Alone versus In-a-group: A Comparative Analysis of Facial Affect Recognition

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. Recent works have started moving from individual to group scenarios. However, little attention has been paid to comparing the affect expressed in individual and group settings. This paper presents a framework to investigate the differences in affect recognition models along arousal and valence dimensions in individual and group settings. We analyse how a model trained on data collected from an individual setting performs on test data collected from a group setting, and *vice versa*. A third model combining data from both individual and group settings is also investigated. A set of experiments is conducted to predict the affective states along both arousal and valence dimensions on two newly collected databases that contain sixteen participants watching affective movie stimuli in individual and group settings, respectively. The experimental results show that (1) the affect model trained with group data performs better on individual test data than the model trained with individual data tested on group data, indicating that facial behaviours expressed in a group setting capture more variation than in an individual setting; and (2) the combined model does not show better performance than the affect model trained with a specific type of data (i.e., individual or group), but proves a good compromise. These results indicate that in settings where multiple affect models trained with different types of data are not available, using the affect model trained with group data is a viable solution.

Keywords

Affect recognition; arousal and valence recognition; individual affect recognition; group settings

1. INTRODUCTION

Automatic affect analysis has attracted increasing attention and has seen much progress in recent years [5, 6, 11, 17, 13]. Recent works in affective content analysis and affect recognition fields have started focusing on the analysis of naturalistic affect displayed and collected in more diverse and complex scenarios, such as dyadic interactions [1] and a group of people in a scene or involved in an interaction [8, 4]. Human behaviours are largely dependent on social context [7, 16]. Specifically, the way humans behave alone is different from the way humans behave in a group setting. Consequently, we hypothesize that being alone versus being in-a-group setting will affect the performance and the effectiveness of the automatic analysers that heavily depend on the type of data utilised for training.

Pioneering works in this direction have recently emerged in disengagement analysis. In [7], the individual disengagement was studied in different types of settings (i.e., individual vs. group human-robot interactions). However, little attention has been paid to these diverse settings in the affective computing community.

In this paper, we introduce a framework to analyse individual affect using both individual and group videos. We analyse the affective states along both arousal and valence dimensions by training and testing three different models. To achieve this, we trained different affect models using two different datasets, i.e., individual and group. Individual dataset refers to data collected from one participant watching affective movie stimuli, while group dataset refers to data collected from a group of participants watching affective movie stimuli together. Using these datasets we trained three affect recognition models: one model is trained with individual data only, i.e., *Individual Model*; one model is trained with group data only, i.e., *Group Model*; and another model is trained using both individual and group data, i.e., *Combined Model*. Our analysis shows that (1) although the model trained with a specific type of data shows the best testing performance within the corresponding data, *Group Model* can perform better with individual data than the other way around; and (2) *Combined Model* provides a good compromise between the *Individual Model* and *Group Model*.

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

2. THE PROPOSED FRAMEWORK

We propose a framework for the prediction of individual valence and arousal in both individual and group videos by using spatio-temporal facial features. The proposed framework is illustrated in Fig. 1. Our goal is to investigate whether and how affect recognition differs when training a model using the expressive data of individuals when they are alone versus when they are in a group setting. More specifically, how do the data for the same individual acquired in different settings influence affect recognition? Is it possible to obtain a similar performance on group data using a model trained with individual data and *vice versa*? To this

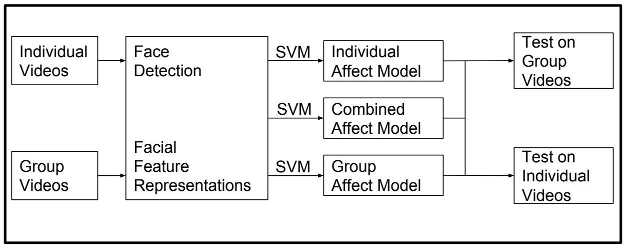


Figure 1: Illustration of the proposed framework.

end, we train three different classification models, including one model trained with data from the individual dataset (i.e., *Individual Model*), one model trained with data from the group dataset (i.e., *Group Model*) and a third model trained with data from both the individual and the group datasets (i.e., *Combined Model*). When training the different models, we use the same feature representation, i.e., facial spatio-temporal representations. We carry out multiple experiments to investigate the performance of different models on different types of data.

2.1 Facial Feature Extraction

Intraface [15] is 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, manual inpsection showed that 99% faces were detected in the individual database, while 96% of faces were detected in the group database.

In our previous work [9] showed that Volume Quantised Local Zernike Moments (*vQLZM*) facial representation outperformed other facial and body representations. Therefore, in this work, we use *vQLZM* for facial representation. After faces are detected, Quantised Local Zernike Moments (QLZM) [12] are obtained from the local patch around each facial landmark point as the appearance representation. QLZM is used as a low-level representation that is extracted by first calculating local Zernike Moments (ZMs) in the neighbourhood of each pixel of the input image. Then the accumulated local features are converted into position dependent histograms. Each ZM coefficient describes the texture variation at a unique scale and orientation. Once the ZMs are computed for all pixels, the QLZM descriptors are obtained by quantising all ZM coefficients around a pixel into a single integer. The QLZM [12], by design, takes into account only static spatial information, that is, it is designed for static images/frames [12]. In this paper, we used the extended volume representation to embed both appearance and temporal information as illustrated in Fig. 2. More details about *vQLZM* representation can be found in [9].

2.2 Fisher Vector Encoding

Fisher Vector (FV) encoding [10] has been widely used in computer vision problems such as action recognition [14] and depression analysis [3]. 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 GMM is then fitted to the extracted descriptors, *vQLZM*. 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. In this way, we obtained Fisher Vectors from *vQLZM* (*vQLZM-FV*).

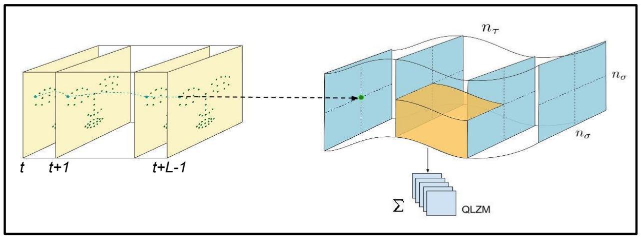
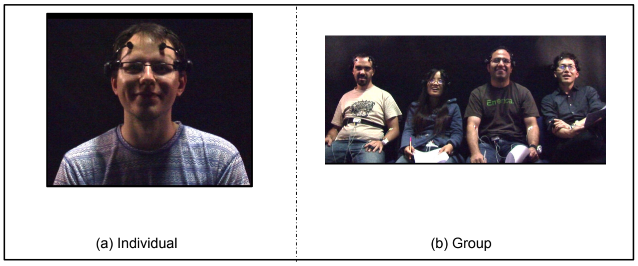


Figure 2: Illustration of the vQLZM feature extraction process. Left: Facial landmark points are detected. Facial landmark point tracking is in the spatial scale over L frames. Right: Appearance and motion information over a local neighbourhood of $N \times N$ pixels along each facial landmark point are extracted, where $N = 24$. 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 [14], parameters are set to $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$.


(a) Individual

(b) Group

Figure 3: The setup for individual and group data acquisition. In individual settings, each participant watched the movie stimuli alone, while in group settings, a group of four participants watched the movie stimuli together.

3. DATASETS AND EXPERIMENTS

3.1 Data Collection and Annotation

Two datasets were collected and annotated, namely the individual dataset and the group dataset. Sixteen participants (8 females and 8 males), aged between 25 and 38 were recorded while watching affective movies. Each participant was recorded in both individual and group settings. In order to avoid familiarity with the stimuli, they watched different movies in these two settings. For the individual dataset, each participant watched sixteen short movies separately. For the group dataset, the participants were arranged into four groups with four participants in each group, watching four long videos together. The setup for these two settings is shown in Fig. 3.

Independent observer annotations were obtained from ex-

Table 1: The distribution of samples for individual and group videos along both arousal and valence dimensions after quantisation.

Dimensions	**Arousal**		**Valence**	
Labels	High	Low	Positive	Negative
Group Data	197	379	171	405
Individual Data	326	937	438	825

ternal human labellers who are all researchers working on affect analysis. An in-house emotion annotation tool, that requires the labellers to scroll a bar between a range of values (0 and 1), was used. Individual (short) videos were divided into 10-second clips. All clips were annotated except the first and last 10s of each video. For the group (long) videos, 10-second clips were annotated 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 that 10-second clip of each participant separately and was asked to observe the non-verbal behaviors without hearing any audio. A single annotation was given by each labeller after watching one 10-second clip. In order to avoid confusion, arousal and valence annotations were obtained separately. The average of annotations from all labellers was calculated and used as thresholds (i.e., 0.5 for both arousal and valence in individual dataset, and 0.4 for arousal and 0.5 for valence in group dataset) to quantize the arousal and valence annotations into two classes – *high* and *low* arousal and *positive* and *negative* valence. The distribution of samples for individual and group videos along both arousal and valence dimensions after quantisation is shown in Table 1. Annotations from three labellers were obtained for group videos, but annotations from only one labeller were available for the individual videos. Details of the annotation and the inter-labeller agreement can be found in [9].

3.2 Experiments

Experimental setup. For the individual setup, data from 16 participants with 14 recordings were used providing us a total number of 1263 clips. For the group setup, data from the same 16 participants (4 groups) were used as follows: 3 groups (12 participants) with 4 recordings and 1 group (4 participants) with 2 recordings. As a result, there were data from 16 participants and 14 sessions in total. During each session, each group watched one movie. From each session, we used 10-second clips extracted every 2 minutes in line with the annotations obtained. The total number of samples from all subjects used in the experiments was 576.

Classifier. As we aim to investigate the effects of different settings (i.e., individual and group) in automatic recognition of affect rather than focus on improving recognition accuracy by testing with different feature representations and different machine learning techniques, we conducted experiments using the same classification technique and the same facial representation. We used Support Vector Machines (SVMs) and the LibSVM library [2], widely used in affect analysis [18, 5].

Evaluation. To evaluate the models, we used *leave-one-subject-out* cross-validation and *subject-specific* validation. Each time the parameters of the classifier were optimized over the training-validation samples. *Leave-one-subject-out* refers to, in each fold, using 15 participants for training-validation and the remaining one participant for testing. *Subject-specific* model was built by applying *leave-one-sample-out* cross-validation for each participant separately. Affect classification was evaluated in terms of average of $F1$ score across all samples (average of $F1$ score for both classes).

Table 2: Classification results obtained with *Individual Model* and *Group Model* on both individual and group datasets in terms of $F1$ score in *leave-one-subject-out cross-validation* and *subject-specific* setups

Models	**Individual Model**			
Test Data	**Individual Data**		**Group Data**	
Dimensions	Arousal	Valence	Arousal	Valence
Sub-specific	0.78	0.80	0.50	0.58
One-sub-out	0.63	0.68	0.51	0.53
Models	**Group Model**			
Test Data	**Individual Data**		**Group Data**	
Dimensions	Arousal	Valence	Arousal	Valence
Sub-specific	0.54	0.58	0.74	0.78
One-sub-out	0.58	0.57	0.61	0.70

Experiments. Linear-SVM was used for classification with respect to the dimensions of arousal (i.e., high and low) and valence (i.e., positive and negative). Prior to feeding the facial features to the classifier, PCA was first applied to reduce the dimensionality. In each case, we trained two binary linear-SVM classifiers, one using the individual dataset and the other one using the group dataset, which we refer to as *Individual Model* and *Group Model* respectively. We also tested the *Individual Model* on the group data and the *Group Model* on the individual data. Specifically, for *leave-one-subject-out* cross validation, we trained the model with group data from 15 participants, and tested with individual data from the remaining one participant, and *vice versa*. For *subject-specific* validation, we trained the model with group data from each participant, and then tested it with individual data from the same participant, and *vice versa*. In addition, we also trained a third model with data from both group and individual settings using *leave-one-subject-out* cross-validation. As the goal is to recognize the affect of each individual separately, the group model does not include any features/information of the other group members. In this way, we can have a fairer comparison between individual and group models, and can avoid some cases in which information of the other participants in the same group is lost due to occlusions or other challenging conditions.

3.3 Results and Analysis

The classification results of the *Individual Model* (M_I) and *Group Model* (M_G) are shown in Table 2. We can see that overall, *vQLZM-FV* proves to be a good representation for classifying affect along arousal and valence dimensions in both individual and group settings.

On one hand, for both *leave-one-subject-out* and *subject-specific* setups, we can see that the M_I shows slightly better performance than the M_G when classifying the data collected from the same type of settings. For instance, the recognition results are 0.78 for arousal and 0.80 for valence in terms of $F1$ score obtained with the M_I in *subject-specific* setup, while those obtained with the M_G are 0.74 for arousal and 0.78 for valence. On the other hand, M_G shows an ad-

Table 3: Classification results obtained with M_{GI} and M_{IG} in terms of $F1$ score for 10 sets of randomly sampled balanced data in *leave-one-subject-out cross-validation* and *subject-specific* setups. (3rd column) p-value obtained for comparisons between M_{GI} / M_{IG} and chance level (0.5); (last column) p-value obtained for each pair of M_{GI} / M_{IG} comparisons.

Arousal	**One-subject-out**										**mean**	**p-value**
M_{GI}	0.57	0.53	0.61	0.55	0.54	0.54	0.51	0.53	0.55	0.57	0.55(p<0.05)	<0.05
M_{IG}	0.46	0.56	0.49	0.48	0.49	0.44	0.44	0.54	0.55	0.35	0.49(p=0.3)	
Valence												
M_{GI}	0.63	0.56	0.58	0.59	0.57	0.58	0.52	0.60	0.59	0.59	0.58(p<0.05)	<0.05
M_{IG}	0.51	0.37	0.56	0.50	0.45	0.38	0.53	0.38	0.60	0.61	0.49(p=0.7)	
Arousal	**Subject-specific**											
M_{GI}	0.54	0.55	0.58	0.57	0.54	0.53	0.53	0.54	0.52	0.57	0.55(p<0.05)	<0.05
M_{IG}	0.49	0.54	0.51	0.46	0.50	0.48	0.53	0.47	0.46	0.56	0.50(p=1)	
Valence												
M_{GI}	0.59	0.57	0.53	0.54	0.59	0.54	0.52	0.52	0.60	0.54	0.55(p<0.05)	<0.05
M_{IG}	0.51	0.37	0.56	0.50	0.45	0.38	0.53	0.38	0.60	0.61	0.49(p=0.7)	

vantage in dealing with data collected from different types of settings (i.e., individual data). For example, arousal recognition results obtained on group test data with the M_I with *leave-one-subject-out* cross-validation is $F1$=0.51, while on individual test data with M_G the result obtained is $F1$=0.58.

In light of these findings we further hypothesise that the *Group Model* tested with individual data (M_{GI}) performs better than the *Individual Model* tested with group data (M_{IG}), and perform t-test to see the statistical significance of these results. We randomly sample the data to make it balanced between individual and group data as well as between the two classes along each dimension. Based on the data we have, each time we randomly select 171 samples for high and low arousal and 197 samples for positive and negative valence for both individual and group data. Then we train the M_{GI} model with randomly selected group data and test it with individual data, and vice versa for M_{IG}, for ten times. We perform (1) two-sample right-tail t-test for each pair of M_{GI} set and M_{IG} set to see whether M_{GI} is significantly better than M_{IG} and (2) one-sample t-test for each M_{GI} set and M_{IG} set to see whether they are significantly different from chance level, i.e., 0.5. The results presented in Table 3 confirm that our hypothesis is indeed correct. M_{GI} is significantly better than M_{IG} and M_{GI} performs significantly different from chance level, while M_{IG} does not. A possible explanation is that although group settings are relatively more complex and challenging, the group model ends up modelling a larger variety of affective behaviors. In group settings, each participant behaves in a multitude of ways, they might behave similarly to when they are recorded alone, or they might end up talking to others expressing their opinions.

The *Combined Model* (M_C) is trained with both individual and group datasets. The classification results obtained with the M_C are shown in Table 4. We can see that the M_C shows better performance than the M_I (M_G) tested with group data (individual data), but not as good as theM_G (M_I) tested with group data (individual data). Note that, the M_C is trained with all data from both individual and group datasets, only excluding the participant used as the test data at each round. Therefore, it is trained with more data than both the M_I and the M_G. However, more training data does not always provide a better model. A possible explanation is that the M_C has to make a compromise when modelling different types of data simultaneously, which results in decreased performance when compared to the M_I and the M_G tested on the same type of data.

Table 4: Classification results of the combined model on both individual and group datasets in terms of $F1$ score using *leave-one-subject-out* cross-validation

Models	**Combined Model**			
Test Data	**Individual Data**		**Group Data**	
Dimensions	Arousal	Valence	Arousal	Valence
One-sub-out	0.59	0.68	0.59	0.61

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework to investigate the effects of affect expressed in different settings (alone vs. in-a-group) on the automatic recognition performance. We conducted a set of experiments on two newly collected datasets. We trained three different SVM-based recognition models using different types of affect data, namely one model trained with data from one participant watching movie stimuli alone (i.e., *Individual Model*), one model trained with data from a group of four participants watching movie stimuli together (i.e., *Group Model*) and another model with the combined dataset (i.e., *Combined Model*).

The experimental results show that the best performance can be achieved by using data from the same type of setting (i.e., individual or group) than using data from different settings for training and testing. The best solution would be to provide multiple affect models trained with different types of data and settings. However, our experimental results also show that a model trained with group data is more generic than a model trained with individual data. Specifically, a model trained with group data has a better performance on individual data than vice versa. Similar results are also reported for the analysis of disengagement in human-robot interaction settings [7]. Our results indicate that in settings where multiple affect models trained with different types of data are not available, using the affect model trained with group data is a viable solution.

Although the experiments reported in this paper have been conducted in an audience context (i.e., participants watching movie stimuli), the proposed feature representation and classification methods can be applied to other types of settings, including human-robot interactions. Despite the promising results obtained, the impact of unbalanced data and feature representation needs to be investigated further by utilising a larger balanced dataset and other feature representations including audio and physiological signals.

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).

5. REFERENCES

[1] 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.

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

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

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

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

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

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

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

[9] W. Mou, H. Gunes, and I. Patras. Automatic recognition of emotions and membership in group videos. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition and Workshops (CVPRW)*, 2016.

[10] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 2013.

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

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

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

[14] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 2013.

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

[16] R. B. Zajonc et al. *Social facilitation.* Research Center for Group Dynamics, Institute for Social Research, University of Michigan, 1965.

[17] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2009.

[18] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2007.