INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Deriving *Dyad-Level* Interaction Representation using Interlocutors Structural and Expressive Multimodal Behavior Features

Yun-Shao Lin and Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan
cclee@ee.nthu.edu.tw

Abstract

The overall interaction atmosphere is often a result of complex interplay between individual interlocutor's behavior expressions and joint manifestation of dyadic interaction dynamics. There is very limited work, if any, that has computationally analyzed a human interaction at the dyad-level. Hence, in this work, we propose to compute an extensive novel set of features representing multi-faceted aspects of a dyadic interaction. These features are grouped into two broad categories: *expressive* and *structural* behavior dynamics, where each captures information about within-speaker behavior manifestation, inter-speaker behavior dynamics, durational and transitional statistics providing holistic behavior quantifications at the dyad-level. We carry out an experiment of recognizing targeted *affective atmosphere* using the proposed *expressive* and *structural* behavior dynamics features derived from audio and video modalities. Our experiment shows that the inclusion of both *expressive* and *structural* behavior dynamics is essential in achieving promising recognition accuracies across six different classes (72.5%), where *structural*-based features improve the recognition rates on classes of *sad* and *surprise*. Further analyses reveal important aspects of multimodal behavior dynamics within dyadic interactions that are related to the affective atmospheric scene.

Index Terms: affect recognition, face-to-face interaction, multimodal behaviors, dyad-level affect

1. Introduction

Human face-to-face interactions, by nature, involve multiple ($\geq$ 2) interacting participants engaged in conversations to exchange information, communicate ideas, express emotional feelings, etc. The individual behavior expressions interweaving with the inter-participants dynamics often shape the overall perceived *tone* and/or the *style* of a given interaction. For example, Sy et al. demonstrated that the behaviors exhibited by the group leader have a strong impact on the perception of the group atmosphere and affect the group performance [1]; Fay et al. showed that the difference in the exchanges of speech patterns around the topic-of-interest within a group discussion is essential in categorizing the interaction as a dialog or a monologue [2].

In terms of group emotion dynamics studies, Barsade et al. proposed a conceptual framework of emotion influence as a "top-down" and a "bottom-up" flow within a group [3]. This framework states that the individual-level emotion and the group-level emotion affect each other; the "bottom-up" aspect focuses on how individual-level emotion shape the group-level emotion, where "top-down" aspect shows vice versa. This framework was elaborated further by Kelly et al. [4]. In this work, our goal is to operationalize a computational framework in representing both interlocutors behaviors and their dynamics in a dyadic setting, and we apply it for automated analysis of the targeted *affective atmosphere* of dyadic interactions.

While there has been a tremendous technical progress in the emotion recognition technology, most of the works have focused mainly on obtaining robust emotion recognition at an *individual*-level by modeling each person's behaviors (e.g., facial expressions [5], speech [6], and multimodal behaviors [7, 8, 9]) or by leveraging dyad's information [10, 11, 12]. Only recently in CVPR2016, Mou et. al. have proposed the use of facial expressions of participants to automatically detect emotion at the group level [13]. There is, however, still very limited research on developing computational approaches in quantifying multimodal expressions and dialogue structures in the automatic recognition of *affective atmosphere* at the group (dyadic) level.

In this work, we propose to derive an extensive set of features from both interlocutors to describe multi-faceted characteristics of an interaction at the dyad-level. The approach is done by first assigning each audio and video frame into one of the three basic *states*. For speech, the three states are defined to characterize the floor holding situation: 1) silence ($s_{1\text{sp}}$), 2) secondary speaker ($s_{2\text{sp}}$), and 3) primary speaker ($s_{3\text{sp}}$) .For video, the three states are defined to characterize the visual attention on movement: 1) stationary ($s_{1\text{bm}}$), 2) one-person movement ($s_{2\text{bm}}$), and 3) simultaneous movement ($s_{3\text{bm}}$). We then compute two attributes, *expressive* and *structural* dynamics, with respect to each state for each modality to compose our final feature set. *Expressive* dynamics are the features representing vocal characteristics and body movements using low-level descriptors on segments of states with activities ($s_{2\text{sp}}$, $s_{3\text{sp}}$, $s_{2\text{bm}}$, $s_{3\text{bm}}$) - quantifying interactive behavior at the micro-scale. *Structural* dynamics involve computing features on durational and transitional statistics of the state's evolution of an interaction - measuring behavior interaction at the macro-scale.

This framework of deriving behavior features can capture individual speaker's behavior manifestation, inter-speaker behavior dynamics, and interaction flow within a dyadic interaction; the formulation further provides a natural interpretability. We apply the framework in task of recognizing the dyad-level *affective atmosphere* in the NTUA database, i.e., a multimodal dyadic emotional database. There are a total of six classes of affective atmosphere, and we obtain the best accuracies of 72.5% with a combination of features from *expressive* and *structural* dynamics using both audio and video modalities. Our analysis shows that without using *structural*-based features, the recognition rate obtained is only 65.6%. These macro-level features provide improvement on classes such as *sad* and *surprise*. Further analysis reveal that the secondary speaker's speech cues provide more information on the affective atmosphere compared to the primary speaker; also, the expression in views of "one-speaker movement" alone can achieve an accuracy of 62.6%. The rest of paper is as follows: section 2 describes

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

http://dx.doi.org/10.21437/Interspeech.2017-569

database and framework, section 3 details experimental setup and results, and section 4 is the conclusion.

2. Research Methodology

2.1. The NTUA Emotion Database

The NTUA Emotion database is a newly-collected Chinese corpus. It is a collaborative work with 44 participants (24 females, 20 males) from the department of drama of the National Taiwan University of Arts, Taiwan. A pair of actors formed a dyad team to perform a 3-minute long face-to-face interactions. The interactions include improvised real-life scenarios without predefined scripts to ensure natural behavior manifestations. In total, we have recorded about 11-hour data, which includes 210 sessions, of both audio and video data. In each session, both actors' voice are captured by using Bluetooth wireless closed-up microphones. The movement of actors is capture by SONY high definition video camera. All of the sessions are directed by two professional directors (also hold appointments as professors at the department). The director plays a significant role in steering the overall targeted interaction *atmosphere* during the collection. The database is designed to target six major categories of affective atmospheres at the dyad-level: *angry*, *frustrated*, *happy*, *neutral*, *sad*, and *surprise*.

In this work, we focus on predicting these dyad-level affective attributes, i.e., the targeted affective atmosphere as our prediction label. The reliability of these targeted dyad-level affective attributes are further assessed by having 42 annotators to rate each interaction to be one of the six categories. Table 1 shows percentages of the perceived ratings (determined through majority vote of the 42 annotators) that result in the same attributes as the targeted affective atmospheres; all six emotion categories achieve over 90%. The total number of samples are 210 with the split between the six classes shown in Table 1.

2.2. Multimodal *Structural* and *Expressive* Features

Figure 2 shows a schematic of our complete multimodal *structural* and *expressive* features used in the detection of dyad-level affective attributes. The framework involves two steps: 1) preprocessing to assign each audio and video frame as one of the three distinct states, and 2) computing *structural* and *expressive* features intended to capture aspects on individual speaker's behavior manifestation, inter-speaker behavior dynamics, interaction states' durational and transitional statistics.

2.2.1. Pre-processing of Audio and Video

Within each dyadic interaction, we can conceptualize the dynamics of behavior flow as a *state-changing* evolution. As a first step, we categorize each session into three pre-defined parts with respect to audio and video separately. For the audio, we separate frames of audio into three distinct states: 1) silence ($s_{1\text{sp}}$), 2) secondary speaker speaking ($s_{2\text{sp}}$), and 3) primary speaker speaking ($s_{3\text{sp}}$), where the primary speaker is defined as the interlocutor that takes up the largest portion of speaking floor in a given interaction. This description of audio states (10ms frame-rate) can be thought of as a basic primitive roughly in characterizing the floor-taking within a dyadic interactions.

Table 1: *It shows the percentages of the perceived ratings (determined through majority vote of the 42 annotators) that result in the same attributes as the target affective atmospheres*

Dyad-level Affect	Targeted	Perceived	Percentage (%)
angry	32	32	100%
frustration	36	38	94%
happy	30	32	93%
neutral	40	36	90%
sad	40	40	100%
surprise	32	32	100%

The analogous splitting in the video data, i.e., body movement of the dyad, is conceptually similar though the floor-taking phenomena happened in the speech modality do not occur in the video modality. Here, we categorize the three types (states) of movements in the viewing screen: 1) stationary ($s_{1\text{bm}}$), 2) one-person movement ($s_{2\text{bm}}$), and 3) simultaneous movement ($s_{3\text{bm}}$). An example of each of these three states is shown in Figure 1. We derive this categorization to provide the basic unit (30 Hz frame-rate) in characterizing the visual attention of the viewing screen, e.g., no one is moving on the screen, one-person moving in a dyad scene, and two-people moving together. However, unlike the splitting of audio states can be easily done with the speaking duration of each speaker, we have to develop the following automatic video pre-processing procedure:

1. **Extracting dense trajectories:**
   We utilize dense trajectory extraction method [14] to compute positions of interests, that is, the (x, y) positions of candidate movements in the video
2. **Cropping and denoising:**
   We consider movement only within the marked space between the two lines shown in Figure 1 and removing spurious trajectories using morphology-based methods [15]
3. **Categorizing states**
   In order to assign a frame into one of the two movement states ($s_{2\text{bm}}$, $s_{3\text{bm}}$), we perform k-means ($k = 2$) clustering on the positions of leftover trajectories to obtain the two centroids within each frame. If the distance between the two centroid's x-coordinate are large, we mark it as $s_{3\text{bm}}$, otherwise as $s_{2\text{bm}}$. This corresponds to the fact that the actors mainly move horizontally on stage. If there is no trajectories left, the frame is categorized as $s_{1\text{bm}}$.
4. **Generating bounding boxes**
   After states categorization, we further put a bounding box around all moving trajectories to enable further feature computation. $S_{2\text{bm}}$ will output one bounding box, and $s_{3\text{bm}}$ will output two bounding boxes

2.2.2. Structural Dynamics

Structural dynamics features are used to represent the macro-evolution of basic audio and video states progression throughout an interaction. In our work, we define two broad types of structural features: **statistics** and **transitions**. In terms of statistics, we compute the normalized sum, average, and standard deviation. *Sum* is computed as the total number of frames occupied by a particular state divided by the total number of frames per session. *Average* is computed by averaging the time duration over homogeneous segments of a particular state within a session. *Standard deviation* is a computation of standard deviation on the time duration over these homogeneous segments. These *statistics*-based computation result in a total of nine features for each modality. In terms of transitions, there are six different states transition possible for each modality. We extract the transition features by computing the state transition probability within a session. In total, *transition*-based features result in a total of six dimensions for each modality.

In summary, *statistics*-based features attempt to capture the overall characteristics of the state occupation both in absolute terms and in terms of consecutive time-duration (e.g., how long

Figure 1: *An example of three different states in video modality. Yellow circles are the dense trajectories extracted. The two blue line roughly limits possible movements on stage. Red rectangles and pink rectangles appear at the position where actors have the movement*

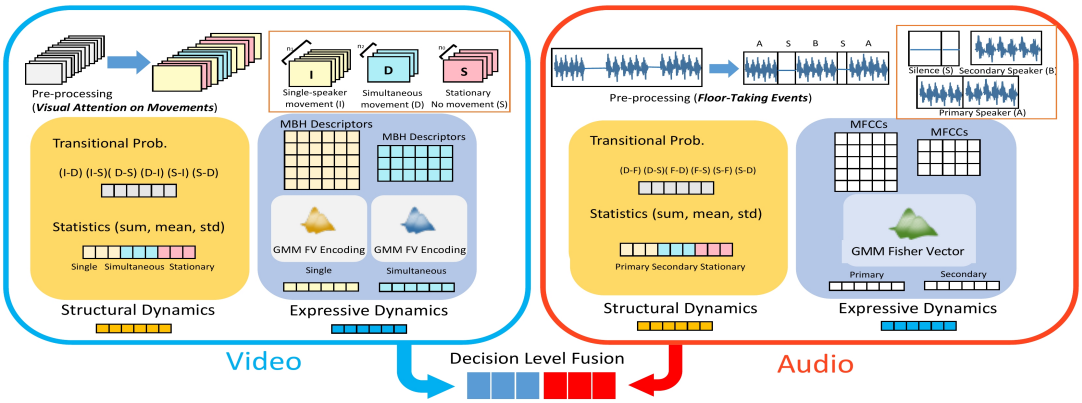

Figure 2: *A schematic of our complete multimodal structural and expressive features. The framework involves two steps: 1) pre-processing to assign each audio and video frame as one of the three distinct states, and 2) computing structural and expressive features to capture aspects on individual speaker's behavioral manifestation, inter-speaker behavioral dynamics, durational and transitional statistics. The final recognition is done via decision-level fusion between the two modalities.*

do the overall *stationary* scenes occupy, how long does the primary speaker speak overall, etc), and *transition*-based features capture the changing dynamics of these states for each modality over a session (e.g., how rapid does turn taking occur between the two speakers, how rapid is the shift in visual attention on movements, etc).

2.2.3. *Expressive Dynamics*

We compute expressive dynamics features to represent the behavior characteristics of each state at the interaction-level for audio and video modality. The expressive features can only be computed for states with activities, i.e., $s_{2\text{sp}}$, $s_{3\text{sp}}$, $s_{2\text{bm}}$, $s_{3\text{bm}}$. In the video modality, expressive features are computed from employing Gaussian Mixutre Model (GMM) Fisher-vector encoding (FV) [16] on motion boundary histogram (MBH); MBH are video descriptors for action characterization that is derived from the tracked trajectories [14]. In the audio modality, we extract a high-dimensional vector to represent the acoustic profile of each actor in the session. We first extract 39 low-level acoustic descriptors, i.e., 13 MFCCs and their delta and delta-deltas. We then perform FV encoding on MFCCs after performing speaker-wise z-normalization to obtain interaction level representation. We compute these representations for each of the two *states* with activities for each behavior modality.

3. Experimental Setup and Results

We conduct dyad-level *affective atmosphere* recognition tasks in the NTUA database using our proposed multimodal *expressive* and *structural* behavior features (section 2.2). The choice of classifier is support vector machine with linear kernel ($C = 1$). The evaluation scheme is done via leave-one-interaction-out cross validation. The performance metric used is unweighted average recall (UAR). The fusions between speech- and video-based *expressive* and *structural* features are carried out in a two-stage process, where the decision scores of each type of features are concatenated and then fed into the second-level classifier. The training of the second-level classifier is done completely using the training set only.

3.1. Experimental Results and Analyses

3.1.1. *Expressive-only and Structural-only Dynamics*

Table 2 summarizes the recognition results using expressive-only dynamics features and structural-only dynamics features in speech-only, video-only, and multimodality fusion.

Performing dyad-level attribute recognition based on *expressive* features can be thought of as the most intuitive go-to method, i.e., extracting behavior representations of each speaker within the interaction as feature inputs to the classifier. In fact, the method proposed in [13] is largely based on this concept. The best recognition rate using multimodal expressive features is 65.6%, i.e., achieved by fusing secondary speaker's acoustic information with the movement characteristics exhibited in the duration of time when only one of the dyad is moving. This fusion of multimodal expressive behaviors improves 14.9% and 3% absolute over the best accuracies achieved in speech (50.7%) and video modality (62.6%). Video-based expressive features outperform audio-based expressive features. The body movement features are better overall with recognition rate higher in attribute of *frustration*, where audio features are better for different attributes, i.e., *happy* and *neutral*.

There are two interesting observations that we see from Table 2. The first one is that the information possessed by the secondary speaker is more than the primary speaker in terms of recognizing the dyad-level affective attributes (50.7% vs. 43.2%). This provides an intriguing result indicating that it may not be the person who talks the most during a dyadic interaction that would shape the affective interaction as a whole the most. Secondly, the state of "one-speaker movement" gives the higher recognition rate in the video modality compared to "simultaneous movements" (62.6% vs. 31.7%). The time portion of the

Table 2: *A summary of affective atmosphere recognition results using multimodal expressive or structural dynamics-only features. The accuracy measure is unweighted average recall, and the per-class recall rate is shown below*

Type	*Expressive* Dynamics					*Structural* Dynamics				
Modality	Speech		Video		Speech + Video	Speech		Video		Speech + Video
States	Primary	Secondary	One	Simultaneous	Secondary+ One	Statistics	Transitions	Statistics	Transitions	All
mean-UAR	43.2	**50.7**	**62.6**	31.7	**65.6**	20.5	14.9	18	14.9	29.0
angry	13.3	33.3	60.0	20.0	40.0	20.5	0.0	0.0	0.0	26.7
frustration	53.6	63.2	79.0	36.8	73.7	21.1	89.5	84.2	89.5	21.1
happy	40.0	73.3	60.0	20.0	86.7	40.0	0.0	0.0	0.0	46.7
neutral	70.6	76.5	41.2	32.3	52.9	0.0	0.0	23.5	0.0	17.7
sad	38.9	38.9	66.7	27.8	77.8	88.9	0.0	0.0	0.0	55.6
surprise	43.8	18.8	68.8	50.0	62.5	0.0	0.0	0.0	0.0	6.25

Table 3: *A summary result on fusion of expressive dynamics and structural dynamics. The accuracy is presented as (best expressive-features / expressive + structural features)*

Type	*Expressive* + *Structural* Dynamics		
Modality	Speech	Video	Speech + Video
Feature	Secondary(S) +Transitions(S)	One(V) +Statistics(V) +Transitions(V)	Secondary(S) +One(V) +Statistics(V) +Transitions(V)
mean-UAR	50.7 / **54.7**	62.6 / **65.6**	65.6 / **72.5**
angry	33.3 / 40.0	60.0 / 60.0	40.0 / 40.0
frustration	63.2 / 68.4	79.0 / 79.0	73.7 / 79.0
happy	73.3 / 80.0	60.0 /66.7	86.7 / 86.7
neutral	76.5 / 52.9	41.2 / 41.2	52.9 / 76.5
sad	38.9 / 55.6	66.7 / 77.8	77.8 / 77.8
surprise	18.8 / 31.3	68.8 / 68.8	62.5 / 75.0

"one-speaker movement" is about twice the total duration of "simultaneous movement". The characteristics in the manifested body movement when only one of the dyad is moving possess the most significant amount of affective information toward the interaction. The relatively low accuracies achieved when using "simultaneous movement" features in the video modality may point to the fact that those movements are associated possibly more with the *logistics*, e.g., walking around the stage or simply changing positions, in carrying out an act and less on expressions of emotion.

Finally, our results indicate that the *structural*-based features, i.e., characterizing the macro-structure of interaction flow in audio and video modality, do not possess adequate discriminability in identifying the six dyad-level affect attributes by themselves; however, we will demonstrate their complementary nature to the *expressive* features in section 3.1.2.

3.1.2. Fusion of Expressive and Structural Dynamics

Table 3 shows a summary results in fusing *expressive* and *structure* dynamics features in speech, video, and multimodality, respectively. In each of the column, we present a comparison to the best accuracies obtained using *expressive* features in each behavior modality separately. In short, by fusing structural-based features, it improves the overall UAR across all three modalities (speech: 4% absolute, video: 3% absolute, and speech + video: 6.9%). The best accuracy achieved after fusing *expressive* and *structure* is 72.5%.

The improvement by integrating structural features comes mainly from a higher recognition rate achieved for the class of *sad* in speech-only and video-only modality. This result seems intuitive as the flow of the actors' dialogue in terms of generating an overall *sad* tone is likely to be different from all the other five emotion classes. In the "speech + video" modality, the improvement of structural features comes mainly in the *surprise* and *neural*. The improvement in *surprise* may have originated in the benefit of structural features in the speech modality. It is also foreseeable that in order to create an overall tone of *surprise*, there needs to be a distinctive back-and-forth macro-interaction turn-taking patterns existed between the dyad. In summary, we obtain a promising accuracy of 72.5% by fusing secondary speaker's acoustic information, one-speaker movements characteristics, and macro-structural information of video states in a six-class affective atmosphere recognition task.

4. Conclusions

In this work, we present a novel framework in deriving an extensive set of features, termed as *expressive* and *structural* dynamics, automatically computed from both interlocutors to measure their multimodal behavior manifestations and interactive dynamics in a task of six-class *dyad-level* affect recognition. The inclusion of *structural* features along with multimodality is essential in obtaining the final promising accuracy of 72.5%. Since the suite of features are naturally-interpretable, we further identify several interesting observations in terms of understanding the specific aspects of interaction that would shape the overall tone of the scene. For example, the secondary speaker and the single-speaker movement hold the most information in terms of the dyad-level affective attributes. Furthermore, *structural* features are useful in terms of characterizing certain types of dyad-level affective attributes, i.e., *sad* and *surprise*. Many of these analyses help provide quantitative evidence in understanding interlocutors' behaviors during face-to-face interaction at the *dyad*-level.

There are multiple future directions. One of the immediate work is to apply and adopt the framework toward quantitative analysis of dyadic interactions occurred within other domains, e.g., autism spectrum disorder, where not only the behavior manifestation but also the overall atypical interaction dynamics play an important role in the current clinical assessment protocol. Furthermore, there is a very limited availability on the publicly-available multimodal databases collected with rigorous design and annotation of *group- (dyad)* level affect. We will explore the opportunity to engage in inter-disciplinary collaborative research [17, 18] with domain experts, such as group scholars or organizational psychologists, to advance our technical framework in measuring human behaviors at multiple abstraction level (individual, dyad, group, etc). At the same time we hope to bring additional novel insights about social and communicative behaviors as human engage in group (dyad) interactions using computational methods.

5. References

[1] T. Sy, S. Côté, and R. Saavedra, "The contagious leader: impact of the leader's mood on the mood of group members, group affective tone, and group processes." *Journal of applied psychology*, vol. 90, no. 2, p. 295, 2005.

[2] N. Fay, S. Garrod, and J. Carletta, "Group discussion as interactive dialogue or as serial monologue: The influence of group size," *Psychological Science*, vol. 11, no. 6, pp. 481–486, 2000.

[3] S. G. Barsade and D. E. Gibson, "Group emotion: A view from top and bottom," *Research on managing groups and teams*, vol. 1, no. 4, pp. 81–102, 1998.

[4] J. R. Kelly and S. G. Barsade, "Mood and emotions in small groups and work teams," *Organizational behavior and human decision processes*, vol. 86, no. 1, pp. 99–130, 2001.

[5] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 966–979, 2012.

[6] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[7] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3687–3691.

[8] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio–visual affective expression recognition through multistream fused hmm," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

[9] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2462–2465.

[10] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions." in *Interspeech*, 2009, pp. 1983–1986.

[11] A. Metallinou, A. Katsamanis, and S. 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.

[12] Z. Yang and S. Narayanan, "Modeling dynamics of expressive body gestures in dyadic interactions," *IEEE Transactions on Affective Computing*, 2016.

[13] W. Mou, H. Gunes, and I. Patras, "Automatic recognition of emotions and membership in group videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 27–35.

[14] 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.

[15] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE transactions on pattern analysis and machine intelligence*, no. 4, pp. 532–550, 1987.

[16] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[17] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[18] S. Narayanan and P. 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.