Automatic Recognition of Emotions and Membership in Group Videos

Wenxuan Mou
Queen Mary University of London, UK
w.mou@qmul.ac.uk

Hatice Gunes
University of Cambridge
Cambridge, UK
Hatice.Gunes@cl.cam.ac.uk

Ioannis Patras
Queen Mary University of London, UK
i.patras@qmul.ac.uk

Abstract

Automatic affect analysis and understanding has become a well established research area in the last two decades. However, little attention has been paid to the analysis of the affect expressed in group settings, either in the form of affect expressed by the whole group collectively or affect expressed by each individual member of the group. This paper presents a framework which, in group settings automatically classifies the affect expressed by each individual group member along both arousal and valence dimensions. We first introduce a novel Volume Quantised Local Zernike Moments Fisher Vectors (vQLZM-FV) descriptor to represent the facial behaviours of individuals in the spatio-temporal domain and then propose a method to recognize the group membership of each individual (i.e., which group the individual in question is part of) by using their face and body behavioural cues. We conduct a set of experiments on a newly collected dataset that contains fourteen recordings of four groups, each consisting of four people watching affective movie stimuli. Our experimental results show that (1) the proposed vQLZM-FV outperforms the other feature representations in affect recognition, and (2) group membership can be recognized using the non-verbal face and body features, indicating that individuals influence each other's behaviours within a group setting.

1. Introduction

Automatic affect analysis has attracted increasing attention and has seen much progress in recent years [34]. However, little attention has been paid to the analysis of the affect expressed by a group of people in a scene or in an interaction setting. From psychological perspective, the affect of each individuals is influenced by the overall group [3]. From the automatic analysis perspective, Leite *et al.* [20] reported that individual disengagement could be modelled differently in individual and group settings; and their results indicated that more diverse types of disengagement behaviours were shown in group settings than in individual settings. Therefore, it would be also interesting to study the individual affect expressed in a group setting. However, to the best of our knowledge, most of the existing works on affect analysis in group settings focus on the automatic recognition of collective group-level emotions in static images [10, 22]; and little attention has been paid to the automatic affect analysis of each individual member in a group setting.

In this paper, we introduce a framework for automatic affect analysis of individual members in group videos. To this end, we extract spatio-temporal face and body information to analyse the affective states along both valence and arousal dimensions. For facial representation, we extend the static Quantised Local Zernike Moments (QLZM) to the temporal domain and extract Fisher Vectors to investigate on their performance in the domain of affect recognition. Our analysis shows that (1) different modalities (i.e., face or body) and different types of information (local appearance deformations or motion) contribute to the different recognition tasks in different ways, and (2) that fusion of information from the different modalities, in our case by an SVM trained on the soft outputs of the individual modality classifiers, almost always leads to performance improvements. Finally, we show that face and body features are informative in predicting contextual information, namely the group each individual belongs to (i.e., group membership).

The rest of paper is organized as follows: the 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.

2. Related Work

Affect analysis in group settings. Recent works in affect recognition fields have started focusing on the analysis of spontaneous affect displayed by multiple people in more naturalistic settings. Pioneering works in this area focused on analysing the overall group-level emotions. 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 [9] and predicted the collective valence level (i.e., positive, neutral and negative) in [10]. An extended framework was introduced in [22] 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. Although the individual information was used in the aforementioned works to predict the group-level affect, all of these works focus on the collective affect expressed by the whole group rather than analysing the affect displayed by each individual. In addition, all of the works focus on static images rather than dynamic videos. Videos instead naturally enable the use of temporal as well as spatial information which are very informative for recognizing human affect. Therefore, in this paper we focus on affect analysis of the individuals in group videos from spatio-temporal face and body features.

Multimedia content evoked affect database and analysis. As one of the primary functions of multimedia content (e.g., music and movies) is to regulate the users' affect, how to represent and predict the users' affective states while they are exposed to multimedia content is becoming an increasingly popular topic. In the literature, a number of databases for decoding the users' affective responses to multimedia content have been introduced already [29, 17, 1, 30]. In these databases, both videos and physiological signals (e.g., EEG and GSR) are provided, but all of them are limited to having only one subject in each session. Instead of using these databases, we have collected and annotated a new database because we are interested in exploring how an individual displays affect when exposed to multimedia content in a group setting.

Face and body features. Face and body expressions are widely used non-verbal information for automatic emotion recognition [13]. The most frequently used face features include geometric features and appearance features. Specifically, geometric features can represent the shape of the facial components (e.g., eyes and mouth) and the location of facial salient points (e.g., corners of the eyes and mouth) [23]; and appearance features represent the facial texture including wrinkles, bulges and furrows [27]. Recently affective computing field has started paying an increasing attention to body expressions, in the form of body movement and postures [12]. Body expressions have shown to be particularly useful for predicting the level of arousal [16]. Therefore, in this paper we extract both face and body features for arousal and valence recognition. However, differently from previous work [28], we (1) extend the static Quantised Local Zernike Moments (QLZM) to volume representation and (2) encode all body and face descriptors to Fisher Vector representations to embed spatio-temporal information.

Context in affect analysis. Studies have shown that the displayed affect heavily depends on context, such as where the person is and what the person is doing at that time [31]. Therefore, in addition to the face and body information, using context information is becoming increasingly popular for automatic affect recognition [21]. Especially group settings with multiple people inherently involve complex contextual situations, not only in terms of each individual's identity, location and task but also in terms of interpersonal dynamics, e.g., who the person is with and what others are doing at that time. The contextual information based on the group structure was used to infer group-level affect in [22] and individual gender and age in [11]; and scene contextual features were utilized to predict group-level affect information in [9]. Contextual information can be directly used as a cue for affect analysis or it can be fused with other attributes, e.g., facial expressions and body motion. In this paper, group membership is referred to as context and is recognized by using dynamic non-verbal behaviours. Specifically, we automatically recognize which group each individual is part of using their face and body features. .

3. The Proposed Framework

We propose a framework for the prediction of individual emotions and group membership in group videos by multimodal analysis of face and body features. The proposed framework is illustrated in Fig. 1. We are interested in investigating (1) the individual affect responses when the participants are watching long-term videos (i.e., 14-24 mins) in group settings; and (2) group membership recognition by using visual cues. To this end, both face and body features are extracted. For representing faces, both geometric and appearance features are utilised. Facial landmark trajectories are used as geometric features and the extended volume QLZM extracted along facial landmark trajectories are used as appearance features. For representing bodies, dense trajectories are first extracted and then Histogram of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF) descriptors are extracted along the trajectories. Prior to being fed to the classifier and regressor, all of the descriptors are encoded into Fisher Vector (FV) representations. Multiple experiments are carried out to investigate the subject-dependent and subject-independent affect classification and regression using unimodal and multimodal visual signals. A set of experiments is also conducted to recognize group membership (i.e., which group each individual is part of) using face and body behavioural cues.

3.1. Low Level Feature Extraction

Face features. Prior to facial feature extraction, Intraface [33] is first used to detect all faces in the videos and 49 facial points are 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, in 96% of the frames the faces of four subjects are detected. To make the feature extraction consistent, in the case that the face is lost

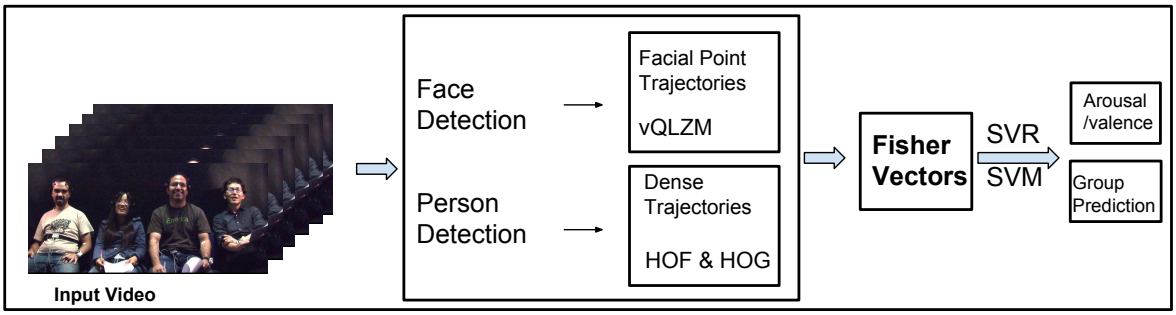

Figure 1. Illustration of the proposed framework.

in a frame, the last position in which the face has been detected is used.

For geometric features, let $X_t = [\ (x_t^1,\ y_t^1),\ (x_t^2,\ y_t^2)\ ...\ (x_t^n,\ y_t^n)\ ]$ denotes the location of the n facial landmark points at the current frame t, where $n = 49$ is the total number of facial landmark points. x_t^k and y_t^k are the coordinates of the k-th facial landmark point at frame t. Facial landmark points of the subsequent frames are concatenated to form the facial landmark trajectories. The shape of the trajectory of each facial landmark point encodes the motion patterns [32]. For the k-th facial landmark point, it is described by a sequence $(\Delta X_t^k,\ \Delta X_{t+1}^k\ ...\ \Delta X_{t+L-1}^k)$ of displacement vectors, where $\Delta X_t^k = (X_{t+1}^k - X_t^k) = (x_{t+1}^k - x_t^k, y_{t+1}^k - y_t^k)$ and L is the length of the facial landmark trajectories. The resulting vector is then normalized by the sum of the displacement vector magnitudes:

$$Y^k = \frac{(\Delta X_t^k, \Delta X_{t+1}^k ... \Delta X_{t+L-1}^k)}{\sum_{j=t}^{t+L-1} ||\Delta X_j^k||} \quad (1)$$

We refer to Y^k as *Facial Landmarks* in the rest of the paper. The fixed length of the facial landmark trajectories is $L = 15$ frames based on [32], therefore, a 30 $(2 \times L)$ dimensional descriptor is obtained around each facial landmark point.

For appearance features, after geometric features are extracted, Quantised Local Zernike Moments (QLZM) [28] are obtained from the local patch around each facial landmark point as the appearance representation. QLZM is 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. The QLZM [28], by design, takes into account only static spatial information, that is it is designed for static images/frames [28, 5]. In this paper, we extend it to the spatio-temporal domain to embed both appearance and temporal information, as illustrated in Fig. 2. This feature is referred as *vQLZM* in the rest of the paper. The size of the volume is $N \times N$ pixels and L frames long. To keep the same volume size with the *Facial Landmarks*, the same length $L = 15$ is used. To embed structure information, the volume is subdivided into a spatio-temporal grid of size $n_\tau \times n_\tau \times n_\sigma$. The QLZM descriptor is computed in each cell of the spatio-temporal grid. The final descriptor is obtained by concatenating these descriptors. In the experiments, N is set to $N = 24$, i.e., the average of the distances between the centroids of two eyes from all of the detected faces.

Body features. In order to extract person-based representations we first need to apply a person detector - in our simplified settings with a fixed number of individuals and a static camera, we use an ad-hoc scheme that divides the frame in equally sized parts. Then, dense trajectories [32] are extracted and, subsequently, HOG and HOF descriptors are extracted around each trajectory. The latter are computed in the spatio-temporal volume aligned with the trajectories similarly to *vQLZM* features. For both HOG and HOF, orientations are quantized into eight bins with full orientations. An additional zero bin is added for HOF for pixels with optical flow magnitudes lower than the threshold (i.e., nine bins in total). Thus, the final descriptor size is 96 for HOG and 108 for HOF with the trajectory length $L = 15$ frames. More details can be found in [32]. These two features are referred to as *body HOG* and *body HOF* in the remaining of the paper.

3.2. Fisher Vector Encoding

Fisher vector (FV) encoding [26] has been widely used in computer vision problems such as action recognition [32] and depression analysis [15, 8]. It encodes both the first and the second order statistics between the low-level (local) video/image descriptors and a Gaussian Mixture Model (GMM). To reduce the dimensionality, Principal Component Analysis (PCA) is first applied to the descriptors. A

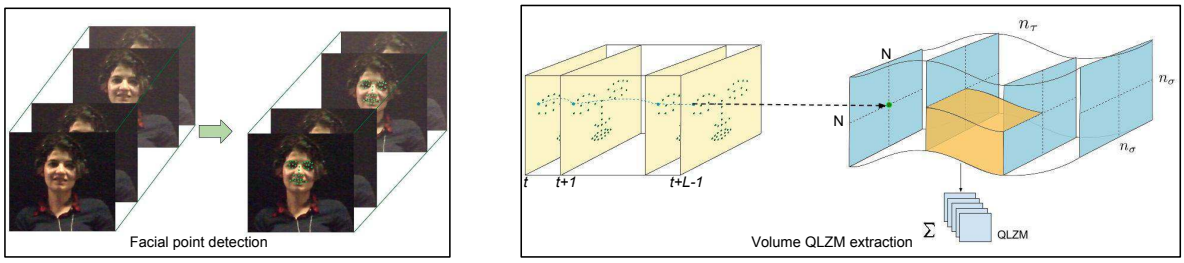

Figure 2. Illustration of our approach to extract the vQLZM feature. Left: Facial landmark points are detected. Right: Facial landmark point tracking is in the spatial scale over L frames. Appearance and motion information over a local neighbourhood of $N \times N$ pixels along the each facial landmark point are extracted. In order to embed the structure information, the local volume is subdivided into a spatio-temporal grid of size $n_\tau \times n_\sigma$. Based on [32], $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$.

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

Table 1. 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.

GMM is then fitted to all of the face and body descriptors. The number of Gaussians is set to $K = 256$ and a subset of 256000 descriptors is randomly sampled to fit a GMM. Subsequently, each clip is represented by a $(2D + 1)K$ dimensional Fisher Vector, where D is the dimensionality of the descriptor after performing PCA. We obtained four types of Fisher Vectors (FVs) from *Facial Landmarks, vQLZM, body HOG* and *body HOF* separately.

4. Experiments and Analysis

Experiments are conducted using a database collected to study affect analysis from multimodal cues in group settings while each group (i.e., four participants) were watching a number of long movie segments.

4.1. Data and Annotation

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 1. Sixteen participants (8 females and 8 males), aged between 25 and 38 were recorded while watching these movies. They were arranged into four groups with four participants in each group watching all of the four videos listed in Table 1 together. Videos were recorded at 1280×720 resolution, 25fps. A representative frame from the database is shown in Fig. 3.

Annotation. Independent observer annotations were obtained from three human labellers who are all researchers working on affect analysis. An internal emotion annotation

Figure 3. A representative frame from the database.

tool that requires the labellers to scroll a bar between a range of values (0 and 1) was used. The labellers were asked to annotate 10-second recordings for every 2 minutes starting from the first minute, e.g., the interval for 00:50∼1:00 min, 2:50∼3:00 min etc. Each labeller was presented with the 10-second recordings of each subject and was asked to observe the non-verbal behaviours without hearing any audio. A single annotation was given by each labeller after watching one 10-second recording. In order to avoid confusion, arousal and valence annotations were obtained separately.

Analysis of Annotations. To assess the inter-labeller agreement, Cronbach's α [6] and Fleiss' Kappa [4] statistic, widely used in literature, were computed. The Cronbach's α was calculated directly from the continuous annotations. As Fleiss' Kappa can only be used for the categorical ratings, prior to computing the Fleiss' Kappa, both arousal and valence annotations were first quantised into two classes using the average of all of the annotations as thresholds (i.e., 0.4 for arousal and 0.5 for valence). In this way, arousal is quantised into *high* and *low* and valence is quantised into *positive* and *negative*. After the first annotation round, the Cronbach's α was computed for each subject and the average of all subjects. The displays of subjects with Cronbach's α below the average were re-annotated through discussions, and each labeller's annotation was subsequently normalised using Equation 2, where

Dimension	Arousal		Valence	
Methods	Cronbach	Kappa	Cronbach	Kappa
Raw	0.85	0.48	0.85	0.56
Reannotated	0.95	0.74	0.85	0.61
Normalized	0.95	0.73	0.85	0.75

Table 2. The measurement of inter-labeller agreement on both arousal and valence dimensions among 3 labellers in terms of Cronbach's α and Fleiss' Kappa for the raw, re-annotated and normalized ratings.

$X = [x_1,\ x_2,\ x_3\ ...\ x_n]$ refers to all annotations from one labeller, and n is the number of the 10-second recordings.

$$z_i = \frac{x_i - min(X)}{max(X) - min(X)} \tag{2}$$

From the Table 2, it can be seen that the results after the re-annotation and normalization indicate a very strong inter-labeller reliability for both arousal and valence dimensions. The average of annotations from three labellers are used as the ground truth.

4.2. Experiments

Experimental setup. Data from 4 groups were used in our experiments, namely 3 groups (12 subjects) with recordings from 4 movies (N1, P1, B1 and U1) and 1 group (4 subjects) with recordings from 2 movies (N1 and P1). As a result, there were data from 16 subjects and 14 sessions in total. During each session, each group watched one movie. From each session, we used 10-seconds clips extracted every 2 minutes in line with the annotations obtained. The number of short clips from each session varies with the length of the movies, i.e., 12 clips for N1 and B1, 9 clips for P1 and 7 clips for U1. Therefore, the total number of clips we used in the experiments is $(12 \times 4 \times 4) + (12 \times 4 \times 4) + (9 \times 4 \times 3) + (7 \times 4 \times 3) = 576$.

Different classification and regression models were trained by applying *leave-one-sample-out and leave-one-subject-out* cross-validation. Each time the parameters of the model were optimized over the training-validation samples. *Leave-one-sample-out* means, in each fold, 575 out of 576 clips were used for training-validation and the remaining one clip was used for testing. *Leave-one-subject-out* refers to, in each fold, using 15 subjects for training-validation and the remaining one subject for testing. *Subject-specific* model was built by applying *leave-one-sample-out* cross-validation for each subject separately.

The experimental results of affect classification were evaluated by the average of $F1$ score (average of $F1$ score for both classes). For affect regression, in addition to the Mean Absolute Error (MAE) and the Mean Squared Error (MSE), we also used the Pearson's Correlation Coefficient (CC) and Concordance Correlation Coefficient (CCC) [24]. CCC combines the CC with the square difference between the means:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3}$$

where ρ is the CC 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, are penalised by the deviation.

The experiments were divided into two parts, affect recognition and group membership recognition. Affect recognition includes both classification and regression along arousal and valence dimensions. The experiments for affect recognition were conducted with both unimodal feature vectors and using decision-level fusion. An SVM classifier, applied on the soft outputs of the single-modality classifiers, was used for decision-level fusion. Group membership recognition aimed to recognize the group that each individual belongs to, and was conducted by using face and body behavioural cues. *Leave-one-subject-out* cross-validation was used for group membership recognition. In both cases, the publicly available LibSVM library [7] was used. Prior to feeding face and body features to any classifier or regressor, PCA was first applied to reduce the dimensionality by preserving 99% of the variance.

Affect recognition. Linear Support Vector Machine (SVM) was used to do classification with respect to the dimensions of arousal (i.e., high and low) and valence (i.e., positive and negative). The classification results with unimodal features and decision-level fusion are illustrated in Table 3. It can be seen that different features provide different classification results. On the one hand, *vQLZM-FV* generally outperforms the other feature representations in all recognition tasks except that *Facial Landmark FV* provides slightly better performance for *leave-one-subject-out* arousal classification. This indicates that the spatio-temporal texture information encoded in the proposed *vQLZM-FV* descriptor is very informative for affect analysis. The results also reveal that body information is more powerful in predicting arousal. For instance, the *F1* scores obtained with *body HOG* and *body HOF* for arousal are 0.63 and 0.61 respectively, but those obtained for valence are 0.55 and 0.58 respectively in *leave-one-sample-out* cross-validation setup. This result is consistent with the literature that reports arousal to be better conveyed by body information than valence [16]. On the other hand, although the *leave-one-sample-out* and *subject-specific* models show better performance than the *leave-one-subject* model due to the subject-dependency problem, there are only slight differences between these results. Overall, the results obtained show that it is possible to build a generic model across different subjects using the proposed framework.

In decision-level fusion, we combined the decision values (i.e., the probability for each class) from individual modalities using a linear-SVM. The fusion results show that when the features are fused at the decision level, all of the classification results are equal to or better than those obtained by using unimodal features. For instance, the best classification results from unimodal features are 0.63 for arousal and 0.70 for valence in terms of $F1$ score in *leave-one-subject-out* experiments; and those obtained from decision fusion are 0.64 for arousal and 0.70 for valence. Thus, the fusion of different features is generally useful for improving the classification performance. However, it can be seen that using the proposed *vQLZM-FV* feature alone can achieve very close results to that obtained by fusing all the different features. In terms of the computational complexity, much more effort is required to generate dense trajectories and encode FVs. Taking both the classification performance and computational complexity into consideration, using *vQLZM-FV* alone proves to be more advantageous. In order to investigate how different features contribute to the decision fusion, we checked the weights assigned to each type of feature as shown in Fig. 5. It can be seen that face / body motion information (i.e., Landmarks and body HOF features) is more informative for arousal analysis, while appearance features, especially local facial appearance (i.e., vQLZM), tend to be more helpful for valence classification.

For affect regression, Support Vector Regression (SVR) was used with a radial basis function (RBF) kernel. The regression results using unimodal and multimodal features are summarized in Table 4. For unimodal features, it can be seen that the regression results are quite similar to the classification ones: *vQLZM-FV* generally has the best performance and body features have better prediction for arousal than valence. For the decision-level fusion, we proceeded in a similar way to the fusion in affect classification. We fused the ratings predicted from individual regressors in an RBF-SVR. The results show that using the proposed *vQLZM-FV* feature alone can achieve very close results to that obtained by multimodal fusion.

Finally, we also looked into the affect recognition results of each group separately for both arousal and valence. We found that in each group setting, there were a number of subjects for which the prediction of affect was more challenging. This is possibly due to the fact that different subjects display affect at different levels of subtlety (as shown in Fig. 4), leading to a more challenging recognition problem. For instance, during fusion with *leave-one-subject-out* cross-validation, prediction was better and relatively easier for subject 1 in Fig. 4 (i.e., $F1 = 0.91$ for valence and $F1 = 0.74$ for arousal) than subject 2 (i.e., $F1 = 0.59$ for valence and $F1 = 0.48$ for arousal).

Membership recognition. This experiment focuses on automatically recognizing the group that each individual belongs to. The experimental results are shown in Table 5 and the confusion matrix with *decision-fusion* is shown in Table 6. From Table 5, we can see that both body features work better than the face features in predicting membership. The decision-level fusion helps improve the membership recognition result. Compared to chance level recognition of 25%, our results obtained with *decision-fusion* of 43.75% is a significant result. This result indicates that non-verbal behaviours, especially body behaviours, are influenced and

Dimensions	Arousal	Valence
	F_1	F_1
One sample out		
vQLZM	**0.73**	**0.79**
Facial Landmarks	0.65	0.54
body HOG	0.63	0.55
body HOF	0.61	0.58
Decision-fusion	***0.73***	***0.79***
One subject out		
vQLZM	0.61	**0.70**
Facial Landmarks	**0.63**	0.51
body HOG	0.58	0.53
body HOF	0.61	0.60
Decision-fusion	***0.64***	***0.70***
Subject-specific		
vQLZM	**0.74**	**0.78**
Facial Landmarks	0.64	0.58
body HOG	0.62	0.51
body HOF	0.64	0.53
Decision-fusion	***0.79***	***0.79***

Table 3. The classification results along arousal and valence dimensions with unimodal features, vQLZM-FV, Facial Landmarks FV, body HOG FV and body HOF FV respectively and the decision fusion results with these four different features.

Figure 4. Ground truth (and predicted) affective states of two different subjects. In Scene 1: Subject 1 high (high) arousal and positive (positive) valence; Subject 2 low (low) arousal and positive (negative) valence. In Scene 2: Subject 1 low (low) arousal and negative (negative) valence, Subject 2 low (low) arousal and negative (negative) valence.

Dimensions	Arousal				Valence			
	MAE(std)	MSE(std)	CC	CCC	MAE(std)	MSE(std)	CC	CCC
Leave one sample out								
vQLZM	0.088(0.07)	0.013(0.02)	**0.62**	**0.53**	0.090(0.08)	0.015(0.03)	**0.54**	**0.47**
Facial Landmarks	0.110(0.07)	0.017(0.02)	0.43	0.24	0.110(0.09)	0.020(0.03)	0.27	0.10
body HOG	0.106(0.08)	0.017(0.02)	0.40	0.32	0.110(0.09)	0.020(0.04)	0.28	0.18
body HOF	0.100(0.07)	0.016(0.02)	0.46	0.32	0.110(0.09)	0.020(0.04)	0.27	0.14
Decision-fusion	*0.086(0.07)*	*0.092(0.02)*	***0.62***	***0.53***	*0.082(0.09)*	*0.014(0.03)*	***0.56***	***0.49***
Leave one subject out								
vQLZM	0.111(0.08)	0.009(0.01)	0.36	**0.30**	0.096(0.09)	0.017(0.03)	**0.43**	**0.36**
Facial Landmarks	0.110(0.07)	0.017(0.020)	**0.39**	0.22	0.110(0.09)	0.020(0.04)	0.22	0.104
body HOG	0.120(0.08)	0.021(0.03)	0.19	0.11	0.118(0.10)	0.023(0.04)	0.08	0.05
body HOF	0.11(0.075)	0.017(0.02)	**0.39**	0.26	0.115(0.09)	0.021(0.04)	0.17	0.09
Decision-fusion	*0.100(0.08)*	*0.016(0.02)*	***0.44***	***0.33***	*0.089(0.10)*	*0.017(0.04)*	***0.45***	*0.33*
Subject-specific								
vQLZM	0.099(0.07)	0.014(0.02)	**0.60**	**0.45**	0.104(0.09)	0.018(0.03)	**0.42**	**0.33**
Facial Landmarks	0.110(0.07)	0.018(0.02)	0.41	0.26	0.120(0.08)	0.020(0.03)	0.24	0.14
body HOG	0.106(0.08)	0.017(0.03)	0.38	0.28	0.115(0.09)	0.022(0.04)	0.23	0.19
body HOF	0.110(0.08)	0.017(0.03)	0.39	0.29	0.117(0.09)	0.022(0.04)	0.20	0.15
Decision-fusion	*0.082(0.08)*	*0.013(0.03)*	*0.58*	***0.55***	*0.081(0.10)*	*0.016(0.04)*	***0.52***	***0.46***

Table 4. The regression results along arousal and valence dimensions with unimodal features, vQLZM-FV, Facial Landmarks FV, body HOG FV and body HOF FV respectively and the decision fusion results with these four different features.

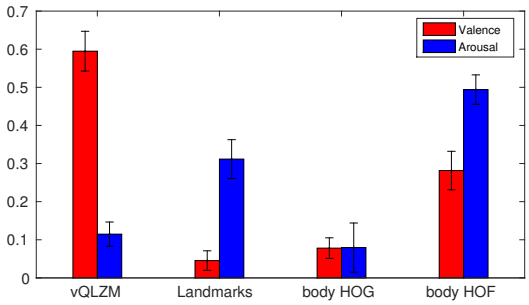

Figure 5. The average and standard deviations of weights assigned to different features in decision-level fusion with *leave-one-subject-out* cross-validation experiment.

	One subject out
Chance level	25.00%
vQLZM	18.96%
Facial Landmarks	13.07%
body HOG	36.20%
body HOF	**41.56%**
Decision-fusion	***43.75%***

Table 5. The group membership recognition results with unimodal features, vQLZM-FV, Facial Landmarks FV, body HOG FV and body HOF FV respectively and the decision fusion results with these four different features.

Table 6. Confusion matrix of the group classification results with *decision-fusion*.

	Group1	Group2	Group3	Group4
Group1	45	7	31	17
Group2	26	52	10	11
Group3	19	4	54	24
Group4	29	17	26	28

shared among different group members. This result is also consistent with what is reported in psychological studies -- people in a group often mirror one another's posture and behaviours [2, 19]. From Table 6, we can see that some groups (i.e., group 1, group 2 and group 3) are more easily to be recognized than others (i.e., group 4). This is possible because that group members in some groups share more information than those in the other groups.

5. Conclusion and Future Work

In this paper, we propose a novel framework for automatic emotion analysis of each individual in group settings. We first extract face and body static and dynamic information to analyse the affective states along both valence and arousal dimensions. For facial expression analysis, we introduce the (*vQLZM-FV*) descriptor to encode spatio-temporal information. We then propose a method to recognize the group membership of each individual by using their face and body behavioural cues. A set of experiments is conducted on a newly collected database. Overall,

we observe that first, the proposed *vQLZM-FV* representation outperforms other unimodal features and generates similar results to decision-level fusion for affect recognition; and second, group membership can be recognized using non-verbal behaviours, indicating that individuals influence each other's behaviours within a group.

Since the data, annotation and evaluation methods are different, most of the results published in the literature are not directly comparable with the results reported in our work. However, we attempt to compare our results with the methods having the most similar setup. For instance, Koelstra and Patras [18] also used a multimedia content evoked affect database (i.e., MAHNOB HCI [29]) to do binary emotion classification. They obtained 0.638 for arousal and 0.628 for valence in terms of $F1$ score by using face features. The results we obtained are 0.63 for arousal and 0.70 for valence in *leave-one-subject-out* setup by using the proposed face feature (i.e., *vQLZM-FV*). For affect regression, we provide a comparison with the 2015 Audio/Visual Emotion Challenge and Workshop (AV+EC) [24]. AV+EC 2015 used the spontaneous RECOLA database [25] that contains recordings of pairs of people in a remote collaborative work setting. They did emotion analysis along arousal and valence dimensions with *subject-independent* setup, and also used the same evaluation metric (i.e., CCC) as ours. In the winner paper [14], although the multimodal regression results were 0.824 for arousal and 0.688 for valence by combining audio, visual and physiological signals, the results obtained for unimodal regression using face appearance features were 0.587 for arousal and 0.346 for valence. The results that we obtained with the proposed *vQLZM-FV* face appearance feature are 0.30 for arousal and 0.36 for valence with *subject-independent* setup (i.e., *leave-one-subject-out* cross-validation).

Despite the promising results obtained in the experiments, analysis of the affect expressed in group videos is very challenging and needs to be further investigated in future work by taking advantage of other fusion techniques and extending the current work to group-level affect analysis.

Acknowledgements

The work of Wenxuan Mou is supported by CSC/Queen Mary joint PhD scholarship. The work of Hatice Gunes and Wenxuan Mou is partially funded by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (grant ref: EP/L00416X/1).

References

[1] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Trans. on Affective Computing*, 2015. 2

[2] S. G. Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 2002. 7

[3] S. G. Barsade and D. E. Gibson. Group affect its influence on individual and group outcomes. *Current Directions in Psychological Science*, 2012. 1

[4] 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. 4

[5] O. Celiktutan, F. Eyben, E. Sariyanidi, H. Gunes, and B. Schuller. Maptraits 2014: The first audio/visual mapping personality traits challenge. In *Proc. of the Workshop on Mapping Personality Traits Challenge and Workshop*. ACM, 2014. 3

[6] O. Celiktutan and H. Gunes. Continuous prediction of perceived traits and social dimensions in space and time. In *ICIP*, 2014. 4

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2011. 5

[8] A. Dhall and R. Goecke. A temporally piece-wise fisher vector approach for depression analysis. In *ACII*, 2015. 3

[9] A. Dhall, R. Goecke, and T. Gedeon. Automatic group happiness intensity analysis. *IEEE Trans. on Affective Computing*, 2015. 2

[10] 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 *FG*, 2015. 1, 2

[11] A. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009. 2

[12] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 2007. 2

[13] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *FG*, 2011. 2

[14] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proc. of the Int. Workshop on Audio/Visual Emotion Challenge*, 2015. 8

[15] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux. Depression estimation using audiovisual features and fisher vector encoding. In *Proc. Int. Workshop Audio/Visual Emotion Challenge*, 2014. 3

[16] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Trans. on Affective Computing*, 2013. 2, 5

[17] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. on Affective Computing*, 2012. 2

[18] S. Koelstra and I. Patras. Fusion of facial expressions and eeg for implicit affective tagging. *Image and Vision Computing*, 2013. 8

[19] M. LaFrance. Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. *Social Psychology Quarterly*, 1979. 7

[20] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati. Comparing models of disengagement in individual and group interactions. In *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2015. 1

[21] L.-P. Morency. The role of context in affective behavior understanding. *Social Emotions in Nature and Artifact*, 2013. 2

[22] W. Mou, O. Celiktutan, and H. Gunes. Group-level arousal and valence recognition in static images: Face, body and context. In *FG*, 2015. 1, 2

[23] 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. 2

[24] 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. 5, 8

[25] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *FG*, 2013. 8

[26] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013. 3

[27] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE TPAMI*, 2015. 2

[28] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro. Local zernike moment representation for facial affect recognition. In *BMVC*, 2013. 2, 3

[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. on Affective Computing*, 2012. 2, 8

[30] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *IEEE Trans. on Affective Computing*, 2012. 2

[31] A. Vlachostergiou, G. Caridakis, and S. Kollias. Context in affective multiparty and multimodal interaction: why, which, how and where? In *Proc. ACM Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, 2014. 2

[32] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 2013. 3, 4

[33] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 2

[34] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE TPAMI*, 2009. 1