Alone vs In-a-group: A Multi-modal Framework for Automatic Affect Recognition

WENXUAN MOU, Queen Mary University of London
HATICE GUNES, University of Cambridge
IOANNIS PATRAS, Queen Mary University of London

Recognition and analysis of human affect has been researched extensively within the field of computer science in the last two decades. However, most of the past research in automatic analysis of human affect has focused on the recognition of affect displayed by people in individual settings and little attention has been paid to the analysis of the affect expressed in group settings. In this paper, we first analyze the affect expressed by each individual in terms of arousal and valence dimensions in both individual and group videos and then propose methods to recognize the contextual information, i.e., whether a person is alone or in-a-group by analyzing their face and body behavioral cues. For affect analysis, we first devise affect recognition models separately in individual and group videos and then introduce a cross-condition affect recognition model that is trained by combining the two different types of data. We conduct a set of experiments on two datasets that contain both individual and group videos. Our experiments show that (1) the proposed Volume Quantized Local Zernike Moments Fisher Vector *(vQLZM-FV)* outperforms other unimodal features in affect analysis; (2) the temporal learning model, Long-Short Term Memory Networks (LSTM), works better than the static learning model, Support Vector Machine (SVM); (3) decision fusion helps to improve the affect recognition, indicating that body behaviors carry emotional information that is complementary rather than redundant to the emotion content in facial behaviors; and (4) it is possible to predict the context, i.e., whether a person is alone or in-a-group, using their non-verbal behavioral cues.

Additional Key Words and Phrases: Affect analysis, multimodal interaction, group settings, non-verbal behaviours, context analysis

ACM Reference format:
Wenxuan Mou, Hatice Gunes, and Ioannis Patras. . Alone vs In-a-group: A Multi-modal Framework for Automatic Affect Recognition.

1 INTRODUCTION

Affect analysis has attracted a lot of attention [50] in recent years. Automatic affect analysis aims to create a system capable of automatically interpreting, understanding and responding to emotions and moods displayed by humans. Building such systems are expected to advance Human-Computer Interaction (HCI) further.

Over the last decades, various methodologies have been proposed to automate the analysis of affect and emotions. However, the majority of the existing works focus on individual settings and little attention has been paid so far to the analysis in group settings, either at the the overall group-level emotion displayed by the entire group or at the individual-level emotion displayed by each individual within that group. From the psychological perspective, affect analysis in group settings is more complex than in individual settings due to the influence of the overall group as well as influences by each group member [2]. From the automatic analysis perspective, it has been shown that the degree of variation between individual and group settings is significant in terms of

differences in facial and bodily behaviors, timing and dynamics [41, 42]. To obtain further insights into this challenging problem, it is important to study the affect expressed in group settings. A few works focus on group-level affect analysis in static images in recent years [13, 14, 26, 40]. However, to the best of the our knowledge, except of our previous work [42], no works pays attention to individual-level affect analysis in group videos.

In this paper, we aim to investigate the following: (1) whether it is possible to recognize the affect expressed by each participant while presented with movie stimuli; (2) whether the affect recognition performance is affected by different settings or databases, i.e., individual vs group setting; (3) what kind of body and face features work better for different tasks; (4) whether the fusion of body and facial features is able to improve the recognition results; (5) whether it is possible to predict the context information (a person being alone or in-a-group) using facial and body behavioral cues. This work is an extended version of our previous works [41] and [42]. Differently from the aforementioned papers, the contributions of this work are:

(1) A novel framework for "individual vs. group" contextual information prediction is proposed. That is to recognize whether a person is alone or in-a-group using their face and body behavioral cues.
(2) The temporal modeling method, Long Short-Term Memory Networks (LSTM) combined with QLZM facial features, is utilized to analyze affect in terms of both arousal and valence dimensions.
(3) We conduct multiple experiments with both the individual and group datasets acquired in coherent setups for:
    (a) Affect classification / regression using both face and body behavioral cues.
    (b) Affect classification / regression using multi-modal fusion.
    (c) Affect classification / regression using both static learning model (i.e., SVM) and dynamic learning model (i.e., LSTM).
    (d) Context prediction using both face and body behavioral cues.

Specifically, in our previous work [41], we introduced a framework to analyze individual affect in individual and group videos along arousal and valence using facial features *only*. In [42], we proposed a method to recognize affect and group membership in group videos. For both of these works, we conducted experiments across small databases, e.g., only 576 samples in group videos. In [41], only one type of facial features was used by the classification model. In this paper, we investigate both affect analysis and contextual prediction in both individual and group videos using a multitude of face and body features combined with temporal learning. For affect analysis, on the one hand, we first train the affect recognition model separately in individual and group videos and then analyze how a combined model trained with data from two databases (i.e., individual and group videos) performs. On the other hand, we first utilize the Fisher Vector representation of the face and body features with a static learning model (Support Vector Machine) and then we use the static facial features with a temporal learning model, that is Long Short-term Memory Networks. Such a comparison for modeling affect in individual and group settings has never been conducted before. To this end, we extract different face and body features and train different affect recognition models using data from two different datasets, i.e., individualDB and groupDB (details of the databases are given in Section 3.1). For the prediction of contextual information, we propose an approach to predict whether a person is alone or in-a-group based on non-verbal visual cues - this has never been studied in previous works.

The remaining part of the paper is structured as follows: we review the previous works in Section 2; we state the proposed method in Section 3; we present and discuss the experimental

Table 1. Databases for individual affect analysis

Database	CK database [55]	JAFFE [36]	GEMEP [1]	SFEW [11]	AFEW [11]
Environment	Lab	Lab	Lab	Movie	Movie
Posed/acting/ spontaneous	Posed	Posed	Acting	Acting	Acting
Annotations	6 basic emotions	6 basic emotions + neutral	18 discrete emotions	6 basic emotions + neutral	6 basic emotions + neutral

6 basic emotions refer to "anger, disgust, fear, happiness, sadness, and surprise".

results and analysis in Section 4; and finally in Section 5 we conclude this work and discuss the future works.

2 RELATED WORKS

Automatic affect recognition has received a lot of attention in recent years with various applications in very diverse areas such as human-computer interaction [10], security [24], healthcare [30] and education [32]. Humans express their emotions through different channels: speech, facial expressions, head motion, body gestures, etc. In the affective computing field, various studies have been carried out to create systems that can recognize affective states by using multiple cues. Most of these works have been carried out in *individual settings*. However, in the real world, people are very often with others, interacting in group settings. More recently an increasing number of works have started focusing on affect analysis in *group settings* and there are challenges organized in this field since 2016 [14, 15]. The literature review below is divided into two parts, i.e., affect analysis in individual settings and group settings.

2.1 Affect Analysis in Individual Settings

The literature for analyzing a single person's emotion, and affective states is rich. We reviewed these works in terms of databases, modalities and methodologies for affect analysis in individual settings. Further details for automatic affect analysis in individual settings please refer to the recent survey studies [8, 50].

Databases. In the early days, most of the databases contain posed expressions, e.g., CK database [55] and JAFFE [36]. Currently, it is widely accepted that the recognition of posed expressions, even though is an interesting research problem, is not very relevant for real world settings. The expressions in the real life are far more diverse and complex. Hence, the focus has gradually shifted to automatic recognition of affect expressed in more naturalistic settings. For instance, AFEW and SFEW databases used in Emotion Recognition in the Wild (EmotiW) challenges are collected from movies [11], and images from the FER-2013 database were collected from the web [20]. In addition, researchers started using not only visual but also physiological signals. Therefore, some databases provide not only visual signals but physiological signals (from such as EEG and MCA modalities) for analysis, such as the Database for Emotion Analysis using Physiological Signals (DEAP) [33].

Features. Face is one of the most important channels of non-verbal communication. Facial expressions are prominent in researches for almost every aspect of emotion. In affective computing community, most of the research in vision-based emotion recognition has centered around facial expressions [7, 31, 32, 50]. However, humans naturally express emotions in a multi-modal way by means of language, vocal intonation, facial expression, body movements, postures etc [21]. Despite

Table 2. Representative works on affect analysis in group settings

	Dhall et al. [13, 16]	Dhall et al. [17]	Huang et al. [28]	Huang et al. [26]	Li et al. [35]	Tan [54]	Our work
Data Source	Web	Web	Web	Web	Web	Web	Recordings
Data Type	Static	Static	Static	Static	Static	Static	Dynamic
Samples	3134	504	3134	3134	3134	6471	7630
Labellers	4	3	4	4	4	-	3
Labels	6 stages of happiness	3 categories for valence	6 stages of happiness	6 stages of happiness	6 stages of happiness	3 categories for valence	Categorical & continuous
Settings	Group	Group	Group	Group	Group	Group	Individual & group
Features	Face & scene	Face & scene	Face & scene	Face, body & scene	Face & scene	Face & whole image	Face & body

the available range of cues and modalities used by humans, the mainstream on automatic emotion recognition has mostly focused on recognition of facial expressions in terms of the basic emotion categories (neutral, happiness, sadness, surprise, fear, anger and disgust). However, other cues, such as body movements and gestures, also play an important role in emotion expression and perception [32]. Facial expressions in combination with body behaviors have been successfully used to predict various emotional states in [40].

Methodologies. Affect recognition can utilize both traditional machine learning methods (e.g., Support Vector Machine) and deep learning methods. Many works on affect recognition are using traditional learning methods. For example, the methodologies that obtained top results [52] in EmotiW 2013 challenge utilized Support Vector Machine. Recently, deep learning has shown a good performance in the conventional computer vision problems, such as action recognition [43] and face detection [63]. Therefore, some works in emotion recognition have also started using deep neural networks. In recent EmotiW series of competitions [14, 15] and AVEC challenges [46, 56], most of the submitted works used deep neural networks. The winner of EmotiW challenge for video-based emotion recognition [18] used a CNN-RNN framework. The winner of AVEC'17 affect sub-challenge used different hand-crafted and deep learned features to predict arousal, valence and likability [6]. They also showed that the temporal learning model, Long Short-term Memory (LSTM), performs better than the non-temporal model SVM, especially in terms of arousal and valence prediction. As we are dealing with dynamic videos in this paper, we utilize LSTM for arousal and valence recognition.

2.2 Affect Analysis in Group Settings

In the early years of affect analysis, most of the works focused on individual settings. However, 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) [41, 42]. Therefore, in the past few years, a number of works have started paying attention to affect analysis in group settings [13, 17, 40, 41]. The representative works on affect analysis in group settings are listed in Table 2.

Databases. The first database for group emotion analysis, named as HAPPEI, was collected by Dhall et al. [16]. This database contains 4,886 images that are collected from Flickr using key words, such as "party + people" and "graduation + ceremony". Each image was labeled with a group-level happiness intensity, face level happiness intensity, occlusion intensity and pose by four human

annotators. The happiness intensity is categorized into six levels of happiness (0-5), i.e., neutral, small smile, large smile, small laugh, large laugh and thrilled. Then Dhall et al. [17] collected another database containing 504 images, GAFF database, which extended the HAPPEI database from positive affect only [13] to other emotion categories, i.e., "positive, neutral and negative". In EmotiW challenge 2017 [14], GAFF database was extended to contain 6,471 images named as GReco, which was labeled in the same way as GAFF. In a further step, Mou et al. [40] collected a dataset for group-level emotion analysis along both arousal and valence dimensions. Each image was annotated by 15 labelers and each labeler was asked to select one label from "low, medium, high" for arousal and one from "negative, neutral, positive" for valence, that best described the group-level emotion expressed by people in each image. However, to the best of our knowledge, there is no publicly available database that contains both individual and group data in the same setup.

Features. Analysis of the affect expressed by people in group settings is challenging due to the challenging situations that involve head and body pose variations. A number of works have already reported multi-modal emotion recognition in group settings [17, 27, 40, 42]. Compared to individual settings, a group setting may contain face, body and other contextual information, such as who the person is talking to and what the person is watching. Facial features are the most widely used cues for automatic affect analysis [22]. 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 [44], while appearance features can represent the texture of the face such as wrinkles and furrows [50]. More and more studies have shown that body features are as powerful as facial features and are complementary information for emotion recognition to facial features [32]. For example, De Gelder and colleagues found that body expressions provide more useful information than the facial expressions for discriminating between happiness and fear [57], and discriminating between anger and fear [37]. It is reported that when affective information conveyed by the different modalities, i.e., face and body, is incongruent, body information appears to be the prominent factor for the recognized emotion [57]. Face and body, as part of an integrated whole, both contribute to the recognition of human affect [23]. Inspired by such works, in this paper we utilize both facial and body information for affect analysis in terms of arousal and valence and report the results when using both the individual modalites and the results of their fusion at decision level.

Pioneering works have also shown that the emotion displayed by people heavily relies on context [58], e.g., whether the person is alone or staying with others and whether the person is in a meeting or in a party at the time. Therefore, in addition to using face and body information, the utilization and analysis of contextual features is getting increasingly popular for automatic affect analysis [39], especially in the case of group settings where there are multiple people inherently involved in more complex contextual situations than individual settings, not only in terms of each individual's information (e.g., the identity of an individual and location of the person) but also in terms of dynamics among group members (e.g., where the person is staying and how the others are feeling at the moment). [19] is a pioneering work using the contextual features based on the distribution of a group of people in an image to infer the individual age and gender. In [64], it is found that social context (i.e., alone or together with others) has an effect on the Quality of the Viewing Experience in terms of five aspects, namely enjoyment, endurability, satisfaction, involvement in the viewing experience and perceived visual quality. Contextual information based on the relative location and scale of the people in an image was used for group-level affect analysis in [40]. Context features / information on one hand can be used as a type of feature to analyze affect and other

social dimensions, on the other hand it can be combined with other features, such as facial and bodily expressions. In this paper, we predict the contextual information - whether a person is alone or in-a-group using non-verbal behavioral cues.

Methodologies. Automatic emotion recognition in group settings can be reviewed under two categories, group-level emotion analysis [17, 26, 40] and individual-level emotion analysis [42]. Psychological studies show that group members are influenced by each other [2]. Specifically, group emotion as a whole is influenced by emotions of the individuals within the group, and the emotions of the individuals are influenced by the emotion of the whole group. Therefore, group-level and individual-level affect analysis are both important for understanding the group dynamics. There are a few works focusing on group-level affect analysis in recent years. For instance, the first framework of group-level affect analysis was proposed by Dhall et al. [13], which aimed to infer the *"overall happiness mood intensities"* displayed by a group of people (i.e., no less than two people) in static images. Subsequently, Dhall et al. [17] introduced a framework to predict the collective valence levels of a group of people, "positive, neutral and negative". Meanwhile, another extended framework was proposed in [40] for recognizing the affect displayed by a group of people in static images along the arousal and valence dimensions. Each dimension was divided into 3 levels, arousal for "high, medium and low" and valence for "positive, neutral and negative". EmotiW [14, 15] has organized a group-based emotion recognition sub-challenge since 2016, which aimed at predicting affect displayed by a group of people in static images. However, none of the aforementioned works focus on individual-level affect analysis in group settings. Furthermore, all of the above works are limited to static images, while dynamic videos naturally include interactions and enable the use of temporal information which makes the recognition of human affect more insightful. In this paper, we focus on individual-level affect analysis in group videos.

A few works focus on group-level affect analysis in static images in recent years [13, 14, 26, 40]. However, to the best of our knowledge, except of our previous work [42], no works pay attention to individual-level affect analysis in group videos.

3 THE PROPOSED METHOD

This paper proposes a framework to recognize (1) the affect of individuals in different settings, i.e., individual and group videos and (2) the prediction of contextual information, i.e., whether a person is alone or in-a-group by using non-verbal behavioral cues, i.e., face and body cues. We illustrate the proposed framework in Fig. 1.

We first adopt an SVM-based multi-modal method using dynamic features and conduct experiments on both individual and group videos. To represent faces, we use geometric and appearance representations. The geometric feature we utilize is facial landmark trajectory, while appearance feature we use is the extended volume Quantized Local Zernike Moments (QLZM) [50, 51] extracted along facial landmark trajectories. In light of the body representations, we first extract dense trajectories and then we extract Histogram of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF) descriptors along each trajectory [60]. Before feeding the features to different classifiers and regressors, we encode the different face and body low-level descriptors into Fisher Vectors (FV). Multiple experiments are carried out for affect analysis using unimodal and multi-modal cues. Secondly, we train an temporal learning model, namely an LSTM, using static features for affect recognition. LSTM is one of the state-of-the-art sequence modeling approaches and has been successfully applied to affect analysis [6, 35].

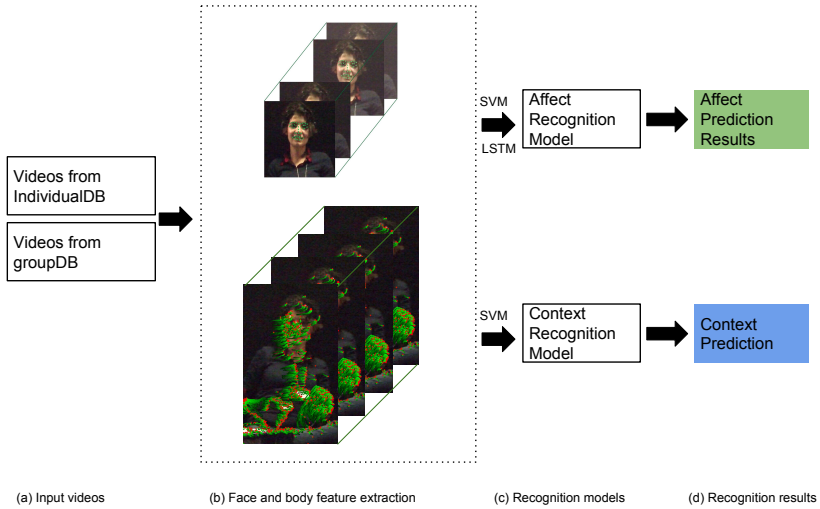


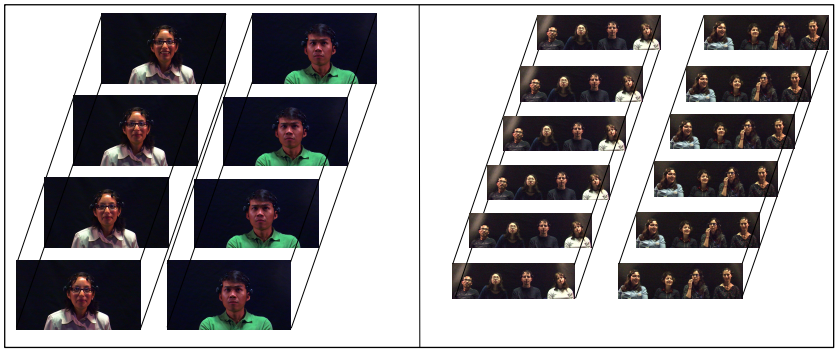
Fig. 1. Description of the proposed framework. (a) Input videos, videos from individualDB, and groupDB; (b) Feature extraction - face and body features extracted; (c) Different recognition models are trained, i.e., affect recognition models and contextual information recognition models.

Table 3. The stimuli of long movies / videos are presented with their sources (movie / video IDs are listed in parentheses and in the remaining part of the paper, the video IDs are used to refer to movies / videos) and the movie / video durations.

Movie / Video	Duration/min
Descent (N1)	23:30
Mr. Bean (P1)	18:38
Batman the Dark Knight (B1)	23:25
Up (U1)	14:01

3.1 Data and Annotation

3.1.1 ***Data Collection.*** Two recently collected databases that are part of the AMIGOS dataset [38] are used in this work, i.e., group database (groupDB) and individual database (individualDB). The main objective of the databases is to study the personality, mood and affective responses of people engaging with multimedia content in two social contexts, (i) when they are alone (individualDB), and (ii) when they are part of an audience (groupDB). During the recordings, the participants were asked to watch stimuli of different affective nature. In both databases, four long movie segments (14-24 mins) were used as movie stimuli, details of which are listed in Table 3. In groupDB, sixteen participants were recorded while they were watching different movies. These sixteen participants were arranged into four groups (i.e., four participants in each group) watching all of the four movies (the information of the four movies are presented in Table 3) together. In individualDB, seventeen participants which were different from the sixteen participants in groupDB, watched these four movies individually. Videos were recorded at 1280×720 resolution, 25fps. Representative frames from these two databases are shown in Fig. 2. In addition, in the recordings, implicit responses, namely, Electroencephalogram (EEG), Galvanic Skin Response (GSR), Electrocardiogram (ECG)



(a) IndividualDB (b) GroupDB

Fig. 2. Representative sequences from both the (a) individualDB and (b) groupDB.

Table 4. The results of the measurement of inter-labeler agreement on the annotations are reported in terms of arousal and valence dimensions among 3 labelers. Two methods of measurement are provided, i.e., Cronbach's α and Pearsonffs correlation coefficient (PCC).

Dimension	Arousal		Valence	
Methods	Cronbach's α	PCC	Cronbach's α	PCC
GroupDB	0.80	0.60	0.89	0.76
individualDB	0.75	0.63	0.88	0.75

and RGB-D full body videos were also recorded. For further details about the database we refer the reader to [38].

3.1.2 ***Data Annotation.*** The annotation was conducted by human labelers, three researchers who are focusing on 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.

In order to assess the inter-labeler agreement, Cronbach's α [9] and Pearson's correlation (PCC), that have been widely used in the literature for agreement assessment on continuous scale [3, 4, 47, 48], were computed. Mean Cronbach's α and PCC over all participants for both groupDB and individualDB along arousal and valence dimensions are listed in Table 4. From Table 4, we can see that the values of Chronbachfis α are all > 0.7, that is considered as an acceptable agreement level [4, 48]. In addition, we can see that high positive relationships among labelers, i.e., PCC > 0.6, which ensures inter-labeler agreement.

3.2 Face and Body Feature Extraction

3.2.1 ***Face Features***. Before extracting facial features, we first utilize Intraface [62] to detect facial landmarks of each face in the video. After applying Intraface, each face obtains 49 facial points. However, not all faces are detected due to illumination, occlusion, and pose variations in such a naturalistic scenario. In order to make the facial feature extraction consistent among all frames, when the face detection fails in a current frame, the position of the last detected face is used.

In terms of facial geometric features, let $X_t = [\ (x_t^1,\ y_t^1),\ (x_t^2,\ y_t^2)\ \ldots\ (x_t^n,\ y_t^n)\ ]$ denotes the position of n landmark points of the face at the current frame t. The number of landmark points on each face $n = 49$. x_t^k and y_t^k refer to the coordinates of the *k-th* landmark point at the current frame t. Then landmark points of the subsequent frames are concatenated to generate the facial landmark trajectories. In this way, the representation of the facial landmark trajectory encodes the motion patterns of the facial points as the body trajectories used in [60]. The *k-th* facial landmark point is described by a sequence $(\Delta X_t^k,\ \Delta X_{t+1}^k\ \ldots\ \Delta X_{t+L-1}^k)$ 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 obtained vector is then normalized by the sum of the displacement vector magnitudes:

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

Y^k is referred as *Facial Landmarks* in the remaining part of the paper. The length of the facial landmark trajectories is fixed as $L = 15$ frames based on [60]. In this way, a 30 ($30 = 2{\times}L$, where $L = 15$) dimensional feature is generated around each landmark point of the face. And for each face, the dimensionality of the descriptor is 49x30 as 49 landmark points are detected for each face.

After the geometric features, Quantised Local Zernike Moments (QLZM) [51] obtained from the local patch around each facial landmark point are extracted as the facial appearance representation. QLZM [51] originally designed for static images. However, as we are focusing on video information processing, temporal information is important. Therefore, it is extended to a volume representation to embed both spatial and temporal information, as described in Fig. 3. We refer the facial appearance feature as *vQLZM* in the remaining part of the paper. The size of the volume is $N \times N$ pixels, while the length is $L = 15$ frames, the same volume size with the *Facial Landmarks*. The volume is then subdivided into a spatio-temporal grid of size $n_\tau \times n_\tau \times n_\sigma$ to encode structure information. The QLZM descriptor is computed in each cell of the spatio-temporal grid. The final descriptor is generated by concatenating these descriptors of each cell. In our experiments, we set $N = 24$ that is the average of the distances between the centroids of two eyes from all of the detected faces across the whole dataset. Note, that as participants are relatively static, at very similar distance from the camera and their faces are at roughly equal sizes (standard deviation between the centroids of two eyes is very small, 2.1 pixels).

3.2.2 ***Body Features***. Body feature extraction is a type of person-based representation, therefore, the first step is to apply a person detector. Constrained by our experimental setups - a fixed number of people in the video (either one in individualDB or four in groupDB) and a static camera, we use an ad-hoc scheme that is to use only the central part where the person is in individualDB and equally divide the frame in four parts in groupDB. In order to avoid the overlap between the participants that are neighboring each other, we leave a space between every two neighbors. The space size is equal to the average size of the faces across all videos, i.e., 64. Then, dense trajectories [60] are extracted. Trajectories capture the local motion information of the video and dense

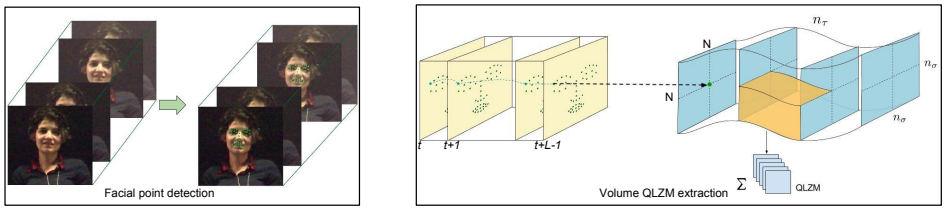


Fig. 3. Details of the approach to extract the facial appearance feature, vQLZM. The left figure shows the detection of facial landmark points. The right figure illustrates the tracking of facial landmark point over L frames. Appearance and motion information is extracted over a local neighborhood of $N \times N$ pixels along each landmark point. To encode the structure information, the local volume is subdivided into a spatio-temporal grid of size $n_\tau \times n_\sigma$. $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$ based on [60].

representation guarantees a good coverage of foreground motion as well as of the surrounding context. Subsequently, HOG and HOF features are obtained along each extracted trajectory. They are computed in the spatio-temporal volume aligned with the trajectories as shown in Fig. 4. HOG and HOF orientations are quantized into eight bins with full orientations. However, as an additional zero bin is added for HOF for pixels with optical flow magnitudes lower than the threshold (i.e., nine bins in total), the final representation size of HOG is 96 and that of HOF is 108 with the trajectory length $L = 15$ frames. We refer these two body related representations as *body HOG* and *body HOF* respectively in the rest of the paper.

The trajectory is extracted based on motion information using optical flow method. The step by step description of the extraction of the trajectories is given below:

(1) **Dense sampling.** Feature points are densely sampled on a grid spaced by $W = 5$ (obtained from experiments [60]) pixels.
(2) **Trajectories extraction.** For the current frame I_t, its dense optical flow field $w_t = (\mu_t, \nu_t)$ is computed with respect to the next frame I_{t+1}, where μ_t and ν_t refer to the horizontal and vertical components of the optical flow respectively. If we refer a point in the current frame I_t as $P_t = (x_t, y_t)$, the point P_{t+1} in the next frame I_{t+1} is smoothed by applying a median filter on w_t :

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)|_{(x_t, y_t)} \tag{2}$$

where M is the kernel of the median filter. Points in the subsequent frames are concatenated to form trajectories, i.e., $(P_t, P_{t+1}, P_{t+2}, P_{t+3}, ...)$.
(3) **Remove static points.** The static points are removed in the post-processing as they can not provide any motion related information.

3.3 Fisher Vector Encoding

Fisher Vector (FV) representation [49] has been widely utilized in traditional computer vision problems (e.g., action recognition [60, 61]) and affect analysis (e.g., depression analysis [12, 29]). The first work that applied Fisher Vector descriptors for the problem of action recognition in videos used local features extracted along dense trajectories [59]. The trajectories are extracted by defining a dense grid of points which are then tracked using optical flow that was estimated offline to include motion information in the pipeline. By encoding the extracted trajectory features with the Fisher Vector descriptor, this approach and its improved version [60, 61] achieved the state-of-the-art results for the action recognition before deep neural networks are widely utilized. It encodes

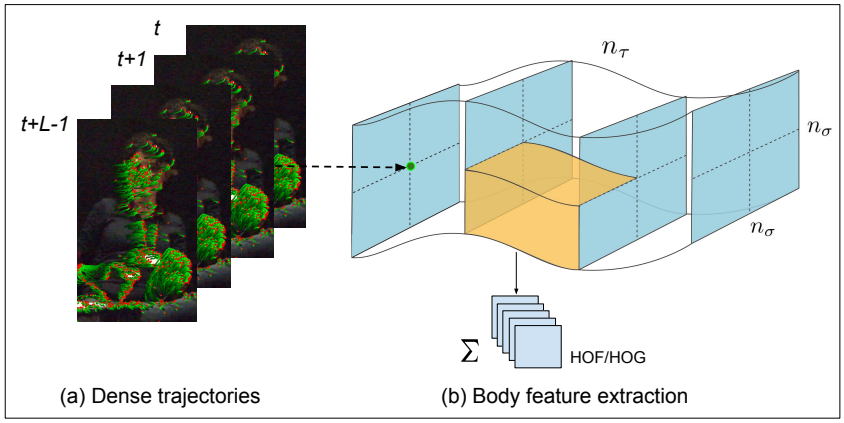


Fig. 4. Description of the method of body HOG/HOF feature extraction. (a) shows the detected dense trajectories. (b) illustrates the HOG/HOF feature extraction along the trajectories in the spatial scale over L frames. Motion information over a local neighborhood of $N \times N$ pixels along each trajectory point are extracted. In order to encode the structure information, the local volume is divided into small spatio-temporal grid of size $n_\tau \times n_\sigma$. Based on [60], $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$.

both the first and second order statistics between the low-level (local) video/image descriptors and a Gaussian Mixture Model (GMM). To obtain the Fisher Vector, firstly, Principal Component Analysis (PCA) is applied to the descriptors to decrease the dimensionality. Secondly, the low-level descriptors (i.e., face and body descriptors in our case) is fitted to a GMM . The covariance matrices for GMM are diagonal. As suggested by [60, 61], the number of Gaussians is set to $K = 256$ and randomly selected 256,000 descriptors are used to fit a GMM. The dimensionality of the Fisher Vector is $(2D + 1)K$ (D refers to the dimensionality of the descriptor before feeding to GMM, i.e., after applying PCA), which is used to represent one clip. Four different types of Fisher Vectors (FVs) are generated based on face and body features, namely, *Facial Landmarks, vQLZM, body HOG* and *body HOF*.

4 EXPERIMENTS AND ANALYSIS

The experiments are carried out using both individualDB and groupDB, two databases for studying affect analysis from multi-modal cues in different settings, i.e., individual settings and group settings respectively. We aim to analyze (1) the recognition of the affect expressed by each individual in individual and group settings; (2) whether it is possible to predict contextual information, i.e., whether a person is being alone or within a group while watching movie clips; and (3) how different face and body cues perform for different recognition tasks.

4.1 Experimental Details

4.1.1 ***Experimental Setup.*** For groupDB, group videos from four groups are used in the experiments, i.e., three groups (twelve subjects) with recordings of people watching four movies (N1, P1, B1 and U1) and one group (four subjects) with recordings of people watching three movies (B1, N1 and U1). In this case, we have data from sixteen subjects and fifteen sessions in total used in the experiments. One session refers to the recording of one group watching one movie. 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,

Table 5. The distribution of samples for individualDB and groupDB along arousal and valence dimensions after quantization

Dimensions	Arousal		Valence	
Labels	High	Low	Positive	Negative
GroupDB	1792	1792	1792	1792
individualDB	2023	2023	2023	2023

70 clips for B1, 56 clips for P1 and 42 clips for U1. As a result, the total number of clips we use in our experiments is $(70(B1) \times 4(4subjects) \times 4(4movies)) + (70(N1) \times 4(4subjects) \times 4(4groups)) + (56(P1) \times 4(4subjects) \times 3(3groups)) + (42(U1) \times 4(4subjects) \times 4(4groups)) = 3584$. In terms of individualDB, videos from 17 participants are used in the experiments. Each participant was recorded while watching 4 movies (N1, P1, B1 and U1). We also use 20-seconds clips. Therefore, the total number of clips we use in the experiments is $(70 + 70 + 56 + 42) \times 17 = 4046$. Classification and regression models are built with different cross-validation setups, such as *subject-specific* and *leave-one-subject-out*. The parameters of each model are optimized over the training-validation data. *Subject-specific* refers to train the model using *leave-one-sample-out* cross-validation for the data of each subject separately. Namely, in each fold, one sample from a certain subject is used as testing data and all the other samples from the same subject are used as training data. In order to avoid the subject-dependency problem caused by the *subject-specific* model, *leave-one-subject-out* cross-validation is also applied. *Leave-one-subject-out* means that we use one subject's data for testing and all other subjects' data for training-validation in each fold. For groupDB, *leave-one-group-out* cross-validation is also applied. *Leave-one-group-out* validation means that we use data from three groups out of four groups as training data, and data from the left one group as the testing data. For affect analysis, we did both classification and regression. Classification is formulated as a binary classification problem by quantizing both arousal and valence annotations into two classes using the median of all of the annotations as thresholds. In this way, arousal is quantized into *high* and *low* arousal and valence is quantized into *positive* and *negative* valence. The distribution of samples for groupDB and individualDB along both arousal and valence dimensions after quantization is shown in Table 5. For contextual information prediction, it is formulated as a binary classification problem. We conduct experiments to predict whether a person is being alone or in-a-group based on face and body behavioral cues using *leave-one-subject-out* cross-validation.

4.1.2 ***Classifier.*** In the first session of affect analysis, we conduct experiments using the same classifier as we did in our previous works [41, 42], i.e., Support Vector Machines (SVM) [5] for classification and Support Vector Regression (SVR) for regression, with all extracted face and body features. In addition, SVM is also used for context prediction. In the second step of affect analysis, we conduct experiments on affect analysis using Long Short-Term Memory (LSTM) Networks [25] with the best performing feature obtained from the first experiment. LSTM is a type of Recurrent Neural Network (RNN) and commonly used for the analysis of sequential signals. An LSTM unit is composed of a cell, an input gate, a forget gate and an output gate. The cell is responsible for memorizing the important information/features, while how much of the information in the cell can be passed depends on the input and forget gates, which enables the model to learn the long-term dependencies in our experiments. LSTM is trained using the frame-level raw features without the Fisher Vector representation and take each 20-seconds clip as a sequence as shown in Fig. 5.

4.1.3 ***Evaluation.*** The classification results of affect analysis are evaluated by the average of $F1$ scores (average of $F1$ scores for both classes). In terms of regression results of affect analysis, the

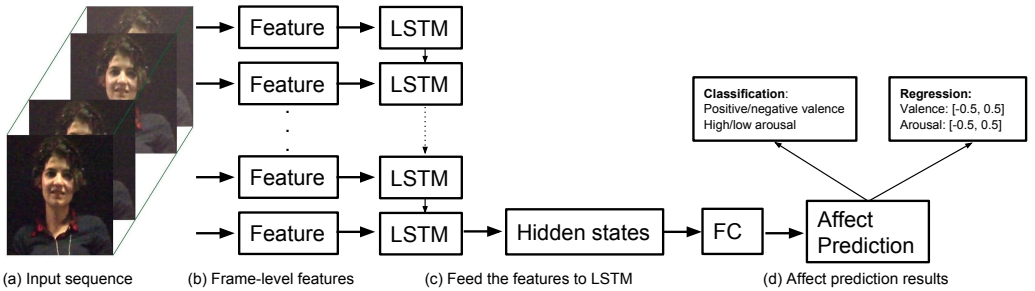


Fig. 5. Illustration of the approach for affect analysis using LSTM. (a) Input sequence, the 20-seconds clip. (b) Frame-level features are extracted. (c) Features extracted from every frame are fed into a one-layer LSTM with 128 hidden states. (d) Affect prediction results obtained for either classification or regression.

Mean Squared Error (MSE) is presented. In addition to the above measure, Pearson's Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC) are also reported. As illustrated in [47], CCC combines the PCC with the squared 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 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, are penalized by the deviation.

Affect analysis is divided into two parts, i.e., affect classification and regression along arousal and valence dimensions. The first part of the experiments is affect recognition that is conducted using (1) different unimodal cues and (2) decision-level fusion method. As we use SVM as the classifier, decision-fusion is applied on the soft outputs of the single-modality classifiers. We utilize the publicly available SVM libary, LibSVM [5] for training and testing. Before the face and body features are fed to any classifier or regressor, we first apply PCA to reduce the dimensionality by preserving 99% of the variance. The second part of the affect recognition is carried out on the best performed unimodal feature, QLZM, using LSTM implemented on Pytorch platform [45].

4.2 Results and Analysis

In this section, the affect recognition results are provided and discussed based on the two databases, individualDB and groupDB separately. In addition, affect recognition across different databases is also analyzed. Finally, the context recognition results are reported in terms of prediction of whether a person is alone or in-a-group.

4.2.1 ***Affect Recognition in individualDB.*** We utilize Linear Support Vector Machine (SVM) to do classification w.r.t. the dimensions of arousal (high arousal vs low arousal) and valence (positive valence vs negative valence). The classification results obtained using unimodal features and decision-level fusion are illustrated in Table 6. We can see that different types of features perform differently. Generally, *vQLZM-FV* shows the best performance in both *leave-one-subject-out* and *subject-specific* cross-validation. It indicates that the proposed *vQLZM-FV* descriptors encode the information of spatio-temporal textures and are informative for tasks of affect analysis. On the other hand, we can see that compared to *leave-one-subject-out* models, *subject-specific* models perform better due to the subject-dependency. We can see that compared to *QLZM Fisher Vectors*

Table 6. The affect classification results in terms of F1 score for **individualDB** using **SVM** with unimodal face and body features and the decision-level fusion. The statistical significance (p-value) is also presented in parentheses.

Dimensions	Arousal	Valence
	$F_1(p-value)$	$F_1(p-value)$
Chance level	0.5	0.5
Leave-one-subject-out		
vQLZM	**0.555** ($p < 0.05$)	**0.59** ($p < 0.05$)
Facial Landmarks	0.548 ($p < 0.05$)	0.57 ($p < 0.05$)
body HOG	0.549 ($p < 0.05$)	0.54 ($p < 0.05$)
body HOF	0.551 ($p < 0.05$)	0.52 ($p < 0.05$)
Decision-fusion of four features	***0.57*** ($p < 0.05$)	***0.60*** ($p < 0.05$)
Subject-specific		
vQLZM	**0.70** ($p < 0.01$)	**0.73** ($p < 0.01$)
Facial Landmarks	0.66 ($p < 0.01$)	0.66 ($p < 0.01$)
body HOG	**0.70** ($p < 0.01$)	0.69 ($p < 0.01$)
body HOF	0.66 ($p < 0.01$)	0.66 ($p < 0.01$)
Decision-fusion of four features	***0.72*** ($p < 0.01$)	***0.75*** ($p < 0.01$)

Table 7. The affect classification results in terms of F1 score for **individualDB** with **QLZM** features using **LSTM**. The statistical significance (p-value) is also presented in parentheses.

Dimensions	Arousal	Valence
	$F_1(p-value)$	$F_1(p-value)$
Chance level	0.5	0.5
Leave-one-subject-out IndividualDB	**0.60** ($p < 0.01$)	**0.61** ($p < 0.01$)

(vQLZM) with SVM and even *the decision-fusion with SVM*, LSTM is more powerful for arousal and valence recognition in dynamic videos. To show the results clearly, we compare the classification and regression results obtained with *QLZM Fisher Vectors (vQLZM) with SVM*, *decision-fusion with SVM* and *QLZM with LSTM* in Fig. 7 and 8. This is the first work to report affect analysis in group videos using temporal models. The results show that LSTM improves affect recognition performance significantly as has previously been reported in single-person videos in [6].

In terms of decision-level, the decision values, that is the obtained probabilities for all classes, from individual features are given as input to a linear-SVM. The results show that the classification performance using decision-fusion of four face and body features is most of the times equal to or better than that obtained with unimodal features. For example, the best affect classification results obtained using unimodal cues are 0.55 (0.70) in terms of arousal and 0.59 (0.73) in terms of valence using the $F1$ score as our evaluation method in *leave-one-subject-out* (*subject-specific*) setups; and those classification results of affect analysis obtained using decision fusion are 0.57 (0.72) in terms of arousal and 0.60 (0.75) in terms of valence. Therefore, the fusion of multiple cues, in general, improves the classification results albeit not by a large margin in comparison to the proposed vQLZM-FV descriptor.

Table 8. The affect regression results in terms of MSE, CC and CCC for **individualDB** using **SVR** with unimodal face and body features and the decision-level fusion.

Dimensions	Arousal			Valence		
	MSE(std)	CC	CCC	MSE(std)	CC	CCC
Leave-one-subject-out						
vQLZM	0.08(0.01)	**0.34**	**0.29**	0.08(0.01)	**0.34**	**0.33**
Facial Landmarks	0.01(0.01)	0.29	0.15	0.05(0.01)	0.25	0.19
body HOG	0.07(0.01)	0.27	0.23	0.06(0.01)	0.13	0.12
body HOF	0.01(0.01)	0.30	0.18	0.06(0.05))	0.26	0.21
Decision-fusion of four features	*0.07 (0.01)*	***0.44***	***0.34***	*0.04(0.01)*	***0.47***	***0.32***
Subject-specific						
vQLZM	0.04(0.01)	**0.76**	**0.62**	0.03(0.01)	**0.69**	**0.60**
Facial Landmarks	0.06(0.01)	0.59	0.39	0.04(0.01)	0.48	0.36
body HOG	0.05(0.01)	0.66	0.53	.0.03(0.01)	0.58	0.52
body HOF	0.06(0.01)	0.59	0.40	0.04(0.01)	0.46	0.35
Decision-fusion of four features	*0.04(0.01)*	***0.75***	***0.66***	*0.02(0.01)*	***0.69***	***0.67***

Table 9. The affect regression results in terms of MSE, CC and CCC for **individualDB** with QLZM features using **LSTM**.

Dimensions	Arousal			Valence		
	MSE(std)	CC	CCC	MSE(std)	CC	CCC
Leave-one-subject-out IndividualDB	0.006(0.003)	**0.60**	**0.59**	0.004(0.002)	**0.62**	**0.61**

For the regression of the affect analysis, we utilize Support Vector Regression (SVR) with a radial basis function (RBF) kernel. The results obtained with unimodal and multi-modal features are presented in Table 8. For the unimodal results, we can see that the regression results are quite similar to the classification ones, i.e., *vQLZM-FV* generally performs best among all unimodal features. As to the decision-level fusion, we proceed in a similar way to the fusion in affect classification. Specifically, we fuse the ratings predicted from unimodal features in an RBF-SVR. The results show that using only the proposed *vQLZM-FV* feature can achieve results that are very close to those obtained by using multi-modal fusion.

Subsequently, we utilize LSTM and facial QLZM feature for affect classification and regression. LSTM is one of the state-of-the-art temporal modeling methods and facial QLZM feature is the best performed unimodal representation as shown in Table 6 and 8. The classification and regression results are reported in Table 7 and 9. We can see that compared to *QLZM Fisher Vectors (vQLZM) with SVM* and even *the decision-fusion with SVM*, LSTM is more powerful for arousal and valence recognition in dynamic videos. To show the results clearly, we compare the classification and regression results obtained with *QLZM Fisher Vectors (vQLZM) with SVM*, *decision-fusion with SVM* and *QLZM with LSTM* in Fig. 7 and 8. This is the first work to report affect analysis in group videos using temporal models. The results show that LSTM improves affect recognition performance significantly as has previously been reported in single-person videos in [6].

4.2.2 ***Affect Recognition in GroupDB.*** Similar to affect recognition in individualDB, Support Vector Machine (SVM) is utilized to do classification and regression w.r.t. the dimensions along arousal and valence. The classification and regression results using four different unimodal features and decision-level fusion are illustrated in Table 10 and Table 12 respectively. It can be seen

that the results are consistent with the results obtained using individualDB: (1) different features provide different classification/regression results and *vQLZM-FV* generally outperforms the other unimodal features in both classification and regression models; and (2) the fusion results are either equal to or better than those obtained with unimodal features, which indicates that the fusion of different features is generally helpful for improving the affect recognition results. On the other hand, compared to the individual settings, affect recognition in group settings performs better using the *leave-one-subject-out* evaluation criteria. The differences between subject-independent conditions (*leave-one-subject-out* cross-validation) and subject-dependent conditions (*subject-specific* cross-validation) in group settings are less pronounced than individual settings. A possible explanation is that for *leave-one-subject-out* experiments, for groupDB, although the subject is left out, there are members from the same group in the training data, which provides some useful information. More specifically, compared to participants from different groups, members in the same group display more similar emotions and present more similar behaviors. In order to test this, we conduct *leave-one-group-out* cross-validation. From Table 10 and 12, we can see that the results obtained with *leave-one-group-out* are not as good as the ones obtained with *leave-one-subject-out*. For each individual in a group, other members' behavioral cues can provide useful information for predicting the affect of that individual, which indicates that (1) group members show similar behaviors and share some common information; and (2) people may behave distinctively when they are in individual settings and when in group settings. For a further analysis of our first hypothesis, we compare the affect recognition results of subjects within the same group and from different groups. Fig. 6 shows the variance of the regression results among subjects in terms of valence dimension obtained using *QLZM Fisher Vectors (vQLZM) and SVR* under *leave-one-subject-out* cross-validation setup along video/movie U1. The red line and blue line refer to two distinct groups respectively while the black line represents results for these two groups. We can see that the variance within the same group tends to be smaller than that of across different groups. This shows that subjects within the same group display more similar affective states than subjects across different groups. In addition, in order to investigate whether people behave differently in different settings, in Section 4.2.4 we propose a method that attempts to predict whether a person is alone or in-a-group using their non-verbal behaviors.

Similar to the individual settings, we then utilize LSTM and facial QLZM feature for affect recognition in group settings. The classification and regression results are presented in Table 11 and 13 respectively. The comparison of the results obtained with non-temporal models (i.e., SVM and SVR) and temporal models (i.e., LSTM) is presented in Fig. 7 and 8. It can be clearly seen that the temporal modeling method, LSTM, outperforms the non-temporal model in terms of both arousal and valence recognition.

4.2.3 ***Cross-condition Affect Recognition.*** For the cross-condition affect recognition, combined models are trained using two databases, i.e., *combined model* trained with individualDB and groupDB. In the experiments, *leave-one-subject-out* cross-validation is applied and the experimental results are illustrated in Table 14. From the results, it can be seen that compared to Table 10, the results obtained with the *combined model* are slightly worse. Note that, the combined models are trained with data from both databases, only excluding the participant used as the test subject at each round. Therefore, it is trained with more data than models trained on each database separately. However, more training data does not always provide a better recognition model. A possible explanation is that the combined models have to make a compromise when modelling different types of data simultaneously, which results in decreased performance. However, compared to Table 6, the results obtained with the *combined model* are slightly better. A possible explanation is that

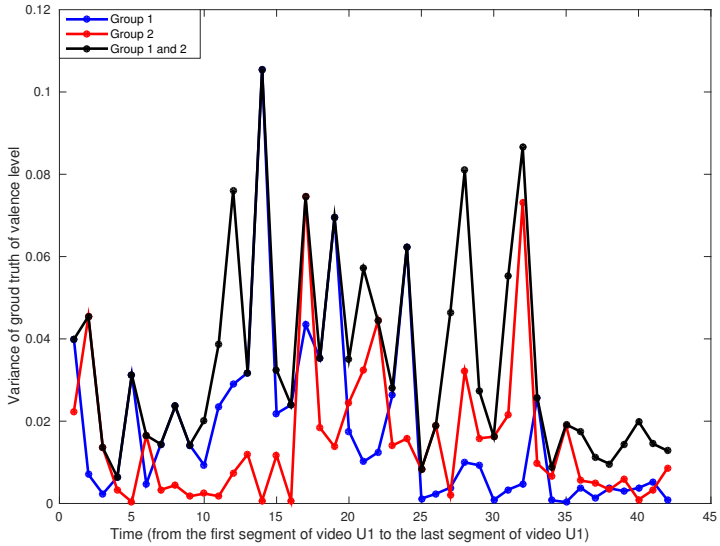


Fig. 6. The red line is variance of the four subjects in terms of the ground truth of valence dimension in group 1 and the blue one is for that of group 2. The black line is for group 1 and group 2.

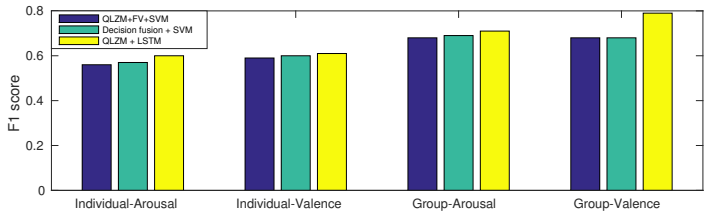


Fig. 7. Illustration of the affect classification results in terms of F1 score using different features and classifiers for individual and group settings along arousal and valence dimensions.

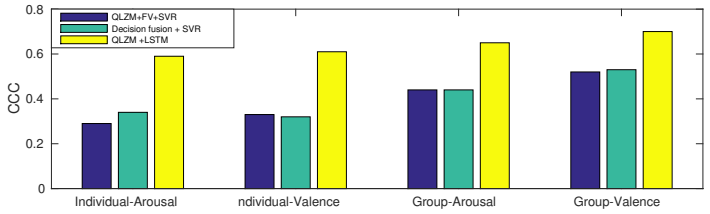


Fig. 8. Illustration of the affect regression results in terms of CCC using different features and regression methods for individual and group settings along arousal and valence dimensions.

people in group settings show more diverse social behaviors than when they are being alone, which helps improve the recognition results for individualDB when tested using the *combined model*.

Table 10. The affect classification results in terms of F1 score for **groupDB** using **SVM** with unimodal face and body features and the decision-level fusion. The statistical significant test (p-value) is also presented.

Dimensions	Arousal	Valence
	$F_1(p-value)$	$F_1(p-value)$
Chance level	0.5	0.5
Leave-one-subject-out		
vQLZM	**0.68** ($p < 0.01$)	**0.68** ($p < 0.01$)
Facial Landmarks	0.61 ($p < 0.01$)	0.58 ($p < 0.01$)
body HOG	0.57 ($p < 0.05$)	0.58 ($p < 0.01$)
body HOF	0.61 ($p < 0.01$)	0.59 ($p < 0.01$)
Decision-fusion of four features	***0.69*** ($p < 0.01$)	***0.68*** ($p < 0.01$)
Leave-one-group-out		
vQLZM	**0.67** ($p < 0.01$)	**0.64** ($p < 0.01$)
Facial Landmarks	0.61 ($p < 0.01$)	0.59 ($p < 0.01$)
body HOG	0.55 ($p < 0.05$)	0.58 ($p < 0.01$)
body HOF	0.61 ($p < 0.01$)	0.57 ($p < 0.01$)
Decision-fusion of four features	***0.68*** ($p < 0.01$)	***0.65*** ($p < 0.01$)
Subject-specific		
vQLZM	**0.80** ($p < 0.01$)	**0.80** ($p < 0.01$)
Facial Landmarks	0.69 ($p < 0.01$)	0.64 ($p < 0.01$)
body HOG	0.70 ($p < 0.01$)	0.68 ($p < 0.01$)
body HOF	0.70 ($p < 0.01$)	0.67 ($p < 0.01$)
Decision-fusion of four features	***0.80*** ($p < 0.01$)	***0.80*** ($p < 0.01$)

Table 11. The affect classification results in terms of F1 score for **GroupDB** with **QLZM** features using **LSTM**. The statistical significance (p-value) is also presented in parentheses.

Dimensions	Arousal	Valence
	$F_1(p-value)$	$F_1(p-value)$
Chance level	0.5	0.5
Leave-one-subject-out GroupDB	**0.71** ($p < 0.01$)	**0.79** ($p < 0.01$)

4.2.4 ***Contextual Information Recognition.*** In a further step, we investigate contextual information prediction using non-verbal behavioral cues. We conduct experiments to recognize whether a person is alone or in-a-group using the extracted face and body features described in Section 3.2. The results are shown in Table 15. We can see that the results we obtained, all above 85%, are significantly better than the chance level of 50%. In addition, it can be seen that body features perform slightly better than face features. It is possibly due to the fact that it is relatively difficult to utilize the facial information in this case as facial information is more subtle than body motion and gestures. Similarly to the affect recognition results shown in Table 6, 8, 10 and 12, fusion of different features again helps improve the performance. Predicting whether a person is alone or in-a-group successfully indicates that people behave distinctly while they are alone compared to being within a group.

Table 12. The affect regression results in terms of MSE, CC and CCC for **groupDB** using **SVR** with unimodal face and body features and the decision-level fusion.

Dimensions	Arousal			Valence		
	MSE(std)	CC	CCC	MSE(std)	CC	CCC
Leave-one-subject-out						
vQLZM	0.02(0.02)	**0.58**	**0.44**	0.01(0.01)	**0.57**	**0.52**
Facial Landmarks	0.02 (0.02)	0.44	0.23	0.01(0.01)	0.39	0.25
body HOG	0.02(0.03)	0.27	0.21	0.011(0.02)	0.29	0.26
body HOF	0.02(0.02)	0.42	0.27	0.01(0.01)	0.31	0.25
Decision-fusion of four features	*0.08(0.02)*	***0.61***	***0.44***	*0.03(0.01)*	***0.58***	***0.53***
Leave-one-group-out						
vQLZM	0.12(0.02)	**0.55**	**0.40**	0.04(0.01)	**0.56**	**0.52**
Facial Landmarks	0.13(0.02)	0.42	0.20	0.05(0.01)	0.36	0.24
body HOG	0.18(0.033)	0.18	0.13	0.07(0.02)	0.24	0.21
body HOF	0.14(0.02)	0.35	0.21	0.06(0.02)	0.28	0.23
Decision-fusion of four features	*0.03(0.02)*	***0.55***	***0.38***	*0.03(0.01)*	***0.65***	***0.55***
Subject-specific						
vQLZM	0.09(0.01)	**0.71**	**0.54**	0.03(0.01)	**0.67**	**0.56**
Facial Landmarks	0.12(0.02)	0.54	0.30	0.05(0.01)	0.46	0.32
body HOG	0.11(0.02)	0.58	0.45	0.04(0.01)	0.56	0.47
body HOF	0.12(0.02)	0.53	0.32	0.05(0.01)	0.48	0.35
Decision-fusion of four features	*0.08(0.02)*	***0.70***	***0.57***	*0.03(0.01)*	***0.69***	***0.62***

Table 13. The affect regression results in terms of MSE, CC and CCC for **GroupDB** with **QLZM** features using **LSTM**.

Dimensions	Arousal			Valence		
	MSE(std)	CC	CCC	MSE(std)	CC	CCC
Leave-one-subject-out groupDB	0.008(0.005)	**0.66**	**0.65**	0.004(0.002)	**0.72**	**0.70**

4.2.5 ***Related works.*** As the data, annotation and evaluation methods utilized in this work are different from existing works, it is difficult to directly compare the results with the published works in the literature. However, for reference, we briefly report here results obtained in other works using similar approaches and setups. For instance, in [34] affect analysis was formulated as a binary classification problem and results on MAHNOB HCI [53], focusing on emotions evoked by the presentation of multimedia content were reported – the $F1$ score by using facial features was 0.638 for arousal and 0.628 for valence. The results we achieved in the *leave-one-subject-out* setup by using LSTM and QLZM features are 0.71 for arousal and 0.79 for valence respectively. So far as affect regression is concerned, we report the results obained in the 2017 Audio-Visual Emotion Challenge (AVEC 2017) - affect sub-challenge, which aimed at affect analysis using the Sentiment Analysis in the Wild (SEWA) database collected 'in-the-wild'. This challenge used a *subject-independent* setup and the same evaluation metric (i.e., CCC) with us. The multi-modal

Table 14. Classification results obtained with ***Combined Model*** trained with **individualDB** and **groupDB** on both individualDB and groupDB in terms of $F1$ score using *leave-one-subject-out* cross-validation.

Models	**Combined Model**	
Test Data	**GroupDB**	
Dimensions	Arousal	Valence
Chance level	0.5	0.5
Leave-one-subject-out		
vQLZM	**0.67** ($p < 0.01$)	**0.67** ($p < 0.01$)
Landmarks	0.54 ($p < 0.05$)	0.57 ($p < 0.05$)
HOG	0.57 ($p < 0.05$)	0.59 ($p < 0.05$)
HOF	0.60 ($p < 0.01$)	0.59 ($p < 0.01$)
Models	**Combined Model**	
Test Data	**individualDB**	
Dimensions	Arousal	Valence
Chance level	0.5	0.5
Leave-one-subject-out		
vQLZM	**0.58** ($p < 0.05$)	**0.59** ($p < 0.05$)
Landmarks	0.55 ($p < 0.05$)	0.57 ($p < 0.05$)
HOG	0.55 ($p < 0.05$)	0.55 ($p < 0.05$)
HOF	0.57 ($p < 0.01$)	0.52 ($p < 0.05$)

Table 15. The contextual recognition (whether a person is alone or in-a-group) results obtained with unimodal face and body features, and the decision-level fusion. The table reports the average recognition accuracy over all subjects. The statistical significance (p-value) is also presented in parentheses. The chance level (50%) is also provided.

	Leave-one-subject-out
Chance level	50%
vQLZM	85% ($p < 0.01$)
Facial Landmarks	90% ($p < 0.01$)
body HOG	93% ($p < 0.01$)
body HOF	91% ($p < 0.01$)
Decision-fusion	***94%*** ($p < 0.01$)

baseline results obtained with multi-modal audio-visual-text features is 0.306 in terms of arousal and 0.466 in terms of valence (using CCC as evaluation method). The best results obtained by the winner paper [46] using unimodal features - facial appearance features were 0.67 for arousal and 0.70 for valence. The results that we obtained with LSTM and QLZM facial appearance feature are 0.65 in terms of arousal and 0.70 in terms of valence using the *subject-independent* setup.

5 CONCLUSIONS AND FUTURE WORKS

A novel framework is introduced in this paper for automatic context recognition and affect recognition in different settings - individual settings (i.e., individualDB) and group settings (i.e., groupDB). Face and body features are first extracted to analyze the affect states in terms of valence and arousal dimensions. In order to use the temporal information, we use two different methods: (1) from feature perspective, to represent facial information in spatio-temporal domain, we introduce a novel

volume based *vQLZM-FV* descriptor; (2) in terms of the learning model, we utilize the temporal modeling method, LSTM. We then propose a method to recognize contextual information, namely whether a person is alone or in-a-group by using their non-verbal behavioral features. A set of experiments is carried out on a database containing both individual and group videos. Firstly, we find that the *vQLZM-FV* descriptor achieves the best performance among all the unimodal face and body features, and generates similar results to decision-level fusion for affect recognition in both databases. Secondly, we find that the temporal learning model outperforms the non-temporal model in terms of affect recognition. Finally, the contextual information of being alone or in-a-group can be successfully recognized using facial and body cues.

Even though the promising results are obtained in our experiments, affect analysis is still a challenging problem, especially when it comes to different settings and needs to be investigated in a further step in future work by taking advantage of other machine learning and deep learning techniques. The current work can be extended by combining information from the group members, and it can also be extended to group-level affect analysis instead of affect recognition of each individual.

6 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). This work is also supported by Nvidia corporation with the donation of a TitanX GPU.

REFERENCES

[1] Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer. 2012. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion* (2012).

[2] Sigal G Barsade and Donald E Gibson. 2012. Group affect its influence on individual and group outcomes. *Current Directions in Psychological Science* (2012).

[3] Oya Celiktutan and Hatice Gunes. 2014. Continuous prediction of perceived traits and social dimensions in space and time. In *Proc. of Int. Conf. on Image Processing (ICIP)*.

[4] Oya Celiktutan and Hatice Gunes. 2017. Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. *IEEE Transactions on Affective Computing* (2017).

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology* (2011).

[6] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*.

[7] Jeffrey F Cohn and Fernando De la Torre. 2014. Automated face analysis for affective. *The Oxford handbook of affective computing* (2014).

[8] Ciprian A Corneanu, Marc Oliu, Jeffrey F Cohn, and Sergio Escalera. 2016. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* (2016).

[9] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* (1951).

[10] Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Trans. of the Royal Society B: Biological Sciences* (2007).

[11] Abhinav Dhall and others. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia* (2012).

[12] Abhinav Dhall and Roland Goecke. 2015. A temporally piece-wise fisher vector approach for depression analysis. In *Proc. of Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*.

[13] Abhinav Dhall, Roland Goecke, and Tom Gedeon. 2015. Automatic group happiness intensity analysis. *IEEE Trans. on Affective Computing* (2015).

[14] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From individual to group-level emotion recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.

[15] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. EmotiW 2016: video and group-level emotion recognition challenges. In *Proc. of ACM Int. Conf. Multimodal Interaction (ICMI)*.

[16] Abhinav Dhall, Jyoti Joshi, Ibrahim Radwan, and Roland Goecke. 2012. Finding happiest moments in a social context. In *Proc. of Int. Conf. on*.

[17] Abhinav Dhall, Jyoti Joshi, Karan Sikka, Roland Goecke, and Nicu Sebe. 2015. 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)*.

[18] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proc. of ACM Int. Conf. Multimodal Interaction (ICMI)*.

[19] Andrew Gallagher and Tsuhan Chen. 2009. Understanding images of groups of people. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[20] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, and Dong Hyun Lee. 2013. Challenges in representation learning: A report on three machine learning contests. *Neural Networks* (2013).

[21] Hatice Gunes. 2010. Automatic, dimensional and continuous emotion recognition. (2010).

[22] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*.

[23] Hatice Gunes, Caifeng Shan, Shizhi Chen, and YingLi Tian. 2015. Bodily expression for automatic affect recognition. *Emotion recognition: A pattern analysis approach* (2015).

[24] Javier Hernandez, Mohammed Hoque, Will Drevo, and Rosalind W. Picard. 2012. Mood meter: counting smiles in the wild. *Association for Computing Machinery* (2012).

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[26] Xiaohua Huang, Abhinav Dhall, Roland Goecke, Matti Pietikainen, and Guoying Zhao. 2018. Multi-modal Framework for Analyzing the Affect of a Group of People. *IEEE Transactions on Multimedia* (2018).

[27] Xiaohua Huang, Abhinav Dhall, Xin Liu, Guoying Zhao, Jingang Shi, Roland Goecke, and Matti Pietikainen. 2016. Analyzing the Affect of a Group of People Using Multi-modal Framework. *arXiv preprint arXiv:1610.03640* (2016).

[28] Xiaohua Huang, Abhinav Dhall, Guoying Zhao, Roland Goecke, and Matti Pietikäinen. 2015. Riesz-based Volume Local Binary Pattern and A Novel Group Expression Model for Group Happiness Intensity Analysis.. In *Proc. of Brithish Machine and Vision Conference (BMVC)*.

[29] Varun Jain, James L Crowley, Anind K Dey, and Augustin Lux. 2014. Depression estimation using audiovisual features and fisher vector encoding. In *Proc. Int. Workshop Audio/Visual Emotion Challenge*.

[30] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. 2012. Continuous pain intensity estimation from facial expressions. In *International Symposium on Visual Computing*.

[31] Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. 2013. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing* (2013).

[32] Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Trans. on Affective Computing* (2013).

[33] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. on Affective Computing* (2012).

[34] Sander Koelstra and Ioannis Patras. 2013. Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing* (2013).

[35] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim. 2016. Happiness Level Prediction with Sequential Inputs via Multiple Regressions. In *Proc. of ACM Int. Conf. Multimodal Interaction (ICMI)*.

[36] Michael J. Lyons, J. Budynek, and S. Akamatsu. 1999. Automatic classification of single facial images. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* (1999).

[37] Hanneke KM Meeren, Corné CRJ van Heijnsbergen, and Beatrice de Gelder. 2005. Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences* 102, 45 (2005), 16518–16523.

[38] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2017. AMIGOS: A dataset for Mood, personality and affect research on Individuals and GrOupS. *arXiv preprint arXiv:1702.02510* (2017).

[39] Louis-Philippe Morency. 2013. The Role of Context in Affective Behavior Understanding. *Social Emotions in Nature and Artifact* (2013).

[40] Wenxuan Mou, Oya Celiktutan, and Hatice Gunes. 2015. 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)*.

[41] Wenxuan Mou, Hatice Gunes, and Ioannis Patras. 2016. Alone versus In-a-group: A Comparative Analysis of Facial Affect Recognition. In *Proc. of ACM Int. Conf. on Multimedia*.

[42] Wenxuan Mou, Hatice Gunes, and Ioannis Patras. 2016. Automatic Recognition of Emotions and Membership in Group Videos. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition and Workshops (CVPRW)*.

[43] Petar Palasek and Ioannis Patras. 2016. Action Recognition Using Convolutional Restricted Boltzmann Machines. In *Proc. of Int. Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction*.

[44] Maja Pantic and Ioannis Patras. 2006. 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).

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[46] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*.

[47] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. AV+ EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proc. Int. Workshop Audio/Visual Emotion Challenge*.

[48] Fabien Ringeval, Andreas Sonderegger, Jens Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*.

[49] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. 2013. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision (IJCV)* (2013).

[50] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* (2015).

[51] Evangelos Sariyanidi, Hatice Gunes, Muhittin Gökmen, and Andrea Cavallaro. 2013. Local Zernike Moment Representation for Facial Affect Recognition. In *Proc. of Brithish Machine and Vision Conference (BMVC)*.

[52] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. 2013. Multiple kernel learning for emotion recognition in the wild. In *Proc. of ACM Int. Conf. Multimodal Interaction (ICMI)*.

[53] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. on Affective Computing* (2012).

[54] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2017. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.

[55] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* (2001).

[56] Michel Valstar, Jonathan Gratch, Bjrn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proc. of ACM Int. Conf. on Multimedia and Workshop on Audio/visual Emotion Challenge*.

[57] Jan Van den Stock, Ruthger Righart, and Beatrice De Gelder. 2007. Body expressions influence recognition of emotions in the face and voice. *Emotion* 7, 3 (2007), 487.

[58] Aggeliki Vlachostergiou, George Caridakis, and Stefanos Kollias. 2014. Context in affective multiparty and multimodal interaction: why, which, how and where?. In *Proc. ACM Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*.

[59] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[60] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (IJCV)* (2013).

[61] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proc. of the IEEE international conference on computer vision (ICCV)*.

[62] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[63] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, and Peter Robinson. 2015. Pose-Invariant 3D Face Alignment. *Proc. of Brithish Machine and Vision Conference (BMVC)* (2015).

[64] Yi Zhu, Ingrid Heynderickx, and Judith A Redi. 2014. Alone or together: measuring users' viewing experience in different social contexts. In *Human Vision and Electronic Imaging XIX*.