Your Fellows Matter: Affect Analysis across Subjects in Group Videos

Wenxuan Mou[1,2], Hatice Gunes[2], Ioannis Patras[1]
[1]Queen Mary University of London, UK
[2]University of Cambridge, UK

Abstract— **Automatic affect analysis has become a well established research area in the last two decades. Recent works have started moving from individual to group scenarios. However, little attention has been paid to investigating how individuals in a group influence the affective states of each other. In this paper, we propose a novel framework for cross-subjects affect analysis in group videos. Specifically, we analyze the correlation of the affect among group members and investigate the automatic recognition of the affect of one subject using the behaviours expressed by another subject in the same group. A set of experiments are conducted using a recently collected database aimed at affect analysis in group settings. Our results show that (1) people in the same group do share more information in terms of behaviours and emotions than people in different groups; and (2) the affect of one subject in a group can be better predicted using the expressive behaviours of another subject within the same group than using that of a subject from a different group. This work is of great importance for affect recognition in group settings: when the information of one subject is unavailable due to occlusion, head/body poses etc., we can predict his/her affect by employing the expressive behaviours of the other subject(s).**

I. INTRODUCTION

Automatic affect analysis has attracted increasing attention and has seen much progress in recent years across various disciplines such as biology, psychology, neuroscience and computer science [1], [2], [3], [4], [5]. Automatic affect analysis aims to create a system capable of automatically interpreting, understanding and responding to emotions and moods displayed by humans, which has various applications in very diverse areas such as human-computer interaction [6], security [7], healthcare [8] and education [9].

Recent works in affective content analysis and affect recognition fields have started focusing on the analysis of naturalistic affect displayed and collected in more diverse and complex scenarios, such as dyadic interactions [10] and a group of people in a scene or involved in an interaction [11], [12], which is more challenging due to various poses of group members and the mutual influences among them. During dyadic/group settings, individuals tend to adapt their behaviours, i.e., synchronizing or mimicking their gestures and expressions with the other individuals in the same group/interaction. Such shared behaviours may help the cross-subjects affect recognition, i.e., analyzing the affect of a subject in a group using the behaviours of the other subject(s) in the same group. Specifically, given the face that the expressivity of subjects in a group is synchronized most of the time (showing similar behaviours), we hypothesize that (1) the affect of subjects in the same group are more correlated than that of subjects across different groups; (2) it is possible to predict the affect of a subject automatically using the behaviours expressed by the other subject(s) in the same group. To address this hypothesis, this work proposes the cross-subject affect analysis using a recently collected dataset containing a number of group videos.

Pioneering works in this direction (i.e., cross-subject affect analysis) have recently emerged in dyadic interaction settings [13], [14]. However, to the best of our knowledge, there are no works focusing on cross-subject affect analysis in group (here group refers to more than 2 people) videos. In this paper, we study the cross-subject affect analysis using two approaches, i.e., correlation analysis of the emotions expressed by the subjects and continuous affect prediction using expressive behaviours of the subjects in group videos.

The rest of the paper is organized as follows: related works are reviewed in Section 2; the proposed framework is illustrated in Section 3; the experiments and results are presented and discussed in Section 4; and conclusions and future work are described in Section 5.

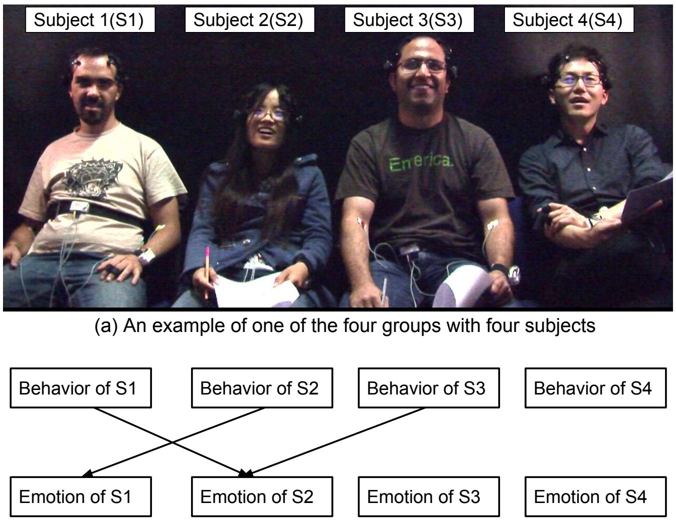
II. RELATED WORKS

A. *Affect analysis in group settings*

Most of the existing works on affect analysis have been carried out in *individual settings*. However, in the real world, people are very often being with others, and interacting in *group settings*. Some preliminary works have shown that the degree of variation and effect between individual and group settings is significant (e.g., differences in facial and body behaviors, timing and dynamics) [15], [16]. Therefore, in the past few years, a number of works have started paying attention to affect analysis in group settings. Dhall et al introduced a database named HAPPEI and inferred the *overall happiness mood intensities* conveyed by a group of people in static images in [12] and predicted the collective valence level (i.e., positive, neutral and negative) in [17]. An extended framework was introduced in [11] for recognizing both the arousal (i.e., high, medium and low) and valence (i.e., positive, neutral and negative) expressed by a group of people in static images. In addition to the aforementioned works, there are challenges organized in this field since 2016 [18], [19]. However, little attention has been paid to analyzing the influences among individuals in a group in terms of their affective states and expressive behaviours.

B. *Non-verbal cues for affect analysis*

Non-verbal behaviours are very important cues for affect analysis [20]. The most frequently used non-verbal behaviours include facial expressions, gaze patterns, body



(a) An example of one of the four groups with four subjects

(b) Predict the emotion of each individual based on the expressive behaviour of the neighbouring subject. For example, predict the emotion of subject 1 based on the behaviour of subject 2, while predict emotion of subject 2 based on the behaviour of subject 1 or 3 etc. (we tested both).

Fig. 1. An illustration of proposed framework.

motion and head movements [21], [22]. Despite various cues and modalities used for affect analysis, the mainstream on automatic affect recognition has mostly focused on facial features [23]. Facial representations include geometric and appearance representations. Facial geometric features are used to represent the shape of facial components and the location of facial salient points, such as the shape of mouth and eyes and the location of corners of a person's mouth and eye brown [24], while appearance features can represent the texture of the face such as wrinkles and furrows [3]. As this work aims at studying the cross-subject affect analysis, and not finding the best feature for affect analysis, the commonly used facial appearance features are utilized.

III. The proposed framework

We propose a framework for the analysis of valence and arousal of each individual in group videos across subjects. Our goal is to investigate (1) whether the affect of subjects in the same group is more correlated than that of subjects across different groups and (2) whether the affect of a subject can be better predicted using the behaviours (e.g., facial behaviours) expressed by another subject in the same group than using the behaviours expressed by a subject from a different group. The proposed framework is illustrated in Fig. 1 (a) and (b). Specifically, in this work, we explore the abovementioned problem in a pair-wised manner. More specifically, we focus on analyzing the affect of a subject using the neighbouring subject in the same group and the randomly paired subjects from two different groups. For example, as shown in Fig. 1, that is to predict the affect of subject 1 using the facial behaviours expressed by subject 2. For comparison purpose, we also provide the predicted affect of a subject using the facial behaviours expressed by a randomly selected subject in a different group. The details of the framework is presented below under four sections.

TABLE I

The stimuli videos listed with their sources (video IDs are stated in parentheses and used to refer to videos in the rest of the paper) and the video durations.

Movie	Duration/min
Descend (N1)	23:35
Mr. Bean (P1)	18:43
Batman the Dark Knight (B1)	23:30
Up (U1)	14:06

A. Data

We utilize a database collected to study affect analysis in group settings while each group (i.e., four participants) were watching a number of long movie segments. Four long movie segments (duration of each longer than 14 mins and smaller than 24 mins) were used as stimuli, details of which are provided in Table I. Twelve participants (7 females and 5 males) from 8 different countries, aged between 25 and 38 were recorded while watching these movies. They were arranged into three groups with four participants in each group watching all of the four videos listed in Table I together. A representative frame from the database is shown in Fig. 1 (a).

The annotation was conducted by human labelers, three researchers who are experienced with affect analysis. Independent observer annotations were obtained by using an in-house affect annotation interface that requires the labelers to scroll a bar between a range of continuous values (-0.5 and 0.5). The labelers were asked to give one label for valence and one label for arousal for every 20 seconds starting from the beginning of each recording (e.g., the interval for 00:00∼00:20 min, 00:21∼00:40 min etc). The labeler annotated arousal and valence separately to avoid the confusion between these two dimensions; the 20-second recordings were played in a random order to each labeler; each labeler was asked to observe the visual behaviors without hearing any audio and rate a single annotation for each 20-second recording along either arousal or valence dimension. Each of the labelers annotated all of the video segments, which means that each video segment obtained three annotations from all of the three labelers.

B. Facial feature extraction

Intraface [25] was used to detect all faces in the videos and 49 facial points were obtained for each face. Due to illumination and head pose variations in such a naturalistic scenario, it is difficult to detect all faces. As a result, manual inspection showed that 96% of faces were detected for this dataset.

Quantised Local Zernike Moments (QLZM) [26] is a recently introduced appearance representation for automatic affect prediction. Quantised Local Zernike Moments (QLZM) facial representation outperformed other facial and body representations in [16]. Therefore, in this work, we use *QLZM* for facial representation. After faces are detected, QLZM [26] are obtained from the local patch around each facial landmark point as the appearance representation.

QLZM is used as a low-level representation that is extracted by first calculating local Zernike Moments (ZMs) in the neighbourhood of each pixel of the input image. Then the accumulated local features are converted into position dependent histograms. Each ZM coefficient describes the texture variation at a unique scale and orientation. Once the ZMs are computed for all pixels, the QLZM descriptors are obtained by quantising all ZM coefficients around a pixel into a single integer. In this way, the frame-level based facial appearance feature is extracted.

C. Correlation analysis of emotions

To investigate whether the affect of subjects in the same group is more correlated than that of subjects across different groups, we start our analysis from the ground truth level, i.e., the emotion label obtained by averaging the labels from the three human labelers. We utilize Pearson's Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC) [27] as the evaluation methods. CCC combines the PCC with the square difference between the means:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{1}$$

where ρ is the PCC between the ground truth and prediction, σ_x^2 and σ_y^2 are the variance, and μ_x and μ_y are the mean of ground truth and prediction respectively. In this way, the predictions that are correlated well with the ground truth but are shifted and penalized by the deviation.

The correlation of subjects in one group between neighbours are first calculated, which are denoted as C_{ii}, where i denotes the group of the subject from, and $i = 1, 2, 3, 4$; and then that of subjects across different groups are denoted as C_{ij}, where i, j denote the group of the subject from and $i, j = 1, 2, 3, 4$, $i \neq j$.

D. Automatic affect prediction

The automatic analysis is designed in the same way as the correlation analysis. It is divided into two parts, i.e., (1) to predict the affect of one subject using the facial behaviours expressed by the neighbouring subject, which is referred to as f_{ii}, where i denotes the group of the subject from, and $i = 1, 2, 3, 4$; and (2) to predict the affect of one subject using the facial behaviours of a randomly selected subject from the other groups, which is referred as f_{ij}, where i, j denote the group of the subject from and $i, j = 1, 2, 3, 4$, $i \neq j$. f_{ij} is randomly paired for 10 times for each subject.

We utilize Long Short-Term Memory (LSTM) [28] and facial QLZM feature for affect regression as shown in Fig. 2. LSTM is one of the state-of-the-art temporal modeling methods. LSTM is trained using the frame-level facial features and take each 20-seconds clip as a sequence as shown in Fig. 2, which is implemented on PyTorch platform [29].

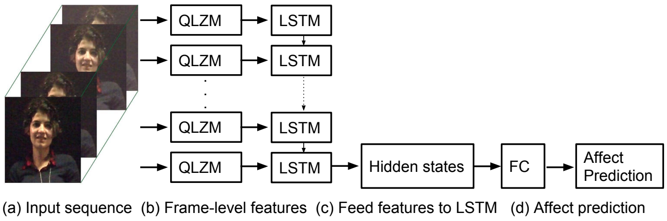


Fig. 2. Illustration of the affect analysis approach using LSTM. (a) Input sequence, the 20-seconds clip. (b) Frame-level features are extracted. (c) Features extracted from every frame are fed into a one-layer LSTM with 128 hidden states. (d) Affect prediction results obtained.

IV. Experiments and analysis

A. Experiments

1) Data: Group videos from four groups are used in the experiments. Four movies were watched by four groups and in total there should have been 16 sessions (4 movies x 4 groups). One session refers to the recording of one group watching one movie. Unfortunately, one session was corrupted, and there are 15 sessions left, i.e., three groups with recordings of people watching four movies (N1, P1, B1 and U1) and one group with recordings of people watching three movies (B1, N1 and U1). For each session, 20-seconds clips in line with the annotations labeled are utilized. The number of the 20-second clips from different sessions varies with the length of the movies, i.e., 70 clips for N1, 70 for B1, 56 for P1 and 42 for U1. As a result, the total number of clips we use in our experiments is $(70(B1) \times 4(4subjects) \times 4(4movies)) + (70(N1) \times 4(4subjects) \times 4(4groups)) + (56(P1) \times 4(4subjects) \times 3(3groups)) + (42(U1) \times 4(4subjects) \times 4(4groups)) = 3584$.

2) Evaluation metrics: The experimental results of affect regression is evaluated with the Pearson's Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC).

B. Experimental results and analysis

1) Correlation analysis of emotions: These correlation results for both C_{ii} and C_{ij} along arousal and valence dimensions are presented in Table II. The correlation of subjects across different groups are paired randomly. Specifically, for each subject, we randomly select a subject from the other three groups to pair with. We randomly selected all pairs for 10 times and the average of this is reported in Table II. We can see that the affect of neighbouring subjects in the same group are much more correlated (i.e., 0.500 and 0.531 in terms of CCC along arousal and valence respectively) than

TABLE II

PCC and CCC of subjects in one group between neighbours (a pair-wise way) and PCC and CCC of subjects across different groups (average of randomly paired for 10 times)

Dimensions	Arousal		Valence	
	PCC	CCC	PCC	CCC
C_{ii}	0.537	0.500	0.578	0.531
C_{ij}	0.365	0.319	0.358	0.325

TABLE III

The recognition results in terms of PCC and CCC along both arousal and valence dimension for neighbours in one group (f_{ii}) and randomly paired across groups (f_{ij}). The significance test (p-value) is reported between f_{ii} and f_{ij}.

Dimensions	Arousal		Valence	
	PCC	CCC	PCC	CCC
f_{ii}	0.469	0.384	0.531	0.440
f_{ij}	0.301	0.232	0.321	0.257
$f_{ii} > f_{ij}$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$

TABLE IV

Recognition results obtained with f_{ij} (randomly selected pairs across different groups) along both arousal and valence in terms of PCC and CCC. All pairs are randomly selected for 10 times and the average across all subjects is reported. The statistical significance (p-value) between f_{ii} and f_{ij} is also presented in parentheses. * refers to $p < 0.05$ and ** refers to $p < 0.01$.

Dimensions	Arousal		Valence	
	PCC	CCC	PCC	CCC
f_{ij}^{1}	0.309(**)	0.232(**)	0.354(**)	0.272(**)
f_{ij}^{2}	0.320(**)	0.265(**)	0.307(**)	0.238(**)
f_{ij}^{3}	0.308(**)	0.240(**)	0.297(**)	0.258(**)
f_{ij}^{4}	0.271(**)	0.212(**)	0.316(**)	0.271(**)
f_{ij}^{5}	0.307(**)	0.246(**)	0.279(**)	0.231(**)
f_{ij}^{6}	0.346(**)	0.281(**)	0.292(**)	0.224(**)
f_{ij}^{7}	0.296(**)	0.225(**)	0.332(**)	0.228(**)
f_{ij}^{8}	0.287(**)	0.202(**)	0.360(**)	0.316(*)
f_{ij}^{9}	0.293(**)	0.211(**)	0.295(**)	0.227(**)
f_{ij}^{10}	0.268(**)	0.204(**)	0.375(**)	0.304(**)

that of randomly paired subjects across different groups (i.e., 0.319 and 0.325 in terms of CCC along arousal and valence respectively).

2) Automatic affect prediction: The recognition results of f_{ii} and f_{ij} and the significance test ($p - value$) for comparison of f_{ii} and f_{ij} are provided in Table III. And the detailed results of the 10 times pairing are presented in Table IV. We can clearly see that the prediction results obtained with f_{ii} (i.e., 0.384 and 0.440 in terms of CCC along arousal and valence respectively) are significantly better than that obtained with f_{ij} (i.e., 0.232 and 0.257 in terms of CCC along arousal and valence respectively). The possible reason is that people in the same group share some facial behaviours, that contributes to the affect prediction in line with our hypothesis. As a result, when the expressive behaviours of a subject is not available due to occlusion or head poses which are challenging for affect analysis in group settings, the behaviours expressed by the other subject(s) can be used for the affect prediction of that subject.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed a framework to investigate the cross-subject affect analysis in group videos. We conducted a set of experiments using a recently collected database that aims to study affect analysis with a group of people watching stimuli movies. The experimental results show that (1) the affect of subjects in the same group is more correlated than that of subjects across different groups and (2) the affect of a subject predicted using facial behaviours expressed by the other subject in the same group is significantly better than that predicted using the behaviours of a subject in a different group. This is possibly due to the fact that people in the same group share some information and are influenced by each other in terms of facial behaviours and emotions. Consequently, this is expected to help address one of the main challenges for affect analysis in group settings, i.e., inability to predict facial affect due to occlusion among subjects or due to head pose variations. Specifically, when the information of one subject is unavailable, we can predict the affect of that subject based on the expressive behaviours of the other subject(s).

Although the experiments reported in this paper have been conducted in an audience context (participants watching movies), the proposed method can be applied to other settings, including human-robot and human-human interactions. Despite the promising results obtained, the impact of other signals such as social context, better feature representations and learning methods need to be investigated further.

VI. ACKNOWLEDGMENTS

The work of Wenxuan Mou is supported by CSC/Queen Mary joint PhD scholarship. The work of Hatice Gunes has partially been supported by the European Unions Horizon 2020 research and innovation programme under grant agreement No 826232. This work is also supported by Nvidia corporation with the donation of a TitanX GPU.

References

[1] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, 2013.

[2] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image and Vision Computing*, 2013.

[3] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[4] D. Matsumoto, "Cultural influences on the perception of emotion," *Journal of Cross-Cultural Psychology*, 1989.

[5] M. Soleymani, S. Koelstra, I. Patras, and T. Pun, "Continuous emotion detection in response to music videos," in *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition and Workshops (FG)*, 2011.

[6] K. Dautenhahn, "Socially intelligent robots: dimensions of human–robot interaction," *Philosophical Trans. of the Royal Society B: Biological Sciences*, 2007.

[7] J. Hernandez, M. Hoque, W. Drevo, and R. W. Picard, "Mood meter: counting smiles in the wild," *Association for Computing Machinery*, 2012.

[8] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *International Symposium on Visual Computing*, 2012.

[9] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. on Affective Computing*, 2013.

[10] 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*, 2008.

[11] W. Mou, O. Celiktutan, and H. Gunes, "Group-level arousal and valence recognition in static images: Face, body and context," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition and Workshops (FG)*, 2015.

[12] A. Dhall, R. Goecke, and T. Gedeon, "Automatic group happiness intensity analysis," *IEEE Trans. on Affective Computing*, 2015.

[13] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Trans. on affective computing*, vol. 4, no. 2, pp. 183–196, 2013.

[14] S. N. Fatima and E. Erzin, "Cross-subject continuous emotion recognition using speech and body motion in dyadic interactions," *Proc. Interspeech 2017*, 2017.

[15] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A comparative analysis of facial affect recognition," in *Proc. of ACM Int. Conf. on Multimedia*, 2016.

[16] ——, "Automatic recognition of emotions and membership in group videos," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition and Workshops (CVPRW)*, 2016.

[17] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2015.

[18] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: video and group-level emotion recognition challenges," in *Proc. of ACM Int. Conf. Multimodal Interaction (ICMI)*, 2016.

[19] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proc. of ACM Int. Conf. on Multimodal Interaction (ICMI)*, 2017.

[20] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative Science Quarterly*, 2002.

[21] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. on Multimedia*, 2012.

[22] U. Avci and O. Aran, "Effect of nonverbal behavioral patterns on the performance of small groups," in *Proc. of ACM workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, 2014.

[23] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2011.

[24] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2006.

[25] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[26] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro, "Local Zernike Moment representation for facial affect recognition," in *Proc. of Brithish Machine and Vision Conference (BMVC)*, 2013.

[27] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. Int. Workshop Audio/Visual Emotion Challenge*, 2015.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.