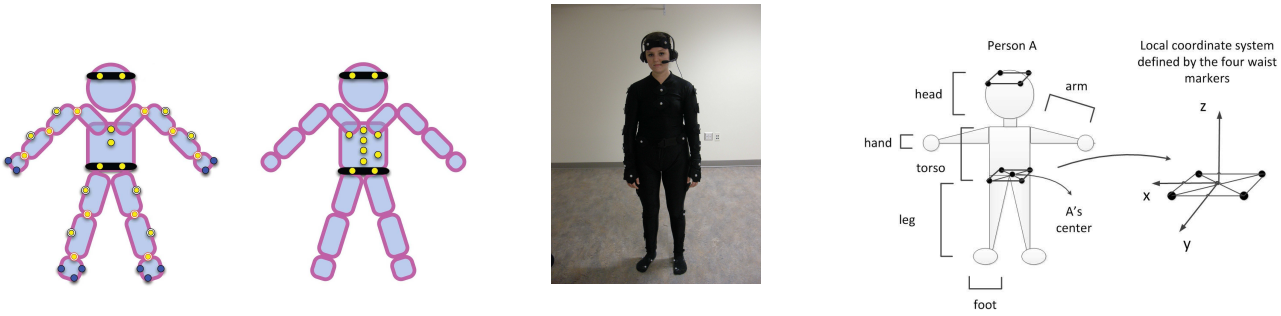
We use the USC CreativeIT database which is a multimodal database that combines engineering and theatrical approaches [6]. It contains a variety of dyadic theatrical improvisations and represents an opportunity to systematically study verbal and non-verbal expressions in affective interactions. Performances are either improvisations of scenes from theatrical plays or theatrical exercises where actors repeat sentences in a manner that conveys specific intent (e.g., accepting or rejecting behavior towards other). However, the actors were not instructed to produce specific emotions. Instead, we expect a variety of emotional expressions and interaction dynamics to occur as part of the performance. This design makes the emotional manifestations of the database especially challenging to analyze, since they are more subtle and diverse. The theatrical design was performed by a theater expert (director/teacher), and the participating actors were senior theater students, who first had to pass an audition. The performances were recorded under the guidance of the theater expert in order to ensure high quality performances. Further data collection details can be found in [6].

The database contains multimodal information from the vocal and body language behavior of the actors obtained through close-up microphones, Motion Capture (MoCap) cameras and HD cameras. Each actor wore a special suit and 45 MoCap markers were placed across his/her body, as illustrated in Figures 2(a) and (b). The performances were recorded by 12 Vicon MoCap cameras placed on the ceiling of the recording room, as well as two HD cameras located at each corner of the room. In this work we use data from 16 actors, 9 female and 7 male; 6 out of 8 dyads performed 6 improvisations, and the remaining two dyads performed 7 improvisations, resulting in 50 improvisations total. The extra improvisations were performed after the theater professor's request, who judged that those dyads' performances were excellent, and asked for an additional performance. Improvisations range from 2 to 10 minutes, while on average about 40% of an improvisation contains speech from one of the two participants. We capture audio-visual data from both actors in each improvisation, therefore we have a total of 100 actor-recordings. Our modeling is based on features which are extracted from MoCap and speech information. The videos of the performances have only been used for the data annotation process.

4.2. Annotation Process

The CreativeIT performances contain a variety of emotional manifestations. Each participant's emotional state is mapped into dimensional labels of activation, valence and dominance, which provides a continuous and generic description of the expressed emotions. Unlike speech-centric emotion databases (i.e., IEMOCAP [35] and VAM [8]), where it is common to segment a conversation into sentences as basic units for examining emotional content, in CreativeIT each performance is characterized by an unfolding flow of body gestures. This makes segmentation into sentences rather arbitrary. Therefore we decided to collect continuous annotations throughout each interaction, without segmenting the recordings, using the Feeltrace instrument, which allows real-time continuous annotation of video content [36]. Annotations are collected for each emotional attribute and for both actors of each performance, by watching the corresponding video recordings.

The problem of emotional data annotation has been addressed in other works including [25], where authors measure evaluator agreement through repeatedly comparing evaluator subsets, and [37] where the notion of implicit annotation is discussed. Here, the continuous nature of the annotation task represents an additional challenge in terms of obtaining agreement. Furthermore, recordings are long and require constant attention from the annotator, while the actors express a wide variety of emotions and have different roles throughout the interaction (speaker, listener, neither). Consequently, inter-annotator agreement is hard to achieve, as we observed in our previous work using a subset of the CreativeIT data [5]. Similar challenges have also been reported in other engineering studies that use continuous annotations [12], or examine expressive body language using discrete labels [38, 25].



(a) Marker positions. (b) Actor wearing markers. (c) Definition of body parts.

Figure 2: The positions of the Motion Capture markers and definitions of the body parts used in feature extraction

For our current study, we recruited psychology students, most of whom had previous experience in emotional annotation. Annotators were further trained by a short instruction session where Feeltrace was introduced and the definitions of activation, valence and dominance attributes were explained through examples. Annotators watched many recorded performances in advance in order to get an idea of the data. They performed their first annotations multiple times to familiarize themselves with Feeltrace and were later encouraged to perform each annotation as many times as needed until they were satisfied with the result. Since annotations are done real-time, there is expected to be a person-specific delay between the time that an event happens and when its emotional content is annotated. In order to reduce this delay, we modified the Feeltrace interface so that annotators can focus on one attribute each time, rather than two attributes, as was initially proposed in [36]. A snapshot of the modified Feeltrace interface for activation annotation is presented in Fig.3. The annotation is perfomed by moving the mouse, shown as a full circle, along the horizontal line, while watching the performance video in a separate window.[1] To further reduce person-specific delays, we also instructed annotators to watch each video multiple times and have a clear idea of the emotional content before starting the real-time annotation.

In Figure 4 we present a segment of the activation annotations of an actor provided by three annotators, and their average which is used as the ground truth. Note that although annotators agree on the trends of the activation curve (mean correlation of 0.67), and recognize

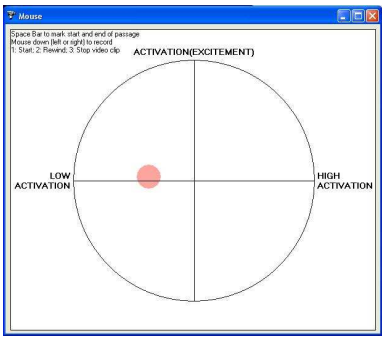


Figure 3: Screenshot of the modified Feeltrace interface.

pronounced activation events, they do not agree on the actual activation values. Similar observations hold true for many of our obtained annotations. This suggests that people tend to agree more when describing emotions in relative terms, e.g., whether there has been an increase or decrease, rather than in absolute terms (an observation which agrees with the literature, e.g., [7]). This motivated us to focus on the annotation trends, and to use correlation metrics, such as linear correlation, to measure evaluator agreement and the performance of the emotion tracking algorithms.

Seven annotators participated in total, rating overlapping portions of the database, so that each actor-recording would be rated by three or four people (88 out of the 100 actor-recordings were rated by 3 people). For computing the annotator correlations we set a cut-off threshold for defining acceptable annotator agreement. For each actor-recording, we take the union of all annotator pairs with linear correlations greater than the threshold; this annotator subset is used to compute the ground-truth for the corresponding actor-recording,

[1]Currently a one-dimensional version of Feeltrace is publicly available, in software Gtrace [39]. This software became available when we were midway in our data annotation, and we decided to keep our own modified Feeltrace version for consistency.

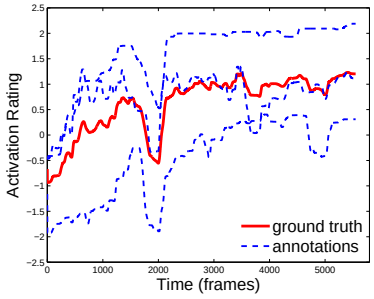


Figure 4: Example of activation rating by three annotators.

by taking the average of the selected annotations. If no annotators are selected then we exclude that actor-recording from our analysis. Our threshold is empirically set to 0.45, which results in selecting 80, 84 and 73 actor-recordings for the activation, valence and dominance class respectively, out of 100 in total (the rest were excluded from further analysis). The annotator agreement measure is computed by first computing the mean of the correlations between the selected annotators per actor recording, and then computing the median over all actor recordings. Median annotator correlations reached 0.59, 0.62 and 0.60 for activation, valence and dominance respectively (these numbers are higher than the ones achieved by our previous annotation effort [5]).

Our choice of averaging multiple annotations to provide ground truth is a common approach in the emotion recognition community, but could be problematic in cases when the actual attribute value is of interest, since different annotators often have different internal rating scales. Here we reduce the extent of this problem by focusing on the trends of the average curve; the trends of the evaluator curves are not affected as much by the mean operation and could be a more robust indicator of the underlying ground truth (see also Fig.4). However, effectively fusing multiple annotators' subjective judgements is an important research problem (e.g., see [40, 41]), and a direction for future research. Also, the issue of person-specific delays is a challenging issue and is worth further investigation in the future, e.g., by means of targeted experiments measuring such delays among annotators.

5. Feature Extraction and Selection

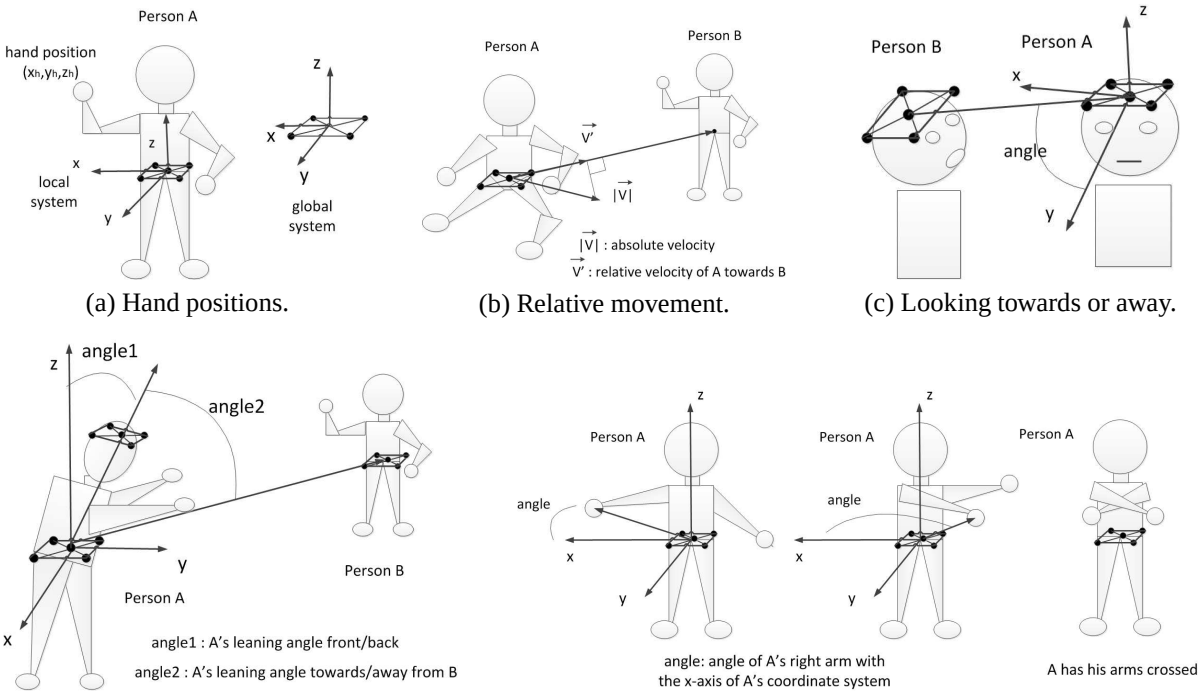
5.1. Body Language Feature Extraction

Our body language features are extracted from full body MoCap data (the performance videos are only used for data annotation). From now on, we will also refer to these MoCap features as visual features, since they are visually perceived. The choice of features is inspired by the psychology literature which indicates that behaviors such as looking at the interlocutor, approaching, touching, as well as body postures such as looking down, and hand gestures carry emotional information [42]. The features are extracted for each person, and they are either absolute descriptions of a person's posture and movement, or relative descriptions of his body behavior with respect to his interlocutor (in the latter case data from both people are used for the feature extraction).

In total, we examine 53 body language features, extracted at the MoCap framerate (60 fps) and smoothed using a median filter. This comprehensive feature set is summarized in Table 1, and may contain correlated or redundant features; decorrelated feature subsets will be later chosen through feature selection. Features are extracted in a geometrical manner from the positions of the MoCap markers, by defining global and local coodinate systems and measuring 3D distances, velocities and angles. These features are potentially informative either individually or in combination with each other. Our modeling framework can exploit such feature relations, as explained in Section 6.1. The origin of the global coordinate system is roughly the center of the recording space, while local coordinate systems for each actor are defined using the four waist markers, as shown in Fig.2(c). The positions of the various body parts are illustrated in Fig.2(c). For example, one's center is defined as the average of the four waist markers.

Certain features are particularly influenced by person-specific bodily characteristics. For example the z-coordinates of a person's upper and lower back, which may reflect crouching and sitting are influenced by the person's height. Therefore, features that are z-coordinate positions are normalized by dividing by the actor's median height in each recording. Additionally, features that are (x,y,z) positions of hands in one's coordinate system are normalized by dividing by the person's median arm length in each recording, measured by the median distance between shoulder and hand markers. All normalized features are denoted as 'norm' in Table 1. Apart from that, we do not perform any normalization of person-specific emotional variability, since our person-independent setup does not assume any prior information about the expressive characteristics of a test subject. Normalizing for such person-specific emotion variability would be an interesting future direction.

Figure 5 illustrates some example features. For instance, as shown in Fig.5(a), the position of one's center



(a) Hand positions. (b) Relative movement. (c) Looking towards or away.

(d) Leaning front/back, and leaning towards/away from Other. (e) Arm angles and description of arms crossed .

Figure 5: Examples of extracted features from MoCap markers.

is measured in the global system to decsribe his location, while positions of one's hands are measured in his local coordinate system to describe his hand gestures. An individual's absolute velocity is computed from the movement of his center, while relative velocity is computed by projecting the velocity vector in the direction between the two participants (Fig.5(b)). A description of one's looking behavior relative to his interlocutor is computed from the angle between the orientation of one's head coordinate system and the direction between the heads of the participants (Fig.5(c)). A description of one's relative body orientation can be obtained similarly by looking at the people's waist coordinate systems instead. In Fig. 5(d), the angle between a person's spine and his local z-axis describes his leaning front/back behavior, while the angle between one's spine and the direction between the centers of the participants describes relative leaning behavior (towards/away). The angles of one's arms with his local x-axis, describe hand position and indicate arms crossing behavior (Fig. 5(e)).

5.2. Feature Selection Approaches

We examine a variety of feature selection approaches to select a subset of decorrelated, informative body language features, tailored to each emotional attribute.

Mutual Information-based and Correlation-based criteria: The minimal redundancy maximal relevance criterion (mRMR), introduced in [43], selects features that maximize the mutual information (MI) between features and the ground truth, and minimize the MI between the selected features. Let $S_M = \{x_i\}_{i=1}^{M}$ be a set of M continuous body language features, y the continuous emotional attribute, and $I(\cdot,\cdot)$ represent MI. Then the mRMR measure is defined:

$$mRMR_I(i) = I(x_i, y) - \frac{1}{M-1} \sum_{x_j \in S_M, j \neq i} I(x_i, x_j) \quad (11)$$

$$where \quad I(x_i, y) = \sum_{x_i \in X_i} \sum_{y \in Y} p(x_i, y) log(\frac{p(x_i, y)}{p(x_i)p(y)}) \quad (12)$$

Estimation of the probability distributions $p(x_i), p(y), p(x_i, x_j)$ and $p(x_i, y)$, which is required for computing the MI values, is performed through uniform quantization.

We also examine the selection of maximal relevance and minimal redundancy features based on correlations rather that MI values. Specifically, if we denote as

Table 1: *Body language features extracted from actor A during his interaction with actor B. Features are denoted as individual when they describe only A's movement and posture information, and as interaction features when they describe the relative movement and posture of A with respect to his interlocutor B. Norm indicates that the corresponding feature has been normalized per actor recording.*

A's velocity (individual)	
• A's velocity (see Fig. 5(b)) • velocity of A's right/left arm • velocity of A's right/left foot	• relative velocity of A's right/left arm w. respect to A • relative velocity of A's right/left foot w. respect to A
A's body posture (individual)	
• A's body leaning angle front/back (see Fig. 5(d)) • A's body leaning angle right/left • A's body position in global coord. system: x,y, norm z coordinates (see Fig. 5(a)) • A's right/left hand position in A's local coord. system: norm x,y,z coordinates (see Fig. 5(a)) • distance between A's right and left hand • angle of A's right/left hand with x -axis in A's system (indicating arms crossed, see Fig. 5(e))	• head angle, looking up/down • distance between A's right/left hand and A's chest • distance between A's right/left hand and A's right/left hip • angle between A's right and left hands • norm z coordinate of A's right/left knee (indicating kneeling) • z coordinate of A's right/left foot (indicating jumping) • norm z coordinate of A's upper back (indicating upward vs crouched posture) • norm z coordinate of A's lower back (indicating sitting down)
A's distance from B (interaction)	
• A's distance from B • Min. distance between A's right/left hand and B's hands • Min. distance between A's right/left hand and B's torso	• Min. distance between A's right/left hand and B's head • Min. distance between A's right/left hand and B's back
A's velocity with respect to B (interaction)	
• A's relative velocity w. respect to B (see Fig. 5(b)) • Relative velocity of A's right/left hand w. respect to B	• Relative velocity of A's right/left foot w. respect to B
A's orientation with respect to B (interaction)	
• Angle of A's face w. respect to B (see Fig. 5(c)) • Angle of A's body w. respect to B (similar to Fig. 5(c), but for waist coord. systems)	• A's leaning angle towards or away from B (see Fig. 5(d)) • Position of A in B's coordinate system

$C(x_i, y)$ and $C(x_i, x_j)$ the linear (pearson) correlations between a feature and the ground truth, and between the two features, respectively, we can define the correlation-based metric as follows:

$$mRMR_C(i) = C(x_i, y) - \frac{1}{M-1} \sum_{x_j \in S_M, j \neq i} C(x_i, x_j) \quad (13)$$

Both approaches perform a ranking of features, where high values are preferred and they denote that the feature shares much information, or has high correlation, with the ground truth and shares little information, or has low correlation, with other selected features.

Fisher Criterion: Alternatively, we select features that discriminate between regions of high, low and medium values of the emotional ground truth. Intuitively these features reflect different body language behaviors across regions of different emotional content. Each attribute is quantized into three levels through k-means clustering, and the features that correspond to each level are collected. Fisher criterion, denoted as F_{value}, is the ratio between the within-class variance and the between-class variance for a feature, and scores highly those features that achieve small within-class variability and large between-class variability [44]. While the previously described correlation and MI based methods favor the selection of feature sets with low redundancy, the Fisher criterion may lead to redundant feature sets. Therefore, we further reduce our feature set, by excluding features, such that no feature pair has a correlation higher than a threshold (here we empirically selected a threshold of 0.8). When choosing between two competing, highly-correlated features, we pick the one with the largest F_{value}.

5.3. Vocal Feature Extraction

In contrast to body language features which are extracted throughout the recording session, the acoustic features are extracted only when the actors are speaking. For this purpose, the microphone signal obtained from each actor is first manually transcribed into regions where that actor is speaking and being silent. We extract 12 Mel Frequency Cepstral Coefficients (MFCCs) along with pitch and energy, using overlapping windows of length 30msec and framerate of 16.67msec (same as MoCap framerate). Such features are standard for speech emotion recognition [45].

6. Tracking Emotion Trends at Multiple Resolutions

6.1. GMM-based tracking at frame and window level

Our GMM-based tracking approach follows the mathematical framework described in Section 3.2. Ad-

ditionally, it takes into account that body language features are available throughout the interaction, while speech features are available only when the actor is speaking. Therefore, when audio-visual features are considered we compute two mappings: a visual mapping trained only with body language features and an audio-visual mapping trained with both body language and speech features. The audio-visual features are fused at the feature-level for training the audio-visual GMM. During testing, we apply the GMM mapping on overlapping windows. When only visual features are used we compute the visual mapping on each window irrespective if whether the window contains speech or not. When audio-visual features are used, we compute an audio-visual mapping for the windows where speech is present, otherwise we compute a visual mapping. Therefore, we again scan the total recording using visual information and, if available, speech information. As a result, the results of the visual and audio-visual experiments are comparable as they are computed on the same recordings, and the audio-visual results provide information about whether speech improves emotion tracking on top of the visual information.

Empirically, we confirmed that including dynamic features produces a smoother emotional trajectory estimate, since it considers a window of the emotional state and the feature vector centered at the frame of interest. In our implementation, the underlying emotional trajectory $y_t, t = 1, \ldots, T$ is estimated over consecutive overlapping windows of length 300 frames, with 150 frames overlap. Then curves obtained from neighboring windows are merged using the add-overlap algorithm, and are smoothed using a low-pass filter.

This approach computes detailed frame-by-frame emotional trajectory estimates. However, emotional states are slowly varying, therefore this degree of accuracy may not be necessary. Modeling body and speech features at such detail may lead to modeling of noise or gestures unrelated to emotion rather that emotionally informative audio-visual manifestations. This motivates the use of window-level tracking, where features and feature functionals are extracted over larger windows in an attempt to capture more meaningful emotional and gestural dynamics. In this case, the mapping function takes as input the functionals computed over a window and outputs the average emotional attribute value of that window. Specifically, we average the ground truth curves over overlapping windows of 3sec length and 2sec overlap. We also apply such windows on the audio-visual features, over which we extract a variety of statistical functionals, specifically: mean, standard deviation, median, minimum, maximum, range, skewness, kurtosis, the lower and upper quantiles (corresponding to the 25th and 75th percentiles) and the interquantile range. Therefore, we extract a potentially richer feature description by including statistical functionals over features. The feature vector dimensionality is reduced by PCA.

We train full covariance GMMs using 4 and 2 mixtures for frame and window-level tracking respectively (the method is not sensitive to the number of gaussian mixtures). The use of a full covariance matrix is important in order to capture relations between the various body language and speech features, and empirically leads to better performance. The joint feature-emotion GMM models were trained using the HTK Toolbox [46], while the subsequent EM equations for computing the statistical GMM-based mapping were implemented in matlab, based on [4].

6.2. Using LSTM neural networks for regression

Long Short Term Memory (LSTM) neural networks were introduced in [47], as a variant of Recurrent Neural Networks (RNN). While RNNs are able to model a certain amount of history through their cyclic connections, it has been shown that longer range history is inaccessible to RNNs since the backpropagated error either blows up or decays over time (vanishing gradient problem). LSTM networks overcome the vanishing gradient problem by storing in their hidden layers information from an arbitrarily long amount of time [47]. LSTM networks have been applied in a variety of pattern recognition applications, including phoneme classification [48], audio-visual emotion classification [49], and regression for tracking continuous emotions [13]. Modeling history seems to be beneficial for the problem of emotion tracking, since emotions tend to be slowly varying over time, and LSTM regression was shown to outperform Support Vector Regression (SVR) for continuously tracking valence and activation over time [13]. Here, we apply LSTM networks for both the frame and the window level regression problems.

LSTM networks for regression are trained using the RNNlib Toolbox [50], without using derivative features. Including derivatives was deemed redundant since temporal information is already captured through the network. The LSTM networks consist of one hidden layer with 128 memory blocks (we also experimented with 64 and 256 memory block configurations, which performed similarly). To improve generalization low Gaussian noise was added to the training features. The produced curves are smoothed using a lowpass filter.

6.3. Baseline based on simple functions of informative features

A relevant question is what would be the tracking performance if we estimated an attribute, e.g., activation, as a simple function of informative features, e.g., velocity of body, of hands, intensity of voice, leaning angle towards interlocutor etc. Indeed such approaches are common in the behavioral sciences, where for instance speech intensity and pitch are sometimes used as indicators of vocal activation [51]. Along these lines, assuming that an interlocutor's emotional attributes and his audiovisual features are normalized to be roughly in the same range, we could compute an estimate of his activation by taking a functional (e.g., mean) of the most activation-informative features. If a feature is negatively correlated with activation then we multiply it with -1 beforehand. This method does not require training a model, however it assumes that we have available a set of informative features for each attribute, which can be chosen through feature selection e.g, by using the approaches of Section 5.2, or through prior knowledge. This simple baseline could be useful for cases were we have few or no annotated data.

In our implementation, we select the K most informative body features for each attribute, based on the F_{value}, and the L most informative speech features based on correlation with the attribute (results based on the $mRMR_C$ criterion are similar and are omitted for lack of space). All the features and the emotional attributes are first normalized to have zero mean and unit standard deviation across the database, and features that are negatively correlated with the emotional attribute are multiplied with -1. Then we compute the mean, median and maximum of these features as different attribute estimates (please refer to the Appendix for a list of the most informative body language features per attribute). For the window-level tracking, we follow the same approach using normalized statistical functionals of body and speech features extracted over windows, so as to directly compare with the methods of Sections 6.1 and 6.2. Again, we select the K most correlated functionals of body features and the L most correlated functionals of speech features.

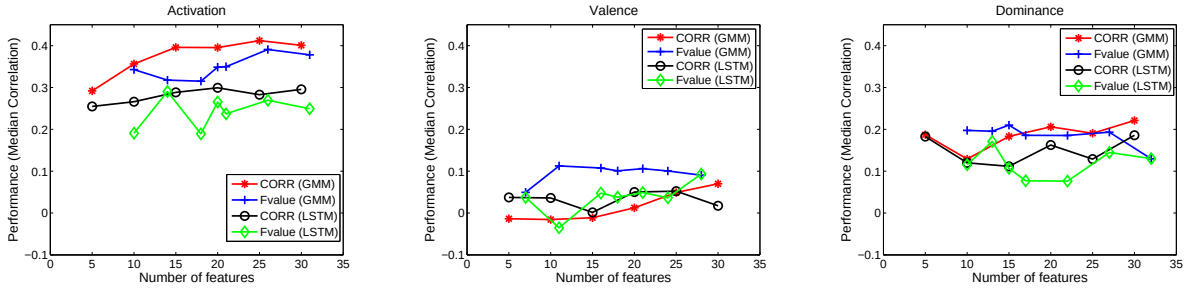
7. Experiments, Results and Discussion

Our experiments are organized in an eight-fold leave-one-dyad-out cross validation. Actors belong only to one dyad, therefore this cross validation ensures that test set actors are not seen during training. Each dyad was recorded in each of eight recording days, and since the selected number of recordings per day vary, this results in 5-12 actor recordings selected for testing at each fold, while the rest are used for training. We focus primarily on tracking the underlying emotional trends, and therefore we compute the correlations between the ground truth and the estimated emotional trajectories as our primary performance metric.

The body language feature sets are selected through the correlation-based criterion $mRMR_C$ or the Fisher criterion F_{value} (the MI-based criterion $mRMR_I$ gave slightly lower performance and is omitted). We systematically examine the effect of the number of body language features on the performance of each tracking approach, by selecting the top $5, 10, 15, \cdots 30$ features for the $mRMR_C$ criterion, and the top $10, 15, \cdots 40$ features for the F_{value} criterion (which are later further reduced after removing highly correlated features). The performance of the GMM-based and LSTM frame-level tracking as a function of the number of selected body features is shown in Fig. 6. This approach represents a principled way to select the final number of features based on visual frame-level tracking performance, although could cause some amount of overfitting. Note however that the selected number of features is not necessarily optimal for the audio-visual and window-level tracking, since some of the body language features that are left out may be important when used in combination with speech, or may have informative statistical functionals. To underscore this point, we present an example of valence tracking at Tables 2 and 4, where selecting 11 features is optimal for the frame-level visual experiments but not for the audio-visual and window-level experiments, where larger feature sets, e.g of 24 features, perform better.

For GMM-based and LSTM frame-level tracking, we select the number of body language features that leads to the best performance. We include speech information, by adding the 14 speech features in our body language feature set (feature-level fusion). We also add the first and second feature derivatives. For window-level tracking, we perform statistical functional computation on the respective optimal frame-level feature set and then Principal Component Analysis (PCA), keeping the first 50 components, which explain about 88-95% of the total variability. To prevent oversmoothing, we only add first derivatives, resulting in 100 dimensional feature vectors for both the visual and the audio-visual case. Both the features and the emotional curves are z-normalized using the global means and standard deviations of the dataset.

Regarding the simple baseline described in Section 6.3 for frame or window-level tracking, we select the



(a) Activation. (b) Valence. (c) Dominance.

Figure 6: Frame-level tracking using body language features: Performance of the various tracking approaches and feature selection algorithms (in terms of median correlation with ground truth) as a function of the number of body language features used.

number of body language and speech features that empirically give the best performance, and we combine them using their mean (which tends to perform better than median and maximum). Our observation it that the performance of this simple approach saturates sooner than the other algorithms, typically around 10 or 15 features.

7.1. Frame-level tracking using audio-visual information

In Table 2, we present the tracking performance of visual and audio-visual features methods for the GMM-based mapping and the LSTM regression approaches. The number of selected body language features is presented in parentheses. For the simple baseline method, the selected number of K body and L speech features is presented in paretheses as (K+L). For each case, we present the median of the correlations between each estimated curve and the ground truth, as a metric of the overall performance. In the last row of Table 2, we also report the median inter-annotator correlations computed at the frame-level, as described in Section 4.2.

For all methods, activation tracking is the best performing task, followed by dominance, while neither of the approaches seems to adequately capture valence trends. Considering speech features increases activation (speech features generally convey activation information [52]) and slightly boosts dominance tracking performance but offers no significant increase for valence. Both feature selection criteria perform comparably.

For valence, the F_{value} resulted in selecting a relatively small body language feature set of 11 features, therefore we also tried a larger feature set of 24 features to see if extra features would increase performance at later stages. Indeed the extra features and their statistical functionals seem to slightly boost performance at window-level tracking (see results of Section 7.2 Table 4), however valence tracking generally remains problematic. This suggests that valence is not adequately reflected on our features, or that body language generally conveys less information about valence, compared to activation and dominance. Valence may be better reflected through other modalities; for instance facial expressions are found to discriminate valence states well [53, 11]. Note that when annotators rated each actor's valence they had access to a variety of cues besides body language and speech, including facial expressions and lexical content, a fact that could explain their good agreement scores.

Between the tracking approaches, the GMM-based mapping achieves consistently higher correlation for activation and dominance. We performed the non-parametric Wilcoxon signed-rank test to examine whether the median of the paired differences between algorithms is significantly different from zero. Specifically, we compared the GMM and LSTM approaches, given same feature selection method, the GMM approach with the simple baseline, and the LSTM approach with the simple baseline (p=0.05). Statistically significant differences are denoted in Table 2 with symbol * for the GMM vs LSTM comparison, † for the GMM vs simple baseline comparison and $^\diamond$ for the LSTM vs simple baseline comparison (symbols are placed next to the method that performs better in the comparison). For example, for the frame-level tracking of activation using bodylanguage features, symbols * and † next to GMM tracking(F_{value}) indicate that the algorithm performs significantly better than both LSTM (F_{value}) and the simple baseline. Overall, the GMM-based mapping significantly outperforms both the LSTM method and the simple baseline for most ac-

Table 2: *Continuous tracking at the **frame-level** of activation, valence and dominance using body language and speech cues. We present the median correlation value between the computed emotional curve and the ground truth. Parentheses indicate the number of selected body features (K), or body and speech features (K+L).*

body language features: median correlations with ground truth			
feature selection	**activation**	**valence**	**dominance**
	GMM-based mapping		
F_{value}	0.3910 (26)$^{\star\ \dagger}$	0.1127 (11)/ 0.1005 (24)	0.2102 (15)*
$mRMR_C$	0.4121 (25)$^{\star\ \dagger}$	0.0699 (30)	0.2212 (30)
	LSTM regression		
F_{value}	0.2905 (14)	0.0934 (28)	0.1712 (13)
$mRMR_C$	0.2994 (20)$^{\circ}$	0.0524 (25)	0.1859 (30)
	simple baseline		
mean	0.2634 (10)	0.0650 (10)	0.1629(15)
body language+speech features: median correlations with ground truth			
feature selection	**activation**	**valence**	**dominance**
	GMM-based mapping		
F_{value}	0.4629 $^{\star\ \dagger}$	0.1178 / 0.1220 $^{\star\ \dagger}$	0.2495 $^{\star\ \dagger}$
$mRMR_C$	0.4692 $^{\star\ \dagger}$	0.0756	0.2582 $^{\star\ \dagger}$
	LSTM regression		
F_{value}	0.2908	0.0619	0.1610
$mRMR_C$	0.3874 $^{\circ}$	0.0842	0.1596
	simple baseline		
mean	0.3000 (10+5)	0.0815(10+5)	0.1829 (5+5)
Median inter-annotator correlation (agreement)			
	activation	valence	dominance
	0.5945	0.6171	0.6028

tivation and dominance tasks. However, LSTM tracking hardly outperforms the simple baseline, which works reasonably well for the activation and dominance tasks.

In order to examine how these methods approximate the actual values of the underlying emotional curves, we also compute the Root Mean Square Error (RMSE) between the estimated curve and the ground truth, which is defined as:

$$RMSE(\hat{y}_{est}, y_{true}) = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\hat{y}_{est}(i) - y_{true}(i))^2}$$

All methods lead to median RMSE methods between 0.8 and 1.2, with the GMM-based mapping usually having a slightly lower RMSE. Those values are considerably higher that the median RMSE values computed between the annotation curves of multiple evaluators, which are 0.37, 0.24 and 0.31 for activation, valence and dominance, respectively.

In Table 3, we also present results based on speech features only. Audio-only GMM-based tracking works reasonably for activation and partially for dominance, which confirms our previous observations regarding the importance of speech for activation trend tracking. Note however that these results are computed only on speech regions, therefore they are not directly comparable with the results of Table 2.

The behavior of the two methods is illustrated in Fig. 7. In Figures 7(a)-(c), we present the multiple annotations (dashed blue lines) along with their mean (red line) which is our ground truth, for two activation and one dominance example. Figures 7(d)-(f) show the estimated curves for GMM-based tracking, LSTM and the simple baseline respectively, for the curve of Fig. 7(a). For this example, GMM-based mapping produces a curve that is smoother and has higher correlation with the ground truth than the other two methods. Figures 7(g)-(i) show the estimated curves for Fig. 7(b), where the GMM-based performance is moderate but the method seems to track the most prominent activation trends. Finally, Figs. 7(j)-(l) show examples of dominance tracking for curve of Fig. 7(c), where all methods perform reasonably well, although the three output curves look quite different. In general, we notice that the GMM method produces smooth and flat curves, while the other two methods produce noisier curves of larger amplitudes.

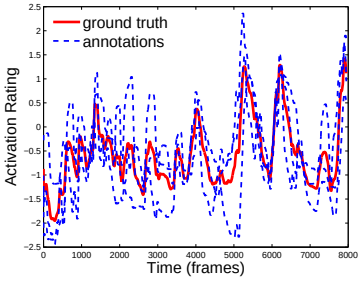
Table 3: *Continuous tracking at the **frame-level** of activation, valence and dominance using speech cues only. We present the median correlation value between the computed emotional curve and the ground truth, **computed only on speech regions**.*

speech features only: median correlations with ground truth		
activation	**valence**	**dominance**
	GMM-based mapping	
0.3866	0.0501	0.1102
	LSTM regression	
0.2237	0.0609	0.0066
	simple baseline (mean)	
0.1823 (5)	0.0529 (5)	0.0093 (5)

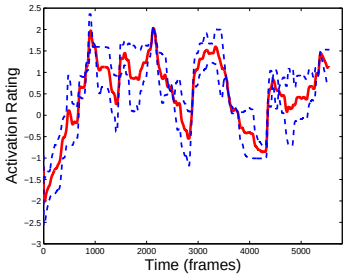
7.2. Window-level tracking using audio-visual information

In Table 4 we present the performance of the low resolution tracking at the window level. The median annotation correlations are re-computed at the window level and are reported at the last row of Table 4. For GMM-based and LSTM tracking we utilize the empirically selected feature sets of Section 7.1, after statistical feature extraction and PCA. For the simple baseline, we present the better performing statistical functionals of K body and L speech features.

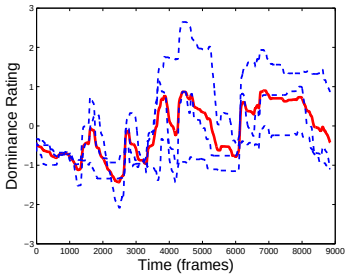
In general, we notice a significant increase from the previous results which can be attributed to the fact that we model less noise and track pronounced trends in the underlying emotional curves. Also we use a richer feature set, consisting of statistical functionals of the frame-level features. The GMM-based mapping results follow similar trends as before; activation is the best performing attribute, followed by dominance. Valence performance is still low, although when we use the Fisher



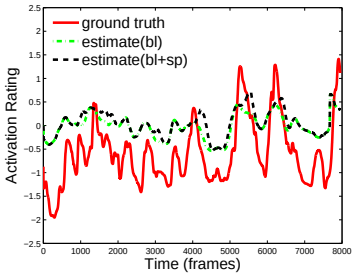
(a)Activation Example Annotations (blue) and their mean (red). Mean evaluator correlation: 0.44



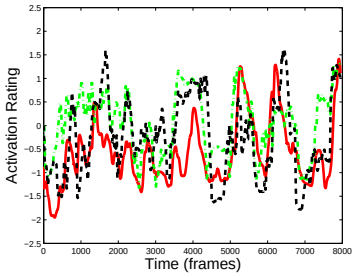
(b)Activation Example Annotations (blue) and their mean (red). Mean evaluator correlation: 0.68



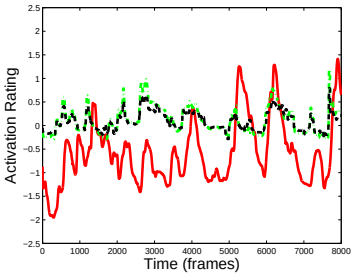
(c)Dominance Example Annotations (blue) and their mean (red). Mean evaluator correlation: 0.57



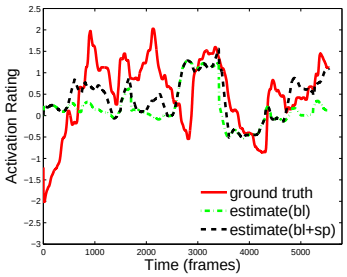
(d)Tracking of Activation Curve (a) using GMM-based mapping, with body (green) and speech+body (black) features. Correlations with ground truth are 0.74 and 0.77 respectively



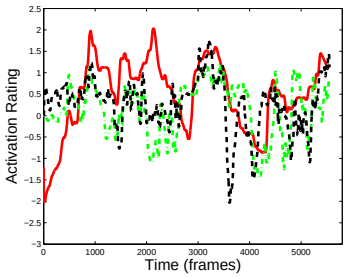
(e)Tracking of Activation Curve (a) using LSTM regression, with body (green) and speech+body (black) features. Correlations with ground truth are 0.60 and 0.52 respectively



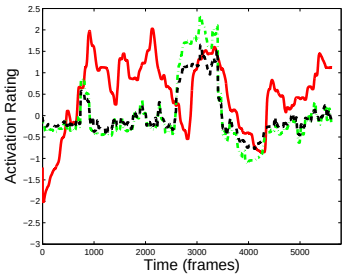
(f)Tracking of Activation Curve (a) using the simple baseline (mean), with body (green) and speech+body (black) features. Correlations with ground truth are 0.26 and 0.30 respectively



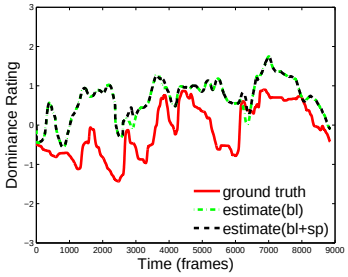
(g)Tracking of Activation Curve (b) using GMM-based mapping, with body (green) and speech+body (black) features. Correlations with ground truth are 0.25 and 0.43 respectively



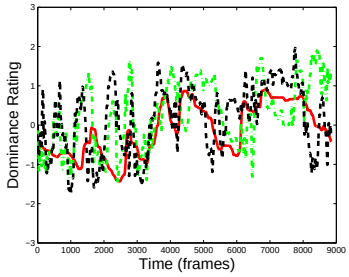
(h)Tracking of Activation Curve (b) using LSTM regression, with body (green) and speech+body (black) features. Correlations with ground truth are 0.25 and 0.29 respectively



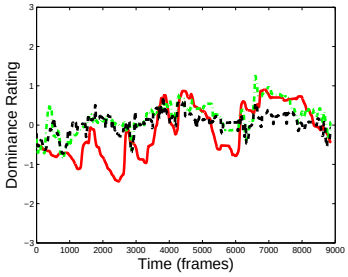
(i)Tracking of Activation Curve (b) using the simple baseline (mean), with body (green) and speech+body (black) features. Correlations with ground truth are 0.32 and 0.32 respectively



(j)Tracking of Dominance Curve (c) using GMM-based mapping, with body (green) and speech+body (black) features. Correlations with ground truth are 0.59 and 0.63 respectively



(k)Tracking of Dominance Curve (c) using LSTM regression, with body (green) and speech+body (black) features. Correlations with ground truth are 0.51 and 0.50 respectively



(l)Tracking of Dominance Curve (c) using the simple baseline (mean), with body (green) and speech+body (black) features. Correlations with ground truth are 0.68 and 0.48 respectively

Figure 7: Results of the three tracking methods, GMM-based mapping, LSTM regression and the simple baseline, for activation, and dominance cases, for frame-level tracking

criterion F_{value} with the larger feature set our performance increases. Adding speech features considerably increases activation and dominance performance. Activation tracking reaches a median correlation of around 0.6, which is similar to the median correlations between human annotators for this task. The LSTM regression and simple baseline results follow similar trends, although median correlations are generally lower.

The statistical significance of these results is examined using the Wilcoxon signed-rank test for paired differences, following the same notation as in Section 7.1. In general, GMM-based tracking significantly outperforms the other two approaches for activation and dominance trend tracking, while LSTM and simple baseline have comparable performance, with LSTM being slightly better.

Again, when looking at the resulting curves we observe smooth and flat curves for the GMM-based method and noisier curves with bigger amplitute for the LSTM and simple baseline methods. Figures 8(a)-(c) illustrate examples of rated activation, valence and dominance respectively. In Figures 8(d)-(f) we present the window-level tracking of activation curve 8(a), where all methods perform well, while the GMM-based curve achieves the highest correlation with the ground truth. In Figures 8(g)-(i) we present less successful tracking results of the valence curve in Fig. 8(b). The GMM-based mapping captures few of the valence peaks, while the other two methods seem to mostly capture noise. Finally, in Figs 8(j)-(l) we present tracking of the dominance curve 8(c), where GMM-based tracking performs better than LSTM, which in turn ourperforms the simple baseline.

Table 4: *Continuous tracking at the **window-level** of activation, valence and dominance using body language and speech cues. We present the median correlation value between the computed emotional curve and the ground truth*

body language features: median correlations with ground truth			
feature selection	**activation**	**valence**	**dominance**
GMM-based mapping			
F_{value}	0.4943 * †	0.1296 / 0.2061	0.3268 †
$mRMR_C$	0.5169 * †	0.0866	0.3219 * †
LSTM regression			
F_{value}	0.4455	0.1348	0.2268 $^{\circ}$
$mRMR_C$	0.4529 $^{\circ}$	0.1480	0.2835 $^{\circ}$
simple baseline			
mean	0.3682(10)	0.0626(15)	0.0953(15)
body language+speech features: median correlations with ground truth			
feature selection	**activation**	**valence**	**dominance**
GMM-based mapping			
F_{value}	0.5979 * †	0.1831 / 0.2247	0.3696 * †
$mRMR_C$	0.5837 * †	0.0563	0.3368 * †
LSTM regression			
F_{value}	0.4882	0.0976	0.2122
$mRMR_C$	0.4934 $^{\circ}$	0.0878	0.2549
simple baseline			
mean	0.4447(10+5)	0.1261 (15+5)	0.1837(5+5)
Median inter-annotator correlation (agreement)			
	activation	valence	dominance
	0.6199	0.6317	0.6200

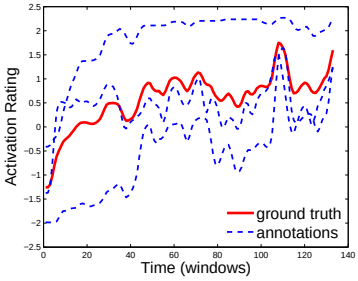
7.3. Discussion of informative body language features

This section discusses the selected body language features, to provide insights about the body language gestures, movements and postures that are informative of the underlying emotional attributes. Details about the top ranking body language features, according to F_{value} criterion, are presented in Appendix A, Tables A.5, A.6 and A.7. We omit detailed analysis of the $mRMR_C$ selected features; similar observations can be made for the activation and dominance tasks.

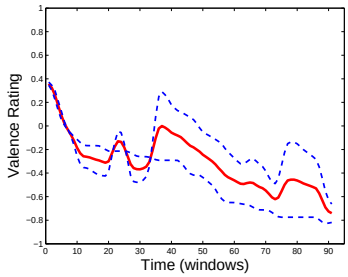
As seen in Table A.5 for activation, many of the the selected features describe absolute velocities, relative body orientation and leaning, posture and hand gestures. Highly activated subjects generally display higher arm and foot velocities (feats 4,20,21), more leaning and body orientation towards the interlocutor (feats 1,5), and more front leaning (feat 9) among others. Also many selected features describe hand gestures, for example hands tend to be further from the body (feats 3,6,7,10,19), further from each other (12,22), and raised higher (24,25) for highly activated subjects. Also, body location in (x,y) coordinates reflects a tendency of activated participants to be at the center of the recording space (feats 11,13).

For the dominance task, according to Table A.6, many of the selected features are common with the activation features, however we notice a preference for features describing relative behaviors like velocity, leaning and orientation. For example dominant individuals tend to lean and have body orientation more towards interlocutor (feats 1,4), and move their body, arms and feet more towards interlocutor (feats 8,17,20,22,24). This seems intuitive since dominance essentially captures relative (interaction) behavior. Also, dominant subjects tend to touch the interlocutor (feat 10), which brings to mind psychological observations relating touching with dominant behavior [54].

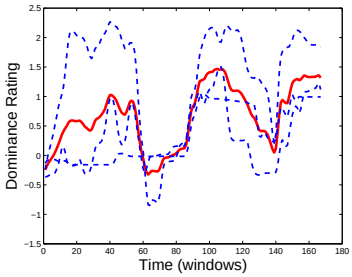
Finally, for the valence task, some features from Table A.7 stand out. For instance, positively valenced subjects tend to place hands on chest (feats 22,23), or touch the interlocutor's hand (feat 15), which seem to be intuitive bodily expressions of valence. Also positively valenced subjects tend to look more towards and move towards the interlocutor (feats 9,21), and move their arms and feet more (feats 2,13,14). Also the com-



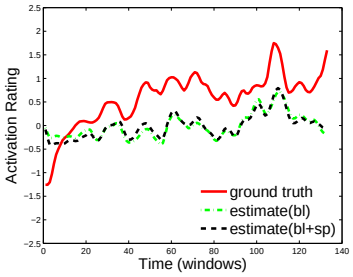
(a)Activation Example Annotations (blue) and their mean (red). Mean evaluator correlation: 0.54



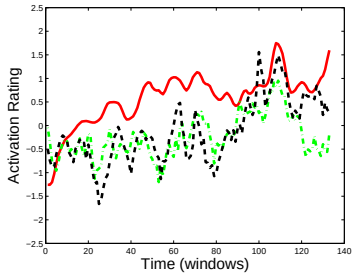
(b)Valence Example Annotations (blue) and their mean (red). Mean evaluator correlation: 0.46



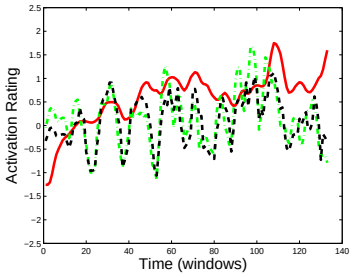
(c)Dominance Example Annotations (blue) and their mean (red). Mean evaluator correlation: 0.51



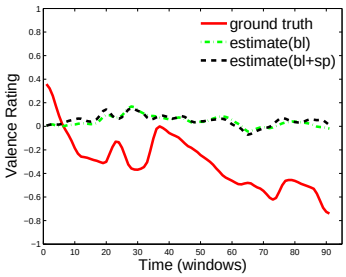
(d)Tracking of Activation Curve (a) using GMM-based mapping, with body (green) and speech+body (black) features. Correlations with ground truth are 0.57 and 0.74 respectively



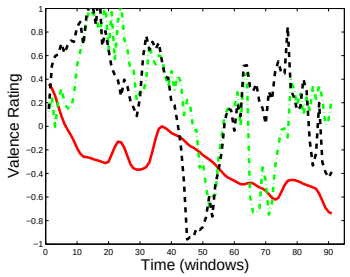
(e)Tracking of Activation Curve (a) using LSTM regression, with body (green) and speech+body (black) features. Correlations with ground truth are 0.49 and 0.58 respectively



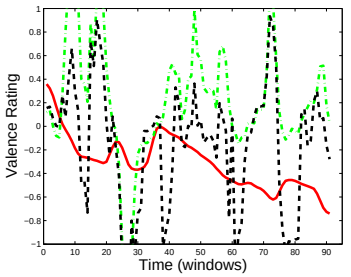
(f)Tracking of Activation Curve (a) using the simple baseline (mean), with body (green) and speech+body (black) features. Correlations with ground truth are 0.11 and 0.41 respectively



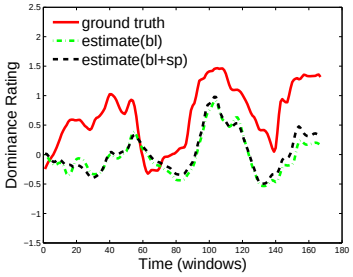
(g)Tracking of Valence Curve (b) using GMM-based mapping, with body (green) and speech+body (black) features. Correlations with ground truth are 0.24 and 0.21 respectively



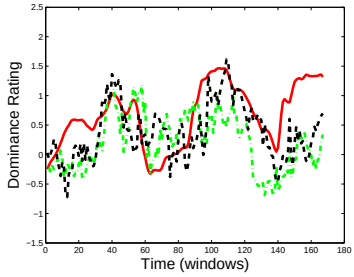
(h)Tracking of Valence Curve (b) using LSTM regression, with body (green) and speech+body (black) features. Correlations with ground truth are 0.31 and 0.17 respectively



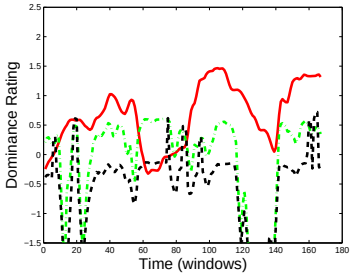
(i)Tracking of Valence Curve (b) using the simple baseline (mean), with body (green) and speech+body (black) features. Correlations with ground truth are 0.02 and 0.03 respectively



(j)Tracking of Dominance Curve (c) using GMM-based mapping, with body (green) and speech+body (black) features. Correlations with ground truth are 0.66 and 0.71 respectively



(k)Tracking of Dominance Curve (c) using LSTM regression, with body (green) and speech+body (black) features. Correlations with ground truth are 0.29 and 0.46 respectively



(l)Tracking of Dominance Curve (c) using the simple baseline (mean), with body (green) and speech+body (black) features. Correlations with ground truth are 0.11 and 0.12 respectively

Figure 8: Results of the three tracking methods, GMM-based mapping, LSTM regression and the simple baseline, for activation, valence and dominance cases, for window-level tracking

bination of more leaning towards others (feat 20), but less front leaning (feat 19) for positive valence, indicates that positively valenced subjects tend to lean more towards the interlocutor, while negatively valenced subjects generally have a more slouched posture.

Some of the above affective body language behaviors agree with the literature, for example arms being far from the body for high activation, or increased body motion for activated emotions such as anger ([31], Table 2). However, direct comparisons are hard to make since most past works on body language examine predefined categorical emotional states rather than continuous emotional attributes. Other aspects that differentiate this work from the literature include examining dominant behaviors, which are generally less discussed, as well as the focus on interaction aspects of body language through the introduction of 'relative' body features.

8. Conclusion and Future Work

We address the problem of tracking continuous emotional attributes of participants throughout affective dyadic improvisations, where participants may be listening, speaking or doing neither. To this end, we have examined interpetable features describing of a person's body language, and speech information. These descriptions complement existing literature, e.g., [22, 25, 28, 29], in capturing a wide range of full body gestures and emphasizing the interactive aspects of body language in dyadic emotional interactions. We propose a statistical mapping approach to automatically track emotional trends based on body language and speech. Our approach outperforms other examined methods, such as LSTM regression [13], and produces smooth emotional curve estimates. Also, the simple baseline represents an interesting, unsupervised alternative, that is worth further investigation. Our results show promising performance for tracking trends of activation and dominance, and also suggest that body language conveys rich activation and dominance related information. For activation trend tracking our correlation-based performance is comparable to human performance. Finally, analysis of our body language features offers quantitative insights on the relations between an underlying emotional state and the displayed bodily behavior in the context of dyadic interaction. This enables us to draw connections with psychological observations regarding body language and emotion.

However, valence trend tracking remains problematic, which might indicate that our features are not adequately reflective of valence. Existing literature indicates that body posture is a better indicator of activation, although the importance of the valence dimension should not be dismissed [25]. Possibly higher-level body features are required to discern valence; we have not incorporated audio-visual cues at the session level, such as the amount and length of pauses, percentage of time that an actor performs an action, turn-taking patterns etc. Such higher-level cues may be informative of valence and dominance, and their investigation is a promising future research direction. Also note that we do not consider facial expressions, which are known to be reflective of valence [53, 11].

Other open questions pertain to our performance metrics; while correlation metrics and RMS errors describe different aspects of tracking performance and are currently used for evaluating systems that produce continuous estimates [13, 12], we may need to find more accurate measures to describe the performance of such systems. Additionally, normalizing for subject-dependent emotional variability in expressive body language is an interesting research direction that could potentially bring significant improvement. A further goal is to extend this work towards examining the produced emotional curves to detect regions of emotional saliency, and study the actual events that occur in such regions. Such vocal, bodily or interaction-based events could give us insights of what consitutes the emotional content of an interaction.

Appendix A. Top Ranked Body Language Features for Emotion Discrimination

In Tables A.5, A.6 and A.7 we present the top ranked 25 body language features for activation, valence and dominance, according to the F_{value} criterion. Detailed results of the $mRMR_C$ criterion are ommited for lack of space, however we include the $mRMR_C$-based rank next to each feature (notice the overlap between the features of the two criteria for activation and dominance, although not for valence). Each feature value represents a meaningful body posture. For performing statistical tests, we quantize each attribute value into 3 classes using the k-means algorithm, and collect the feature instances that correspond to the high and low classes, over the total database. For each feature, we perform a t-test to compare the mean feature value between low and high emotional attribute classes. We also include a description of the corresponding difference in body language that each feature value represents, always comparing high (or positive) versus low (or negative) attribute values. For example, the first line (feature of rank

1) of Table A.5 can be interpreted as 'more leaning towards the interlocutor when subject is characterized by high activation vs no leaning when subject is characterized by low activation'. All feature mean differences are statistically significant, although in some cases mean differences are so small that do not correspond to a recognizable difference in body language (e.g., see feat. 8 at A.5, or feat. 3 at A.7).

References

[1] J. Russell, A. Mehrabian, Evidence for a three-factor theory of emotions, Journal of Research in Personality 11 (1977) 273–294.

[2] H. Schlosberg, Three dimensions of emotion, Psychology Review 61 (1954) 81–88.

[3] M. Greenwald, E. Cook, P. Lang, Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli, Journal of Psychophysiology 3 (1989) 51–64.

[4] T. Toda, A. W. Black, K. Tokuda, Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model, Speech Communication 50 (2008) 215–227.

[5] A. Metallinou, A. Katsamanis, Y. Wang, S. Narayanan, Tracking changes in continuous emotion states using body language and prosodic cues, in: Proc. of ICASSP 2010.

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

[7] Y.-H. Yang, H. H. Chen, Ranking-based emotion recognition for music organization and retrieval, IEEE Trans. on Audio, Speech and Language Processing 19 (2011) 762 – 774.

[8] M. Grimm, E. Mower, K. Kroschel, S. Narayanan, Primitives based estimation and evaluation of emotions in speech, Speech Communication 49 (2007) 787–800.

[9] D. Wu, T. Parsons, E. Mower, S. S. Narayanan, Speech emotion estimation in 3d space, in: Proc. of IEEE Intl. Conf. on Multimedia & Expo (ICME) 2010.

[10] H. Meng, N. Bianchi-Berthouze, Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models, in: Proc. of ACII, 2011.

[11] A. Metallinou, M. Woellmer, A. Katsamanis, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive learning for enhanced audiovisual emotion classification, IEEE Trans. of Affective Computing to appear (2012).

[12] N. Malandrakis, A. Potamianos, G. Evangelopoulos, A. Zlatintsi, A supervised approach to movie emotion tracking., in: Proc. of ICASSP 2011.

[13] M. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, IEEE Trans. on Affective Computing (2011).

[14] M. Woellmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies, in: Proc. of Interspeech 2008.

[15] M. Wollmer, B. Schuller, F. Eyben, G. Rigoll, Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening, IEEE Journal of Selected Topics in Signal Processing 4 (2010) 867–881.

[16] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, IEEE Trans. On Multimedia 7 (2005) 143–154.

[17] A. Hanjalic, Extracting moods from pictures and sounds: Towards truly personalized TV, IEEE Signal Processing Magazine (2006) 90–100.

[18] N. Sebe, I. Cohen, T. Huang, Multimodal Emotion Recognition, Handbook of Pattern Recognition and Computer Vision, World Scientific, 2005.

[19] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, Trans. of Pattern Analysis and Machine Intelligence 31 (2009) 39–58.

[20] H. Gunes, M. Piccardi, Bi-modal emotion recognition from expressive face and body gestures, Journal of Network and Computer Applications 30 (2007) 1334–1345.

[21] H. Gunes, M. Piccardi, Automatic temporal segment detection and affect recognition from face and body display, IEEE Trans. on Systems, Man, and Cybernetics - Part B, Special Issue on Human Computing 39 (2009) 64–84.

[22] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, K. Scherer, Towards a minimal representation of affective gestures, IEEE Trans. on Affective Computing 2 (2011) 106–118.

[23] G. Castellano, L. Kessous, G. Caridakis, Multimodal emotion recognition from expressive faces, body gestures and speech., in: Proc. of ACII, 2007.

[24] G. Castellano, S. Villalba, A. Camurri, Recognising human emotions from body movement and gesture dynamics, in: Proc. of ACII, 2007.

[25] A. Kleinsmith, N. Bianchi-Berthouze, A. Steed, Automatic recognition of non-acted affective postures, IEEE Trans. on Systems, Man and Cybernetics, Part B 41 (2011) 1027–1038.

[26] G.Varni, G.Volpe, A.Camurri, A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media, IEEE Trans. on Multimedia 12 (2011) 576–590.

[27] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. Georgiou, S. Narayanan, Estimation of ordinal approach-avoidance labels in dyadic interactions: ordinal logistic regression approach., in: In Proc. of ICASSP.

[28] D. Bernhardt, P. Robinson, Detecting affect from non-stylised body motions, in: Proc. of ACII 2007.

[29] A. Kleinsmith, N. Bianchi-Berthouze, Recognizing affective dimensions from body posture, in: Proceedings of the 2nd Intl Conf on Affective Computing and Intelligent Interaction (ACII).

[30] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. McOwan, A. Paiva, Automatic analysis of affective postures and body motion to detect engagement with a game companion, in: Proc. of ACM/IEEE Intl Conf. on Human-Robot Interaction, 2011.

[31] A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: A survey., IEEE Transactions on Affective Computing (2012) 1–20.

[32] T. Toda, A. Black, K. Tokuda, Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, IEEE Trans. on Audio, Speech and Language Processing 15 (2007) 2222–2235.

[33] Y. Minami, A. McDermott, E.and Nakamura, S. Katagiri, A theoretical analysis of speech recognition based on feature trajectory models, in: Proc. of Interspeech, 2004.

[34] L. Zhuang, Xiaodanand Wang, F. K. Soong, M. Hasegawa-Johnson, A minimum converted trajectory error (MCTE) approach to high quality speech-to-lips conversion, in: Proc. of Interspeech, 2010.

[35] C. Busso, M. Bulut, C.-C. Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S.Lee, S.Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (2008) 335–359.

[36] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, M. Schröder, FEELTRACE: An instrument for

Table A.5: *Statistical analysis of the top 25 activation features, according to the F_{value} criterion (each feature's rank according to the $mRMR_C$ criterion is included in the second column). The feature descriptions under the statistical tests column are describing high activation behavior compared to low activation behavior of a subject A. The statistical test performed is difference of means of the feature values between high and low activation classes (t-test)*

Activation: Comparison of high vs low activation classes			
		F_{value} **(and** $mRMR_C$**)**	
rank F_{value}	**rank** $mRMR_C$	**feature**	**description of statistical tests results** $p \approx 0$
1	1	A's body lean towards/away from B	more lean towards vs no leaning
2	9	norm x coord of A's right hand in A's system	x coord higher (further from body towards right, see also Fig. 5(a))
3	7	distance of A's left hand from A's hip	greater distance
4	6	abs velocity of A's right arm	higher velocity
5	22	relative angle of A's body towards B	body orientation more towards B vs sideways
6	27	norm y coord of A's left hand in A's system	y coord higher (further from body towards front, see also Fig. 5(a))
7	12	distance of A's right hand from A's hip	greater distance
8	2	A's body leaning angle, left/right	slightly lean right vs straight (though angle in both cases is close to zero)
9	4	A's body leaning angle, front/back	more lean front vs no leaning
10	28	norm y coord of A's right hand in A's system	y coord higher (further from body towards front, see also Fig. 5(a))
11	3	x coord of A's center	x abs value lower (x more towards center (0,0,0) of the recording space)
12	34	distance between A's right and left hand	hands wider apart
13	11	y coord of A's center	y abs value lower (y more towards center (0,0,0) of the recording space)
14	14	norm z coord of A's upper back	higher, more upwards posture, also indicates less sitting
15	48	distance between A's right hand and B's back	smaller, more touching, could indicate hugging depending on the interlocutors orientation
16	23	norm z coord of A's left knee	lower, may indicate kneeling
17	8	A's head angle, up/down	more straight vs more downwards
18	38	angle of A's right hand with x coord in A's system	hand more in front vs slightly towards left (see also Fig. 5(e))
19	37	norm x coord of A's left hand in A's system	x coord lower (further from body towards left, see also Fig. 5(a))
20	18	abs velocity of A's right foot	velocity higher
21	17	abs velocity of A's left foot	velocity higher
22	20	angle between A's hands	hands wider apart
23	40	distance between A's left hand and A's chest	bigger distance, hand further from chest
24	41	norm z coord of A's right hand in A's system	hand is higher, indicates raised hand (see also Fig. 5(a))
25	42	norm z coord of A's left hand in A's system	hand is higher, indicates raised hand (see also Fig. 5(a))

Table A.6: *Statistical analysis of the top 25 dominance features, according to the F_{value} criterion (each feature's rank according to the $mRMR_C$ criterion is included in the second column). The feature descriptions under the statistical tests column are describing high dominance behavior compared to low dominance behavior of a subject A. The statistical test performed is difference of means of the feature values between high and low dominance classes (t-test)*

Dominance: Comparison of high vs low dominance classes			
		F_{value} **(and** $mRMR_C$**)**	
rank F_{value}	**rank** $mRMR_C$	**feature**	**description of statistical tests results** $p \approx 0$
1	7	relative angle of A's body towards/away from B	body orientation more towards other vs sideways
2	1	A's head angle, up/down	more straight vs more downwards
3	6	norm z coord of A's center	higher, indicates less sitting
4	3	A's body leaning angle towards/away from B	more lean towards vs no leaning
5	13	distance of A's left hand from A's hip	greater distance, hand further away from hip
6	2	z coord of A's right foot	lower
7	17	norm x coord of A's right hand in A's system	x coord higher (further from body towards right, see also Fig. 5(a))
8	12	relative velocity of A towards/away from B	move more towards vs away
9	24	distance of A's right hand from A's hip	greater distance, further from hip
10	48	min dist between A's left hand and B's torso	smaller, indicates more touching
11	38	norm z coord of A's right hand in A's system	hand is lower
12	29	distance between A's hands	hands wider apart
13	10	A's body leaning angle, left/right	more lean right (though angle in both cases is close to zero)
14	28	norm x coord of A's left hand in A's system	x coord lower (further from body towards left, see also Fig. 5(a))
15	5	y coord of A's center	y abs value lower (y more towards center (0,0,0) of the recording space)
16	36	angle of A's right hand with x coord in A's system	hand more in front vs slightly towards left (see also Fig. 5(e))
17	14	relative velocity of A's right hand towards/away from B	move more towards vs away
18	45	norm z coord of A's left hand in A's system	hand is lower
19	5	A's body leaning angle, front/back	more lean front vs slighly less lean front
20	21	relative velocity of A's left hand towards/away from B	move more towards vs away
21	37	distance between A's right hand and A's chest	greater, hand further from chest
22	15	relative velocity of A's right foot towards/away from B	move more towards vs away
23	11	norm z coord of A's upper back	higher, more upwards position, also indicates less sitting
24	16	relative velocity of A's left foot towards/away from B	move more towards vs away
25	9	x coord of A's center	x abs value higher (x further from center (0,0,0) of the recording space)

Table A.7: *Statistical analysis of the top 25 valence features, according to the F_{value} criterion (each feature's rank according to the $mRMR_C$ criterion is included in the second column). The feature descriptions under the statistical tests column are describing positive valence behavior compared to negative valence behavior of a subject A. The statistical test performed is difference of means of the feature values between positive and negative valence classes (t-test)*

Valence: Comparison of positive vs negative valence classes			
		F_{value} **(and** $mRMR_C$**)**	
rank	**rank**	**feature**	**description of stat. tests results** $p \approx 0$
F_{value}	$mRMR_C$		
1	33	norm z coord of A's lower back	higher, indicates less sitting
2	42	abs velocity of A's right arm	higher velocity
3	15	A's head angle, up/down	slightly more downwards vs straight (though the two angles are almost the same)
4	34	distance between A's hands	hands closer together
5	41	distance of A's left hand from A's hip	greater distance, further from hip
6	36	distance of A's right hand from A's hip	greater distance, further from hip
7	28	norm x coord of A's left hand in A's system	x coord higher (closer to body towards right, see also Fig. 5(a))
8	20	norm z coord of A's upper back	lower, less upward position
9	11	relative angle of A's face towards B	face orientation more towards other
10	31	norm x coord of A's right hand in A's system	x coord lower (closer to body towards left, see also Fig. 5(a))
11	30	angle of A's left hand with x coord in A's system	left hand more towards front rather than left (see also Fig. 5(e))
12	27	norm y coord of A's right hand in A's system	y coord higher (further from body towards front)
13	21	abs velocity of A's right foot	higher velocity
14	23	abs velocity of A's left foot	higher velocity
15	43	distance between A's right hand and B's hand	lower, indicates more touching of B's hand
16	3	norm z coord of A's left knee	higher, indicates less kneeling
17	4	A's direction relative to B	slightly more towards right-front of B vs more in front
18	29	norm y coord of A's left hand in A's system	y coord higher (further from body towards front)
19	19	A's body leaning angle, front/back	less leaning front vs more leaning front, indicates less slouched posture
20	24	A's body leaning angle, towards/away from B	more leaning towards vs less leaning towards
21	12	relative velocity of A towards/away from B	more moving towards vs moving away
22	37	distance between A's right hand and A's chest	lower, indicates hand touching chest
23	38	distance between A's left hand and A's chest	lower, indicates hand touching chest
24	30	angle of A's left hand with x coord in A's system	hand more towards front vs towards right
25	1	y coord of A's center	y abs value bigger (y further from center (0,0,0) of the recording space)

recording perceived emotion in real time, in: ISCA Workshop on Speech and Emotion, 2000, pp. 19–24.

[37] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Transactions on Affective Computing 3 (2012) 42–55.

[38] A. Camurri, I. Lagerlof, G. Volpe, Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques, International Journal of Human-Computer Studies 59 (2003) 213–225.

[39] R. Cowie, M. Sawey, GTrace - General trace program from Queen's, Belfast, http://www.dfki.de/ schroed/feeltrace/, 2011.

[40] K. Audhkhasi, S. Narayanan, A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels., IEEE Trans. on Pattern Analysis and Machine Intelligence. (2012).

[41] H. Meng, A. Kleinsmith, N. Bianchi-Berthouze, Multi-score learning for affect recognition: the case of body postures, in: Proc. of ACII 2011.

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

[43] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. of Pattern Analysis and Machine Intelligence 27 (2005) 1226–1238.

[44] R. Duda, P. Hart, D. Stork, Pattern Classification, Springer-Verlag New York, Inc., 2007.

[45] D. Ververidis, C. Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Communication 48 (2006) 1162–1181.

[46] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book, Entropic Cambridge Research Laboratory, Cambridge, England, 2006.

[47] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9(8) (1997) 1735–1780.

[48] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18 (2005) 602–610.

[49] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling, in: Proc. of Interspeech, Japan, 2010.

[50] A. Graves, RNNLib toolbox, http://sourceforge.net/projects/rnnl/, 2010.

[51] P. Juslin, K. Scherer, Chapter 3: Vocal Expression of Affect, The new handbook of Methods in Nonverbal Behavior Research, Oxford Univ. Press, 2005, pp. 65–135.

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

[53] I. Kanluan, M. Grimm, K. Kroschel, Audio-visual emotion recognition using an emotion space concept, in: Proc. of EUSIPCO, 2008.

[54] N. Henley, Body politics revisited: What do we know today? In Gender, Power, and Communication in Human Relationships., Hillsdale, NJ: Lawrence Erlbaum, 1995.