Emotion recognition from embedded bodily expressions and speech during dyadic interactions

Philipp M. Müller*†, Sikandar Amin*‡, Prateek Verma†, Mykhaylo Andriluka*† and Andreas Bulling*

*Max Planck Institute for Informatics, Germany, {pmueller,andriluk,bulling}@mpi-inf.mpg.de
†Stanford University, USA, prateekv@stanford.edu
‡Technical University of Munich, Germany, sikandar.amin@in.tum.de

***Abstract*—Previous work on emotion recognition from bodily expressions focused on analysing such expressions in isolation, of individuals or in controlled settings, from a single camera view, or required intrusive motion tracking equipment. We study the problem of emotion recognition from bodily expressions and speech during dyadic (person-person) interactions in a real kitchen instrumented with ambient cameras and microphones. We specifically focus on bodily expressions that are embedded in regular interactions and background activities and recorded without human augmentation to increase naturalness of the expressions. We present a human-validated dataset that contains 224 high-resolution, multi-view video clips and audio recordings of emotionally charged interactions between eight couples of actors. The dataset is fully annotated with categorical labels for four basic emotions (anger, happiness, sadness, and surprise) and continuous labels for valence, activation, power, and anticipation provided by five annotators for each actor. We evaluate vision and audio-based emotion recognition using dense trajectories and a standard audio pipeline and provide insights into the importance of different body parts and audio features for emotion recognition.**

I. Introduction

Emotions are an integral part of human communication and manifest themselves in vocal prosody but also in body movements, facial expressions, and gestures. Particularly body movements induced by emotional responses, colloquially referred to as body language, play a key role in non-verbal human communication that is believed to represent a substantial part of all human communication [1]. In contrast to facial expressions, speech, as well as physiological parameters, such as heart rate or galvanic skin response, analysis of body language and recognition of emotions from bodily expressions is less well-explored in affective computing. This is mainly due to the significant challenge of recording and annotating natural bodily expressions of emotions in everyday environments. Consequently, previous works in affective computing mainly focused on datasets recorded in artificial laboratory settings. In these settings, individual actors were either positioned directly in front of the camera [2] or couples of actors were recorded using intrusive motion capture equipment to track their body movements [3] (see Figure 2 for examples).

In this paper we investigate multimodal emotion recognition from bodily expressions and speech recorded using unobtrusive ambient cameras and microphones in a real kitchen environment during naturalistic dyadic (person-person) interactions.

Fig. 1: Sample bodily expressions associated with different emotions from our dataset.

We propose an experimental setup and methodology that allows us to systematically record such bodily expressions embedded in regular interactions and background activities. To this end, we develop a set of scenarios that evolve around daily-life events and that lead to an emotionally charged conversation between two people. Each scenario is endowed with background information on the attitude of each person towards the event. We then film multiple pairs of actors role-playing and improvising each scenario in a fully functional apartment kitchen to closely resemble natural everyday living conditions (see Figure 1 for examples).

We took particular care to not script actors' performance, i.e. the only information we provided was a high-level background description of the scenario and emotional responses each of the actors was supposed to exhibit. In particular, we did not instruct actors how to role-play each scenario, or which bodily expressions or motions they should use to express a particular emotion. Instead, actors were free to interact with the environment and move around inside the kitchen. They were also not encumbered by wearing motion capture equipment, which made their bodily expressions more natural. The resulting MPIIEmo dataset including all annotations will be made publicly available upon publication.

The contributions of this work are threefold. First, we present

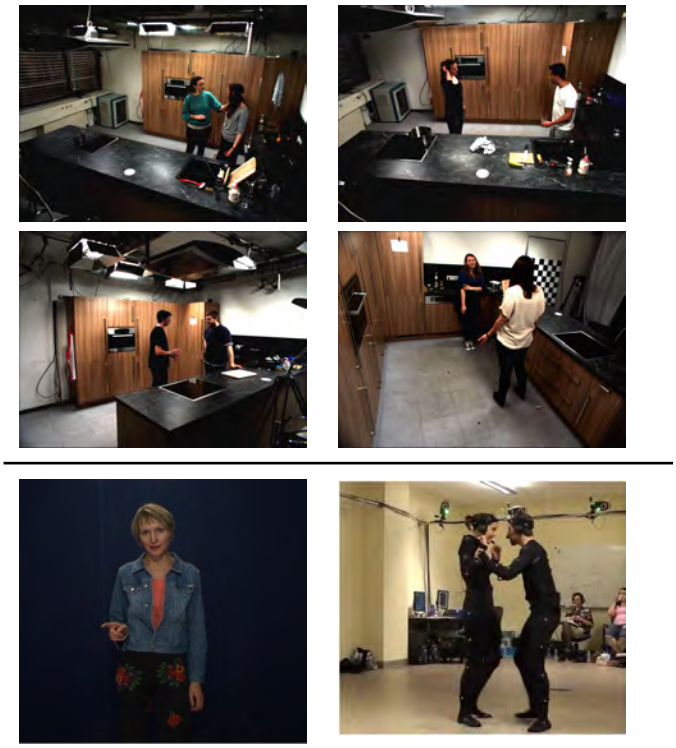
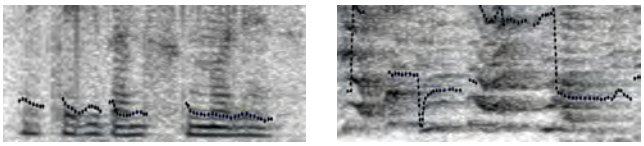


Fig. 2: Sample scenes of emotionally charged person-person interactions from our dataset (top). Samples from GEMEP [2] (bottom left) and CreativeIT [3] (bottom right) datasets.



GEMEP MPIIEmo (ours)

Fig. 3: Examples of audio spectrograms computed on GEMEP [2] and our MPIIEmo dataset. Blue curves correspond to pitch trajectories extracted with the approach described by Kasi and Zahorian [4]. Note that the spectrogram on GEMEP is cleaner which enables more robust pitch extraction.

an experimental set-up and methodology to unobtrusively collect video and audio data of actors engaged in person-person interactions in an everyday environment. The set-up and methodology were specifically developed to balance between the realism of the exhibited bodily expressions as well as the ability to study emotions that are difficult to record in real-world situations. Using this methodology, we further introduce the MPIIEmo dataset that contains 224 high-quality video and audio recordings of eight couples of actors engaged in emotionally-charged, natural interactions revolving around everyday scenarios. To the best of our knowledge, this is the first dataset of full-body videos of dyads of unaugmented people in affective interactions. It features a mix of emotional expressions embedded in a regular conversation and background activities, is fully-annotated with both categorical and continuous emotion labels, and provides multiple synchronised camera views. Third, we evaluate an approach for emotion recognition from video based on dense trajectories [5] and body part detections and provide insights into the relative importance of body parts for emotion recognition. We further study emotion classification from audio [6], highlighting difficulties of audio feature extraction on our dataset compared to the more constrained GEMEP dataset.

II. Related work

Our work is related to previous work that 1) explored the close link between emotions and body movements, 2) focused on recording bodily expressions of emotions, and 3) developed computational methods for human behaviour analysis.

A. Emotions and body movements

Humans are skilled in expressing emotions through non-verbal signals and in interpreting signals of others but relating body movements to specific emotional expressions is challenging given the subtleness of these movements. Efforts to clarify connections between emotions and body movements have a long history in behavioural science and suggested a strong connection exists between emotions and body movements [7, 8, 9]. Analysis of emotional body movements has since sparked a large body of work in social signal processing and affective computing, focusing on both encoding body movements in a similar fashion as facial expressions [10, 11] as well as inferring emotions from body movements [12, 13, 14] (see [15, 16, 17, 18] for reviews).

B. Recording bodily expressions of emotions

Several previous works investigated bodily expressions of emotions of individuals, either involving directed or at least carefully executed bodily expressions while facing the camera [2, 19, 20] or using body motion capture suits in controlled laboratory settings [21]. Previous works that studied bodily expressions during person-person interactions were either limited to only one person showing affective behaviour [22], to sitting people [23] or also used artificial settings and required sophisticated human augmentation [3, 24] (see Čereković [25] for a recent survey). Previous works also often only included short snippets of isolated bodily expressions that were neither embedded in natural background activities nor interactions with the other person or the environment [2, 21].

Our methodology is most similar to the one described by Metallinou et al. [3] but aims to increase realism of the recorded data while still retaining the ability to obtain laboratory-standard recording quality and accurate ground truth annotations. Specifically, our dataset contains bodily expressions of emotions during naturalistic person-person interactions, i.e. interactions that develop around everyday events and that are therefore embedded in casual body movements as well as interactions with the other person and the environment. As described in [2, 3] and following guidelines from [26, 27] we rely on recruited actors to improvise emotional expressions. Our dataset further contains two schemes for representation and annotation of emotional content, namely both categorical emotional labels [2] and continuous affect dimensions [28]. Sample scenes from two existing datasets as well as our own are shown in Figure 1.

C. Human behaviour analysis

Computational methods to analyse human behaviour either rely on on-body sensors, such as inertial measurement units, or ambient sensors, such as video cameras. On-body sensors are widely used in human activity and gesture recognition [29]. While current activity recognition systems achieve good performance for many activity recognition tasks, the majority of research focuses on recognising "which" activity is being performed at a specific point in time. More closely related to the problem investigated in this work, is qualitative activity recognition that studies means to extract qualitative information from inertial data, such as the quality or correctness of executing an activity. Such qualitative assessments are more challenging to perform automatically and have so far only been demonstrated for constrained settings, such as in sports. Specifically, previous works studied qualitative assessment of activities such as weight-lifting [30, 31, 32], rowing [33] or balance board exercises [34]. Recent computer vision works on human behaviour analysis mainly focused on basic recognition tasks, such as people detection [35], pose estimation [36, 37, 38], and recognition of fine grained details, such as appearance attributes [39], body and head orientation [40], gaze direction [41], detection of facial key-points [42], or social signals, such as holding hands or hugging [43]. In this work we investigate how recent advances in computer vision enable recognition of bodily expressions of emotions in video. In particular, we build on [5] that was previously used for activity recognition and [44] for body pose estimation.

III. The MPIIEmo Dataset

Collecting video and audio footage of bodily expressions of emotions in everyday settings is challenging. In addition to the scarcity of such situations in daily life, legal and ethical issues pose significant challenges for the collection of real-world behavioural data. Similar to Busso and Narayanan [26] and Scherer and Bänziger [27] we therefore opted to rely on acted performances and recorded couples of actors interacting with each other in a naturalistic environment (an apartment kitchen).

A. Data recording

We designed the data recording with two main objectives in mind: 1) to record video and audio footage of person-person interactions in a real kitchen setting and without on-body motion capture equipment that could affect the realism of these interactions, and 2) to record bodily expressions of emotions that are embedded in regular interactions and background activities commonly performed in the kitchen.

1) Recording setup: We recorded video and audio footage using eight ceiling-mounted, frame-synchronized machine vision cameras recording at 29.4 fps and four microphones, covering the whole interaction space inside the kitchen (see Figure 6). In total, we recorded eight pairs of actors (three female only, two male only, three mixed), with each pair performing seven scenarios, each consisting of four subscenarios. This resulted in 224 video clips with a total length of 143 minutes or 252,457

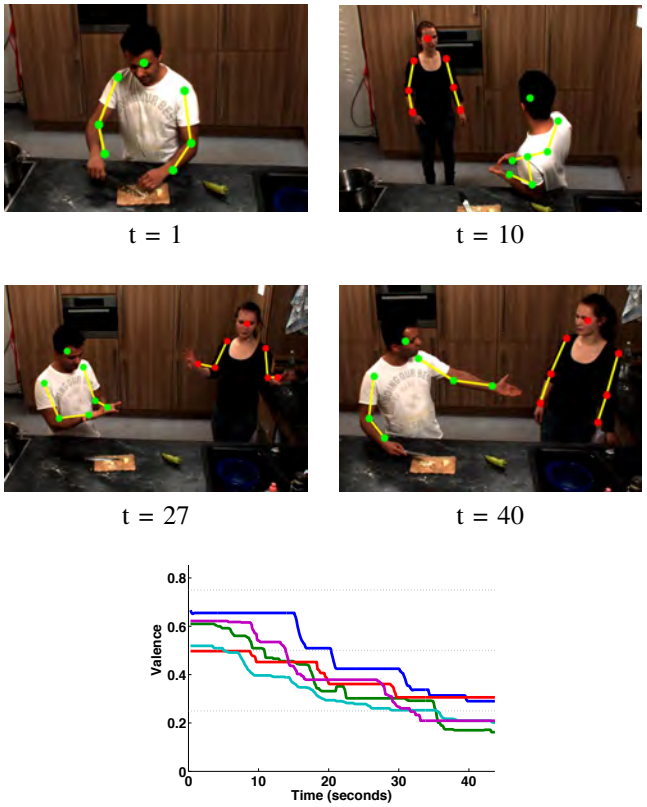


Fig. 5: Sample frames from a sequence with annotations for valence of the female subject with pose estimates. At about 20 seconds the couple gets into an argument about throwing away the garbage, as indicated by a more negative valence rating (cf. Figure 4).

Fig. 6: Kitchen environment used for recording our dataset. The kitchen was fully functional and instrumented with ceiling-mounted cameras (red circles) and microphones (blue circles).

frames. The average length of the recorded video clips in our dataset was around 38 seconds. The subscenarios were different variations of the overall scenario covering the display of different emotional responses. We designed each subscenario to correspond to a short conversations with the overall objective to record a diverse set of interactions that felt natural to the actors. According to these criteria, scenarios and subscenarios were selected from a pool of proposals by testing them in trial runs.

A sample scenario from our dataset is shown in Figure 4. In this scenario, one person reminds the other that it is his turn to empty the waste bin. The subscenarios then evolve around different reactions of the second person. He is either happy to be reminded, angry at the annoying reminder, angry at himself for forgetting about it again, or surprised because it's not his

He is **happy** with her.

He also applied for the same job. He is **angry**, because she talks happily about her new job.

He is **sad** because her new job means, she will soon move away. They had a close relationship. She also becomes **sad**.

He is **surprised** because she told before, how hard it is to get the job. She is proud (occurs later in the sequence).

Fig. 4: Sample scenario from our dataset. Each picture illustrates one subscenario. The high-level scenario description was: *She just received an offer for the job she always wanted. She enters the kitchen and tells the news* ***happily.***

turn. The first person then reacts accordingly. The full list of all scenarios and subscenarios will be released with the dataset.

2) Recording methodology: Actors were recruited from local student theatre groups and selected based on their acting abilities. All actors had at least one year of theatre training and were practising improvisation as a part of it. However, most actors were much more experienced, and half of them were, among others, part of a group dedicated solely to improvisational theatre. A director from a local theatre group worked with the actors during the recording. Actors were given short descriptions (about 1-3 sentences) of the scenarios and subscenarios, including explicit statements about the emotions and feelings involved in the interaction. Four of the six basic emotions by [45] were explicitly referred to: Happiness, Anger, Sadness and Surprise. Due to the freedom of improvisation, emotional expressions not covered by those four were shown as well as mixtures of several emotions. If necessary for the actors to become more familiar with a subscenario, additional background information was provided by the experimental assistant. In case the actors had problems to access the required emotions, the director used reenactment techniques as in [2]. Otherwise the actors were free to improvise. In particular, no instructions concerning concrete verbal or non-verbal expressions or gestures were given. Actors did not wear any specific clothes, could move freely inside the kitchen, and were free to interact with the kitchen itself as well as all objects, tools etc. in it. Each subscenario was repeated until the actors and the director were satisfied with the performance.

B. Groundtruth annotation

Two well-known models to describe emotions are the categorical and the dimensional emotion model. Categorical representations discretise the space of emotions and put labels like "Happiness" or "Surprise" to individual emotions. Recently, researchers in affective computing argued for the use of dimensional emotion models understanding affect as a real-valued vector [46]. Our dataset provides annotations for both emotion models. We used a subset of the well-known six basic emotions [45], namely anger, happiness, sadness, and surprise. For the dimensional model, we used the four dimensions valence, activation, power and anticipation as suggested by [47].

Scale	macc	map	F 1	F 2	F 3	F 4	F 5
Happiness	91.01	88.61	32.72	26.44	21.07	17.06	11.25
Anger	94.98	91.26	24.63	20.17	17.56	15.33	11.62
Sadness	94.62	79.60	18.46	12.25	9.14	6.71	5.41
Surprise	84.66	44.86	43.35	21.40	13.53	7.09	2.41

TABLE I: Performance evaluation of human annotators. "macc" and map" correspond to "mean accuracy" and "average precision". "F k" is the relative frequency of the emotion class when agreement of k annotators is required to mark a sample as positive.

In our setting *valence* intuitively describes whether the actor felt "good" or "bad", *activation* concerned his activeness vs. in-activeness, *power* referred to whether he felt in control of events, and *anticipation* varied between the actor being surprised and feeling he could foresee the future.

We extended the annotation tool GTrace [48] to support both emotion models. Annotations were performed by five psychology students (three female), instructed with descriptions of the different emotions as well as example video clips from a separate test recording. We then asked the annotators to label all video clips in randomized order, sequentially for both actors, and for each actor using first the dimensional and then the categorical emotion model. For the dimensional model we obtained continuous annotations across the whole sequence. For the categorical model, annotators were asked to select a subset of the six basic emotions for each actor in a clip. If an emotion was selected, its intensity was rated continuously in the same way as for the scales of the dimensional emotion model.

C. Analysis of annotations

We analysed the quality of both dimensional and categorical ground truth annotations:

a) Dimensional emotion model.: We quantified the agreement between annotators by computing the median of the Pearson correlation coefficients between all pairs of annotators. The correlations were computed across all frames. We obtained a high median correlation for valence (0.84), moderate median correlations for Power (0.67) and Activation (0.65), and a low correlation for Anticipation (0.42).

b) Categorical emotion model.: We aggregated annotations over several annotators to get discrete labels for all

emotion categories. Prior to aggregation we normalized the intensity ratings of each annotator by subtracting the mean and dividing by the variance across all videos. When computing the mean and variance for a particular emotion we assigned a zero intensity rating to all videos where this emotion was not labelled as present. For simplicity, we afterwards discretized the intensity ratings into binary emotion category labels separately for each annotator by thresholding the normalized intensity at 1. Finally, we defined the emotion label for a frame by requiring $k = 2$ annotators to agree that the emotion is present in this frame.

Table I (left) shows the mean accuracy and mean average precision for emotion recognition achieved by our annotators. The results are obtained by using three annotators to generate ground-truth labels and comparing it with the output of the remaining two annotators, repeating the process for all combinations of annotators. To calculate mean average precision, we used the ordering of data points induced by the annotators ratings. As can be seen from the table, while Happiness, Anger and Sadness are quite accurate, Surprise has a low mean average precision. The reason for this is that Surprise (like the related Anticipation scale) is often very strictly localized in time. In consequence, person-specific delays of the annotators introduce a lot of label uncertainty. Table I (right) shows the relative frequency of the positive class on the whole dataset for different values of k. We observe that agreement across annotators varies significantly across emotions. For example there is approximately two-fold decrease between $k = 2$ and $k = 5$ for Anger (20.17 vs. 11.62) whereas we observe nearly ten-fold decrease for Surprise (21.40 vs. 2.41). We can observe a similar pattern when quantifying agreement of annotators by computing the median of their correlations. We get high correlations for Anger (0.87), Happiness (0.82) and Sadness (0.83), but a low correlation for Surprise (0.48).

IV. Emotion classification from video and audio

To establish baseline performances for emotion classification on our MPIIEmo we evaluated approaches from computer vision and speech analysis. In the visual domain, we further examine the influence of different body parts and the interlocutor. In the audio domain, we compare several features that are commonly used for emotion classification from speech with respect to their performance on MPIIEmo as well as the well-established GEMEP dataset [2].

A. Video

We use dense trajectories, a recently introduced video descriptor which showed state-of-the-art performance for human activity recognition [5]. By using pose estimates, we can in- or exclude dense trajectories from different body parts or persons, which allows us to estimate their importance for emotion recognition in our framework.

1) Pose estimation: To estimate poses of people we train a set of body part detectors building on the convolutional neural network architecture of [44]. The detector in [44] is trained by minimizing a multi-task loss function that combines detection accuracy and accuracy in prediction of the object bounding boxes. When applying this approach to pose estimation we substitute the bounding box prediction component with a component that predicts locations of the neighboring body part. We train detectors for the head, shoulder and wrist. Wrist and shoulder detectors are trained to also predict the location of the elbow joint. Each shoulder and wrist detection thus generates a pair of body joints corresponding to either upper arm or lower arm segments respectively.

In the first step each part detector is densely evaluated in each camera view resulting in an initial set of candidate part hypothesis. In the second step we refine the body part detections using multi-view constraints. To that end we match the body segments across camera views and generate a set of 3D body segment candidates using triangulation. In the process we also discard segments with high reconstruction error, which allows us to filter out false positive detections. In the final step we assemble 3D full-body configurations from the available pool of body segments using constraints on the relative position of head, and upper and lower arms. This process results in high-quality 3D pose estimates for both subjects in the majority of the images. The remaining failures in pose estimation correspond to cases with particularly strong occlusions or rare poses such as subjects bending under the kitchen counter.

2) Identity annotation: The pose estimates are not associated with individual actors. To add personal identities, we annotate which actor is rightmost in one of the camera views and match this annotation to one of the two estimated body configurations.

3) Dense trajectories on body parts: We compute dense trajectories from a single view (the one in Figure 5), and associate them with 2D person and body part bounding boxes. A trajectory is associated with a bounding box if its starting point (x, y, t) is inside the bounding box at time t. We build separate codebooks for each of the five feature channels of the dense trajectory descriptor using k-means clustering with $N = 4000$ centroids on 100,000 trajectories randomly sampled from the training set. Depending on the experimental condition, we build separate codebooks for different body parts. For training and testing, we compute histograms over a time window of 2 seconds separately for each actor.

4) Classification: We apply SVM with a RBF-χ^2 kernel k as in [49]. The L feature channels are combined by normalizing their corresponding χ^2 distances separately using the means of the χ^2 distances of the feature channels on the training set:

$$k(x, y) = exp(-\frac{1}{L}\sum_{c=1}^{L}\frac{\chi^2(x_c, y_c)}{A_c}). \quad (1)$$

Here $\chi^2(x, y)$ denotes the χ^2 distance between x and y, x_c the c-th feature channel of example x and A_c the mean χ^2 distance for feature channel c on the training set.

B. Audio

We compute three features commonly used for emotion recognition from speech [6]: (1) non-zero pitch values, (2) spectral centroid and spectral flatness of the timbre and (3)

Method	Happiness	Anger	Surprise	Sadness	Average
full	48.0	28.4	24.6	16.8	29.5
full-hw	41.5	26.0	23.2	15.5	26.5
full-head	46.5	26.6	23.8	15.6	28.1
wrist	44.3	25.2	21.8	16.2	26.7
head	**50.7**	**32.9**	**26.2**	**18.2**	**32.0**
head-single	46.9	27.9	20.0	15.7	27.6
posrate	21.7	18.5	13.8	10.2	16.1

TABLE II: Mean average precision in percent for leave-one-recording-out cross-validation on our MPIIEmo dataset. "head", "wrist" and "full" denote using trajectories on the head, wrist or the full body, respectively. "full-head" denotes using all trajectories except head, and "full-hw" all trajectories except head and wrist. "head-single" denotes using trajectories from the target person only.

short time energy of the audio signal. The mean and standard deviation of these features are computed for frames of 30ms with 10ms hops, yielding an 8 dimensional feature vector for each frame. Classification is performed by using SVM with an RBF kernel and cross-validating the hyperparameters C and γ on the training set.

V. Experiments

A. Video

We report results on the detection of four types of emotional states. The detection is performed using a sliding window approach with a stepsize of 2 seconds. To compare detection results to human performance in Table I we generate labels from a fixed set of 3 annotators. Each window is then considered as positive for a given emotion category if at least half of the frames in that window are labelled positively by at least two annotators. A separate classifier is trained for each emotion category against other emotions and background. The regularization parameter C is selected by cross-validation in the training set. We report performance using the average precision metric as is common in human activity detection [5].

To quantify the contribution of different body parts, we compare different ways of selecting trajectories (see Table II). First, we investigate differences in performance due to exclusion of trajectories associated with certain body parts (conditions *full*, *full-head*, *full-hw* in Table II). We observe, that removing trajectories from the head lowers the performance for all emotions, whereas additionally removing trajectories from the wrists only results in a significant performance drop for the Happiness class. Secondly, we investigate the performances of classifiers based exclusively on trajectories associated with the head and the wrist. We find, that using trajectories from the head results in better performance and, more surprisingly it even outperforms *full* on all emotions. Finally, we pick the best performing condition to quantify the contribution of features extracted from the interlocutor. When removing those features (*head-single*), performance drops significantly.

When comparing these results with the performance of human annotators in Table I, we note that human performance is strongly superior for all emotion classes.

Dataset	Pitch	Timbre	Energy	All	MLK
GEMEP	53.9	58.9	48.7	64.1	26.0
MPIIEmo	36.5	41.1	41.0	43.2	35.0

TABLE III: Results for emotion classification using audio features. MLK denotes the probability of the most likely class in percent points.

B. Audio

We compare the performance of pitch-, timbre- and energy-based features on MPIIEmo. As a reference we also report results on the more controlled, single-actor GEMEP dataset. To align the experimental setups, we pick the 4 classes from GEMEP that are most similar to the 4 emotions on MPIIEmo (Anger, Joy, Sadness, Surprise), resulting in 39 examples. For MPIIEmo, we extract two second long training windows, with 10Hz sampling frequency, excluding all windows that had either multiple labels per actor, or no label at all (background). Note, that we construct examples without speaker separation, as first experiments using ICA indicated that this is a difficult task on MPIIEmo. As a result, the same features might appear in different classes if the two actors were given different labels in one window, making our task inherently more difficult than the one we defined on GEMEP. To compute the test error, we use leave-one-sequence-out cross-validation on GEMEP and leave-one-couple out cross-validation on MPIIEmo. The results (Table III) show, that combining all features achieves the best performance. Surprisingly, although pitch performs well on GEMEP and in prior research [50], it is near chance on MPIIEmo. Upon closer inspection, the bad performance on MPIIEmo can be explained by the difficulties for pitch extraction arising from the more realistic recording situation with microphones being at a distance from the speakers (*cf.* Figure 3).

VI. Conclusion

In this paper we proposed a new experimental setup and methodology to record bodily expressions of emotions embedded in everyday person-person conversations as well as background activities. Using this methodology, we presented the fully annotated MPIIEmo dataset that contains 224 high-resolution, multi-view video clips and audio recordings of emotionally charged interactions between eight couples of actors. We established baseline performances for emotion classification from both video and audio. We found that visual features computed from the head as well as the interlocutor were particularly important to achieve good performance, and that the more naturalistic recording setup on MPIIEmo poses challenges for audio feature extraction. To spark further research on this challenging emotion classification problem, the full dataset including all body pose estimates as well as categorical and continuous emotion annotations is publicly available.

Acknowledgments

The authors would like to thank Johannes Tröger for working as a director in the recordings as well as all involved actors.

REFERENCES

[1] K. Hogan, *Can't Get Through: Eight Barriers to Communication*. Pelican Publishing, 2003. 1

[2] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012. 1, 2, 4, 5

[3] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database: a multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality 18 May 2010*, p. 55, 2010. 1, 2

[4] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–361. 2

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013. 2, 3, 5, 6

[6] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015. 2, 5

[7] M. De Meijer, "The contribution of general features of body movement to the attribution of emotions," *Journal of Nonverbal behavior*, vol. 13, no. 4, pp. 247–268, 1989. 2

[8] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998. 2

[9] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51–B61, 2001. 2

[10] N. Dael, M. Mortillaro, and K. R. Scherer, "The body action and posture coding system (BAP): Development and reliability," *Journal of Nonverbal Behavior*, vol. 36, no. 2, pp. 97–121, 2012. 2

[11] E. Velloso, A. Bulling, and H. Gellersen, "AutoBAP: Automatic Coding of Body Action and Posture Units from Wearable Sensors," in *Proc. ACII*, 2013, pp. 135 – 140. 2

[12] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen, "Gesture-based affective computing on motion capture data," in *Proc. ACII*, 2005, pp. 1–7. 2

[13] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *Proc. ACII*, 2007, pp. 59–70. 2

[14] ——, "Detecting emotions from connected action sequences," in *Visual Informatics: Bridging Research and Practice*, 2009, pp. 1–11. 2

[15] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009. 2

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE TPAMI*, vol. 31, no. 1, pp. 39–58, 2009. 2

[17] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE TAC*, vol. 4, no. 1, pp. 15–33, 2013. 2

[18] M. Karg, A.-a. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Trans. Affect. Comp.*, no. 99, p. 1, 2013. 2

[19] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334 – 1345, 2007. 2

[20] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a minimal representation of affective gestures," *IEEE TAC*, vol. 2, no. 2, pp. 106–118, 2011. 2

[21] W. Wang, V. Enescu, and H. Sahli, "Towards real-time continuous emotion recognition from body movements," in *Human Behavior Understanding*, 2013, pp. 235–245. 2

[22] K. Bergmann, R. Böck, and P. Jaecks, "Emogest: investigating the impact of emotions on spontaneous co-speech gestures," *Multimodal Corpora: Combining applied and basic research targets*, 2014. 2

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008. 2

[24] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1766–1778, Oct 2014. 2

[25] A. Čereković, "An insight into multimodal databases for social signal processing: acquisition, efforts, and directions," *Artificial Intelligence Review*, vol. 42, no. 4, pp. 663–692, 2014. 2

[26] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *2nd International Workshop on Emotion, ICLRE'08*, 2008. 2, 3

[27] K. R. Scherer and T. Bänziger, "On the use of actor portrayals in research on emotional expression," in *Blueprint for affective computing: A sourcebook*, 2010. 2, 3

[28] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Proc. ICASSP*, 2011, pp. 2288–2291. 2

[29] A. Bulling, U. Blanke, and B. Schiele, "A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors," *ACM Computing Surveys*, vol. 46, no. 3, pp. 33:1–33:33, 2014. 3

[30] E. Velloso, A. Bulling, and H. Gellersen, "MotionMA: Motion Modelling and Analysis by Demonstration," in *Proc. CHI*, 2013, pp. 1309–1318. 3

[31] E. Velloso, A. Bulling, H. Gellersen, W. Ugulino, and H. Fuks, "Qualitative Activity Recognition of Weight Lifting Exercises," in *Proc. AH*, 2013, pp. 116–123. 3

[32] E. Velloso, A. Bulling, and H. Gellersen, "Towards Qualitative Assessment of Weight Lifting Exercises Using Body-Worn Sensors," in *Proc. UbiComp*, 2011, pp. 587–588. 3

[33] B. Tessendorf, F. Gravenhorst, B. Arnrich, and G. Tröster, "An imu-based sensor network to continuously monitor rowing technique on the water," in *Proc. ISSNIP*, 2011, pp. 253–258. 3

[34] A. Moller, L. Roalter, S. Diewald, J. Scherr, M. Kranz, N. Hammerla, P. Olivier, and T. Plotz, "Gymskill: A personal trainer for physical exercises," in *Proc. Percom*, 2012, pp. 213–220. 3

[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, 2010. 3

[36] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE TPAMI*, vol. 35, 2013. 3

[37] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. ICCV*, 2013. 3

[38] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. CVPR*, 2014. 3

[39] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proc. CVPR*, 2014. 3

[40] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. CVPR*, 2011. 3

[41] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-Based Gaze Estimation in the Wild," in *Proc. CVPR*, 2015. 3

[42] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. CVPR*, 2012. 3

[43] Y. Yang, S. Baker, A. Kannan, and D. Ramanan, "Recognizing proxemics in personal photos," in *Proc. CVPR*, 2012. 3

[44] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. ICLR*, 2014. 3, 5

[45] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969. 4

[46] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8. 4

[47] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007. 4

[48] R. Cowie and M. Sawey, "GTrace-General trace program from Queens, Belfast," 2011. 4

[49] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. ICCV*, 2013, pp. 3192–3199. 5

[50] N. A. Madzlan, J. G. Han, F. Bonin, and N. Campbell, "Automatic recognition of attitudes in video blogs - prosodic and visual feature analysis," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1826–1830. 6