A THIN-SLICE PERCEPTION OF EMOTION? AN INFORMATION THEORETIC-BASED FRAMEWORK TO IDENTIFY LOCALLY EMOTION-RICH BEHAVIOR SEGMENTS FOR GLOBAL AFFECT RECOGNITION

Wei-Cheng Lin and Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

Human's judgment has been shown to be *thin-sliced* in nature, i.e., accurate perception can often be achieved for a short duration of exposure to expressive behaviors. In this work, we develop a mutual information-based framework to select the most emotion-rich 20% of local multimodal behavior segments within a 3-minute long affective dyadic interaction in the USC CreativeIT database. We obtain a prediction accuracy of 0.597, 0.728, and 0.772 (measured by Spearman correlation) for an actor's global (session-level) emotion attributes (activation, dominance, and valence) using Fisher-vector encoding and support vector regression built on these 20% of multimodal emotion-rich behavior segments. Our framework achieves a better accuracy over using the interaction in its entirety and a variety of other data selection baseline methods by a significant margin. Furthermore, our analysis indicates that the highest prediction accuracy can be obtained using only 20% - 30% of data within each session, i.e., additional evidences for the thin-slice nature of affect perception.

Index Terms— thin-slice theory, behavioral signal processing, emotion recognition, multimodal signal processing

1. INTRODUCTION

Thin-slice theory of judgment [1, 2] states that accurate perception of another person's attributes, e.g., personality [3], intelligence [3], affect [4], and even negotiation outcome [5], can be obtained within a short duration of interactions; these personal attributes are often reflected in the multimodal behavior manifestation [6], e.g., speech, facial expressions, and body movements. Psychologists have analyzed this phenomenon in various interaction scenarios; for example Ambady *et al.* found that there exists an association between first impressions and sale's effectiveness [7], and Houser *et al.* found a similar conclusion in successful speed dating [8]. Recently, there has been an emerging research effort centered around developing algorithms to automate human's perceptual judgment. Some notable applications exists in systems for emotion detection [9, 10], social behaviors prediction [11, 12], and subjective attributes recognition [13, 14].

Thanks to Ministry of Science and Technology for funding (103-2218-E-007-012-MY3).

Behavioral signal processing (BSP) is one such research fields that aims at providing objective computational frameworks for domain experts in order to facilitate their decision-making process [15]. In BSP-related application domains such as mental health and education, it is common for domain experts make a global and holistic perceptual judgments after observing long-durational audio-video data. For example, in couple therapy, trained experts annotate couples' behaviors after watching 10-minute long interactions [16], and in educational training program, coaching principals grade the pre-service principals' speech after listening to their 3-minute long impromptu talk [17]. Knowing that there exists a *thin-slice* nature of human's perception, identifying which salient *slice* of behaviors that contributes to the final overall perception becomes critical in advancing the design of diagnostic instruments/training materials, the development of efficient engineering recognition algorithms, and even the understanding of human perceptual mechanisms [18].

Prior works in BSP have utilized computational frameworks, such as multiple instance learning [19] and sequential probability ration test [20], to locate salient behavior segments in order to perform automated session-level behavioral coding. In this work, we build upon this idea to recognize global (session)-level affective attributes (valence, activation, dominance) by identifying the behavior segments that are locally emotion-rich within a mutual information framework. In this work, we use the USC CreativeIT database [21], where each actor in a dyadic play is annotated with both local time-continuous (frame-level) and global session-level emotion attributes. Our recognition system for global emotion attributes leverages the availability of frame-level emotion annotation to identify the informative portions of behavior segments.

Our proposed framework utilizes only 20% of the most emotion-rich multimodal behavior segments in each session and obtains a prediction accuracy of 0.597, 0.728, and 0.772 (measured by Spearman correlation) for global activation, dominance, and valence, respectively. Comparing to using the data in its entirety, our framework improves the correlation by 0.234, 0.09, 0.244 absolute; comparing to using the segments associated with the most-frequent-seen local emotion attribute, our framework also achieves an improvement of 0.255, 0.174, and 0.235 (absolute). Our analysis also shows

that the highest recognition accuracy can be obtained by using approximately 20% - 30% of the entire session, which further reinforces the nature of thin-slice judgment of affect.

The rest of the paper is organized as follows: section 2 describes about research methodology, section 3 details the experimental setup and results, and section 4 concludes with discussion and future works.

2. RESEARCH METHODOLOGY

2.1. Database

We use the USC CreativeIT database for the present work [21]. This database includes dyadic improvisations based on an established theatrical acting technique called Active Analysis [22] in order to help elicit natural affective interactions. The behavior modalities included in the database are audio-recording from lapel microphones and full body motion capture on each actor (i.e., a recording of 45 markers' (x, y, z) coordinates using 12 Vicon cameras at 60 frames per second). There are a total of 8 pairs of actors (16 actors in total) with 50 total interaction sessions. Each actor in a session is annotated with both local time-continuous and global session-level emotion labels (a total of 100 samples of emotion annotations available). Emotion labels of interest are valence, dominance, and activation. Local time-continuous label for each attribute takes on a real value between $[-1, 1]$ sampled at every 10 ms frame. Global emotion labels are assigned at the session-level for each actor and take on an integer value between $[1, 5]$. There are at least 3 annotators per session.

In this work, there are a total of 100 full body motion capture data, 90 audio data, and 90 samples with both audio and motion capture. We compute the average of annotator scores for both local time-continuous and global session-level emotion labels to serve as our ground truths in our experiments.

2.2. Locally Emotion-Rich Behavior Segments

A complete workflow of our computational framework in identifying locally emotion-rich behavior (*thin-slice*) segments is illustrated in Figure 1. The procedure can be summarized in the following steps:

1. Multimodal behavior feature extraction and clustering representation using Gaussian mixture model (X)
2. Local time-continuous emotion annotation discretization per frame (Y)
3. Partition the session for each actor into 100 equally-spaced non-overlapping segments ($\approx$ 2 seconds each)
4. Compute mutual information, $I(X;Y)$, between X_i and Y_i, $\forall\, i \in [\,1, 100]$
5. Select top K segments to form the locally emotion-rich behavior segments (i.e., relevant *thin-sliced* behaviors)

2.2.1. Multimodal Behavior Feature Representation

The two modalities of behavior features that we extract correspond to vocal and body language information. 13 mel-frequency cepstral coefficients (MFCCs), including deltas and delta deltas, are extracted to quantify vocal information of each actor (a total of 39 vocal features).

We adopt the similar body language feature extraction method from a previous work done by Metallinou *et al.* on the same database [23]. Body language features are extracted in a geometric manner from the coordinates recorded on the 45 motion capture markers. We extract a total of 95 features per frame quantifying information such as individual body movement of the actor and interactive movement of the actor with his/her interlocutor. The choices of features are designed to capture behaviors such as looking at the interlocutor, approaching, touching, as well as body postures such as looking down and hand gestures. Out of the 95 features, 70 features are identical to the previous work (details are in [23]). We additionally compute 25 more features: 14 acceleration-based features (we compute acceleration for all velocity-based features), 5 distance-based features (left/right hand to head, left/right hand to torso, and left leg to right leg), 3 features of head's coordinates (x, y, z), and 3 angle-based features (angle between left and right leg, angle between left/right leg and global origin).

Lastly, we perform GMM ($m = 128$) on each behavior modality separately to quantize individual behavior streams into m clusters at each frame, i.e., denoted as X, where X can take on value between 1 to m.

2.2.2. Selection of Top K Emotion-Rich Segments

We first discretize the original local time-continuous emotion attributes (E_i), where i indicates {valence, dominance, and activation}, into 5 quantized levels (Y):

- Level 0: $-1.0 <= E_i < -0.6$
- Level 1: $-0.6 <= E_i < -0.2$
- Level 2: $-0.2 <= E_i < \ \ 0.2$
- Level 3: $\ \ 0.2 <= E_i < \ \ 0.6$
- Level 4: $\ \ 0.6 <= E_i <= 1.0$

Then, we split each actor's behavior data into 100 equally-spaced segments. We can, hence, compute the mutual information easily between X_i and Y_i, where i indicates the segment index, to quantify the amount of information jointly present in the behavior expression and local emotion rating.

$$I(X_i, Y_i) = \sum_{Y} \sum_{X} p(X_i, Y_i) \log \frac{p(X_i = x, Y_i = y)}{p(X_i = x)p(Y_i = y)}$$

Finally, we rank $I(X_i, Y_i)$ for each actor in that session and retrieve the top K number of i's to be the locally emotion-rich behavior segments.

2.3. Global Affect Recognition

Section 2.2 describes our approach in identifying sub-portions of data to be used to train global emotion recognition system. The process can be thought as a *data reduction/selection* process within each session. After retrieving K number of segments, we concatenate the original behavior features, i.e., 39 and 95-dimensional feature vector, from all of the chosen segments to represent that particular actor's behaviors.

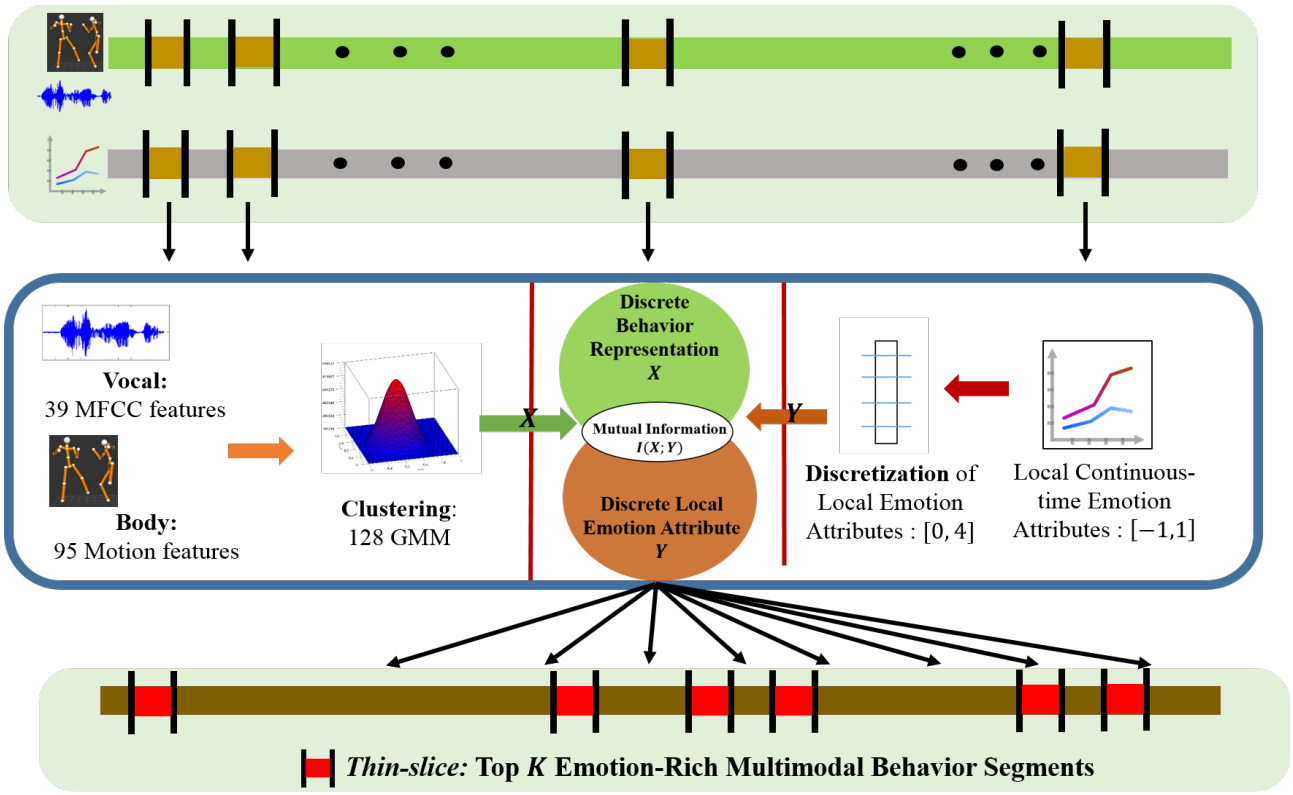

Fig. 1: A complete workflow of our mutual information-based framework for identifying locally emotion-rich behavior segments as our *thin slice* representation

Since each interaction has different lengths, the total number of frames selected varies across different sessions. In this work, we employ Fisher-vector encoding method to map the varying-length sequence of features into fix-length vector of features [24, 25]. Fisher-vector encoding is operated by first trains an overall GMM and further calculates the gradient vector using FIM (Fisher Information Matrix) approximation to describe the direction changed needed for the trained GMM parameters to obtain a better fit on the data of interest, i.e., individual actor's sequence of feature vector per session.

We use support vector regression (SVR) recognize the global emotion attributes. The final fusion between audio and body language information is done at the decision-level. The final predicted emotion value, E_{predAB}, is the following,

$$E_{predAB} = a \times E_{predA} + (1 - a) \times E_{predB}$$

where E_{predA} refers to prediction using audio features, and E_{predB} refers to prediction using body language features.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental setup

We evaluate our global emotion recognition accuracy using leave-dyad-out cross validation and measure the accuracy using Spearman correlation. We select $K = 20$ segments of 1% data partition (approximately 35 - 40 seconds total worth of data) for each actor in the session as our locally emotion-rich behavior segments to perform the prediction task. The number of mixtures used for Fisher-vector encoding is chosen empirically for each emotion attribute separately. We compare against eight different baseline models (listed below).

I. **Entire Session**: Use 100% of data from each session
II. **Random Sample 20%**: Randomly sample 20% of each session
III. **Five-ordered Segments**: Split each session into five parts in sequence, and average the prediction results from these five parts
IV. **Random Five-ordered Segments**: Same as III, but randomly select only one from the five parts
V. **Random 20×1% Sample**: Randomly sample 20 segments of the 1% data partition
VI. **Mean of Local Labels**: Select the data portion corresponds to the mean values (rounded to the nearest integer) of quantized local time-continuous emotion labels (≈ 60% of data in each session)
VII. **Mode of Local Labels**: Same as VI, but choose data corresponds to the mode (≈ 60% of each session)
VIII. **Median of Local Labels**: Same as VI, but choose data corresponds to the median (≈ 60% of each session)

Method I is the conventional method to perform prediction, method II - V are based solely on various random approaches of data reduction, and method VI - VIII leverage the human annotated local time-continuous labels for data selection, sim-

Table 1: *Summary of global emotion prediction attributes results for 9 different methods (measured by Spearman correlation) for audio, body language, and fusion model (all p-values are less than* 10^{-5}*)*

Audio									
	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	Proposed
Activation	0.251	0.240	0.323	0.247	0.207	0.302	0.261	0.294	**0.382**
Dominance	0.260	0.202	0.224	0.264	0.174	0.210	0.207	0.238	**0.545**
Valence	0.281	0.208	0.119	0.041	0.204	0.259	0.235	0.258	**0.521**
Body Language									
	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	Proposed
Activation	0.380	0.373	0.244	0.187	0.179	0.411	0.363	0.395	**0.604**
Dominance	0.656	0.643	0.366	0.472	0.581	0.608	0.613	0.629	**0.685**
Valence	0.547	0.471	0.351	0.110	0.371	0.453	0.525	0.449	**0.755**
Fusion Model									
	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	Proposed
Activation	0.363	0.398	0.359	0.216	0.230	0.398	0.342	0.381	**0.597**
Dominance	0.638	0.633	0.416	0.490	0.586	0.549	0.554	0.603	**0.728**
Valence	0.528	0.430	0.359	0.223	0.278	0.432	0.537	0.459	**0.772**

ilar to our proposed method.

3.2. Results and Discussions

Table 1 summarizes our global emotion attributes prediction results using audio-only, body language-only, and fusion model. There are several major points to note. The first is that the fusion model achieves the best accuracy, i.e., 0.728, and 0.772 for, dominance, and valence, respectively, signifying the importance of multimodal behavior modeling. Furthermore, our proposed mutual information-based method to select locally emotion-rich behavior segments obtain the best accuracy when comparing all other eight different baseline models. In specific, by using only 20% of emotional-rich behavior segments, we obtain an 0.234, 0.09, and 0.244 absolute improvement in the correlation comparing to using behavior data from the session in its entirety.

Another interesting point to note is that method VI - VIII also leverages local time-continuous human annotation to perform data selection. The mean, median, and mode values of local time-continuous (frame-level) emotion label by itself correlate quite well with the global-level emotion attributes, e.g., mode value of local time-continuous emotion attribute correlates 0.611, 0.804, and 0.846 with global emotion attributes for activation, valence, and dominance, respectively. However, by selecting their corresponding behavior segments to perform the global recognition task, unlike our proposed framework, there is no significant improvement over just using the session in its entirety. This result underscores an important feature of our mutual information framework, which considers the joint informational content between local emotional judgment and expressive behaviors.

Lastly, we plot the different percentages of data selected in the fusion model of our framework and their prediction accuracies in the three global emotional attributes in Figure 2. It is interesting to note that it only requires about 30%, 30%, and 20% (valence, dominance, and activation) of the each session to obtain the best prediction accuracies. This result further strengthens the *thin-slice* theory of affect judgment.

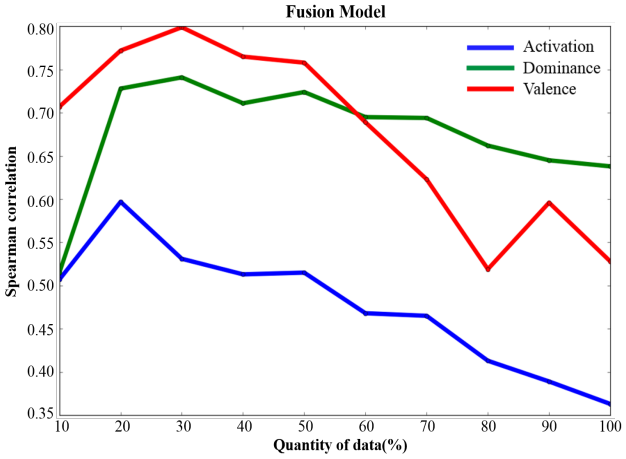

Fig. 2: A plot between different percentages of selected data in each session and global emotion attributes (valence, activation, and dominance) prediction accuracies

4. CONCLUSIONS AND FUTURE WORK

In this paper, we present a mutual information-based framework to identify locally emotion-rich multimodal behavior segments to recognize global (session-level) emotion attributes in the USC CreativeIT database. By using just 20% of each session, we obtain an significant improvement over using the entire session. Our results further demonstrate that only approximately 20 - 30% is needed of behavior data in each session to achieve the best overall prediction - reinforcing the *thin-sliced* nature of human perceptual judgment.

One of our immediate future works is to derive and incorporate the local time-continuous emotion recognition jointly within this framework to complete an end-to-end system for automatic identification of emotion-rich behavior segments for global emotion attributes' prediction. Furthermore, we will also investigate these *thin slices* of behaviors to understand their underlying reasons for triggering human perception of affect when observing long-durational interactions.

5. REFERENCES

[1] Nalini Ambady and Robert Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.," *Psychological bulletin*, vol. 111, no. 2, pp. 256, 1992.

[2] Nalini Ambady and Robert Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness.," *Journal of personality and social psychology*, vol. 64, no. 3, pp. 431, 1993.

[3] Peter Borkenau, Nadine Mauer, Rainer Riemann, Frank M Spinath, and Alois Angleitner, "Thin slices of behavior as cues of personality and intelligence.," *Journal of personality and social psychology*, vol. 86, no. 4, pp. 599, 2004.

[4] Nalini Ambady and Heather M Gray, "On being sad and mistaken: mood effects on the accuracy of thin-slice judgments.," *Journal of personality and social psychology*, vol. 83, no. 4, pp. 947, 2002.

[5] Jared R Curhan and Alex Pentland, "Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes.," *Journal of Applied Psychology*, vol. 92, no. 3, pp. 802, 2007.

[6] Alex Pentland, "Social dynamics: Signals and behavior," in *International Conference on Developmental Learning*, 2004, vol. 5.

[7] Nalini Ambady, Mary Anne Krabbenhoft, and Daniel Hogan, "The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness," *Journal of Consumer Psychology*, vol. 16, no. 1, pp. 4–13, 2006.

[8] Marian L Houser, Sean M Horan, and Lisa A Furler, "Predicting relational outcomes: An investigation of thin slice judgments in speed dating," *Human Communication*, vol. 10, no. 2, pp. 69–81, 2007.

[9] Chul Min Lee and Shrikanth S Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[10] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[11] Maja Pantic and Leon JM Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[12] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang, "Automatic analysis of multimodal group actions in meetings," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 305–317, 2005.

[13] Gelareh Mohammadi and Alessandro Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *Affective Computing, IEEE Transactions on*, vol. 3, no. 3, pp. 273–284, 2012.

[14] BjöRn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan, "Paralinguistics in speech and languagestate-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[15] Shrikanth Narayanan and Panayiotis G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[16] Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[17] Shan-Wen Hsiao, Hung-Ching Sun, Ming-Chuan Hsieh, Ming-Hsueh Tsai, Hsin-Chih Lin, and Chi-Chun Lee, "A multimodal approach for automatic assessment of school prinicipals' oral presentation during pre-service training program," in *Interspeech*, 2015, p. in press.

[18] Dana R Carney, C Randall Colvin, and Judith A Hall, "A thin slice perspective on the accuracy of first impressions," *Journal of Research in Personality*, vol. 41, no. 5, pp. 1054–1072, 2007.

[19] Athanasios Katsamanis, James Gibson, Matthew P Black, and Shrikanth S Narayanan, "Multiple instance learning for classification of human behavior observations," in *Affective Computing and Intelligent Interaction*, pp. 145–154. Springer, 2011.

[20] Chi-Chun Lee, Athanasios Katsamanis, Panayiotis G Georgiou, and Shrikanth Narayanan, "Based on isolated saliency or causal integration? toward a better understanding of human annotation process using multiple instance learning and sequential probability ratio test.," in *INTERSPEECH*, 2012.

[21] Angeliki Metallinou, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan, "The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language Resources and Evaluation*, pp. 1–25, 2015.

[22] Sharon Marie Carnicke, "The knebel technique: active analysis in practice," *Actor Training*, pp. 99–116, 2010.

[23] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[24] Florent Perronnin and Christopher Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[25] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010*, pp. 143–156. Springer, 2010.