Generic to Specific Recognition Models for Membership Analysis in Group Videos

Wenxuan Mou[1], Christos Tzelepis[1,3],
Vasileios Mezaris[3], Hatice Gunes[2], Ioannis Patras[1]
[1] Queen Mary University of London, UK, [2] University of Cambridge, UK
[3] Information Technologies Institute/Centre for Research and Technology Hellas (CERTH), Greece

***Abstract*— Automatic understanding and analysis of groups has attracted increasing attention in the vision and multimedia communities in recent years. However, little attention has been paid to the automatic analysis of group membership – i.e., recognizing which group the individual in question is part of. This paper presents a novel two-phase Support Vector Machine (SVM) based *specific recognition model* that is learned using an optimized *generic recognition model*. We conduct a set of experiments using a database collected to study group analysis from multimodal cues while each group (i.e., four participants together) were watching a number of long movie segments. Our experimental results show that the proposed *specific recognition model* (52%) outperforms the *generic recognition model* trained across all different videos (35%) and the *independent recognition model* trained directly on each specific video (33%) using linear SVM.**

I. INTRODUCTION

Automatic analysis of a group of people has received much attention in computer vision community for different research purposes. Gallagher et al. [1] propose a framework to predict the age and the gender of individuals in group images. Ibrahim et al. [2] focus on group activity recognition. More recently, other research fields, including emotion recognition, have also started to shift their focus from individual to group settings [3], [4]. Research works focusing on the analysis of social dimensions, such as engagement and rapport in group settings have also been introduced [5], [6]. Most of the aforementioned works analyze what is happening within the group. Only recently, works focusing on automatic analysis of the relationship between the members of different groups have emerged. Correa et al. [7] predicted whether a person is in alone or in a group using neuro-physiological signals. In our previous work [8] we introduced group membership recognition using non-verbal behaviors, where group membership recognition refers to recognizing which group each individual is part of.

In this paper we aim to investigate whether we can predict the group membership of each individual when they are part of a group of four participants sitting together and watching four movies. Group here refers to the four people who sit and watch movies together. We form three groups with twelve participants in total and there is no overlap between the group members of these three groups. Even though they are performing the same task, individuals in different groups may behave very distinctly due to differences in group composition and dynamics. According to cognitive and behavioral science researchers, individuals in one group tend to affect the behaviors of each other – i.e., mimic one another or display similarities in non-verbal behaviors [9]. Such shared behaviors within the group, and possible differences between different groups, enable the automatic recognition of group membership [8].

In this paper, we propose a novel solution to the group membership recognition problem. We introduce a novel two-phase Support Vector Machine (SVM) based *specific recognition model* that is learned using an optimized *generic recognition model*. More specifically, the data at hand consists of recordings (videos) of different groups watching different movies. Previous work [8] focused on group membership recognition across all different videos, which in our two-phase framework is referred to as the *generic recognition model*. However, we note that group members behave distinctly while watching different movies, which limits the performance of the *generic recognition model*. If we attempt to solve the membership recognition problem with an *independent recognition model* using only samples from the same video, it becomes very challenging due to the small number of samples available from each video. When the group members are watching different movies, they may react differently; however, they are still part of the same setting performing the same task (i.e., sitting in front of the screen watching movies), which enables them to share some common behavioral characteristics. Therefore, we hypothesize that the *generic recognition model* can provide a useful baseline for the optimization of the *specific recognition model* via a two-phase learning. In order to optimize the *specific recognition model*, we first train a *generic recognition model* using all videos and, then, optimize the *specific recognition model* for each specific video based on the optimization results obtained from the *generic recognition model*. The group membership recognition results obtained through this framework show that the proposed *specific recognition model* outperforms both *the generic recognition model* that was trained across all videos using standard linear SVM, and *the independent recognition model* that was trained directly on each video using standard linear SVM.

The rest of the paper is organized as follows. The related works are reviewed in Section II; the proposed framework is introduced in Section III; the experiments and results are

978-1-5090-4023-0/17/$31.00 ©2017 IEEE

presented and discussed in Section IV; and conclusions and future work are discussed in Section V.

II. Related Work

Individual-level analysis. In [8], the authors proposed a framework for individual affect analysis in group videos along arousal and valence. Leite et al. [5] studied the individual engagement estimation in group settings in the context of human-robot interaction. Hagad et al. [6] automatically predicted rapport in dyadic interactions based on posture and congruence. Gallagher et al. [1] introduced a framework to perform individual analysis, i.e., age and gender recognition by using contextual features that captured the structure of people in the image instead of using features from each individual. Ramanathan et al. [10] proposed to recognize the social roles played by individuals in an event, e.g., instructor and student. In addition to these works analyzing what happened within a group, there are also works focusing on the analysis of the relationship between members across different groups, such as group membership analysis from a social psychological perspective [11], [12]. However, little attention has been paid to automatic analysis of group membership.

Group-level analysis. From a psychological perspective, a large number of group analysis focus on group emotion [13] and group cohesion [14]. Automatic analysis has also moved from individual-level to group-level analysis. Pioneering works on affect recognition analysed the overall affect displayed by the whole group [15], [16], [17], [4], [18]. In addition, some previous works on group-level analysis focused on group activity recognition [19], [20]. Although information about the member was used to predict the group-level attributes, all of these works aim to analyse the collective attributes expressed by the whole group rather than analyse what was displayed by each individual.

Non-verbal cues for group analysis. Non-verbal behaviors are very important cues for group analysis [9]. The most frequently used non-verbal behaviors include gaze patterns, body motion, head movements and facial expressions [21], [22]. Sanchez-Cortes et al. [21] used nonverbal behaviors (both audio and visual non-verbal behaviors) to automatically identify emergent leaders in small group scenarios. Hung and Gatica-Perez [23] did group cohesion estimation by utilizing non-verbal behaviors, e.g., activity of each person and motion informations. Mou et al. [8] analysed the affect of individuals and group membership by using the non-verbal face and body features and reported that body behaviors showed better performance for group membership recognition. Thus, in this work we use body behaviors for group membership recognition.

III. The Proposed Framework

We propose a novel framework for the recognition of group membership in group videos by analysing body behaviours. The proposed framework is illustrated in Fig. 1. We propose a novel two-phase learning approach to learn a *specific recognition model* upon a *generic recognition model*. The first step is to learn a *generic recognition model* using all data across all videos. The second step is to learn the *specific recognition model* using data from only one specific video based on the optimized *generic recognition model*. As the data across different videos are all under the same scenario, that is sitting in front of the screen watching movies, we hypothesize that, the two recognition models share some common knowledge and therefore the *generic recognition model* can provide a baseline for optimizing the *specific recognition model*.

A. The Generic Recognition Model

Our recognition models are based on linear support vector machine (SVM). For training each of the linear models we use a stochastic gradient descent (SGD) algorithm motivated by Pegasos and first proposed by Shalev-Shwartz et al. [24]. Pegasos is a well-studied algorithm [25], [26] providing both state-of-the-art classification performance and great scalability.

The first step of the proposed framework is to learn the *generic recognition model* using the standard linear SVM. In this model, we use all of the available training samples, which are from all subjects across all videos. We denote this training set as $\mathcal{X} = \{(\mathbf{x}_i, z_i), i = 1, \ldots, \ell\}$, where $\mathbf{x}_i$ denotes the i-th training sample and z_i the corresponding ground truth label, being equal to $+1$ if the sample belongs to the respective positive class, or -1 otherwise.

The generic optimization problem, which we denote as $\mathcal{P}_{generic}$, can be cast as follows:

$$\mathcal{P}_{generic}: \quad \min_{\mathbf{w}_0, b_0} \frac{\lambda}{2}\|\mathbf{w}_0\|^2 + \frac{1}{\ell}\sum_{i=1}^{\ell}\mathcal{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i)), \quad (1)$$

Where λ is the regularization parameter and $\mathbf{w}_0$, b_0 are the optimization parameters. $\mathcal{L}$ denotes the loss function and is given by the hinge-loss, as follows

$$\mathcal{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i)) = \max(0, 1 - z_i(\mathbf{w}_0^\top \mathbf{x}_i + b_0)). \quad (2)$$

We use the Pegasos [24] SGD algorithm for solving the above optimization problem and we arrive at the optimal solution $(\mathbf{w}_0, b_0)$, which describes the optimal hyperplane $\mathcal{H}_0: \mathbf{w}_0{}^\top \mathbf{x} + b_0 = 0$. Then, we use the optimal $\mathbf{w}_0$ to construct the *specific recognition model*, as described below.

B. The Specific Recognition Model

The *specific recognition model* is learned utilizing the optimization results obtained from the *generic recognition model*. That is, we use the optimal value for $\mathbf{w}_0$ (by solving the optimization problem in equation (1)) in order to construct the specific optimization problem, which we denote as $\mathcal{P}_{\text{specific}}$ and is given as follows

$$\mathcal{P}_{\text{specific}}: \quad \min_{\mathbf{w}, b} \frac{\mu}{2}\|\mathbf{w}\|^2 + \frac{\nu}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 + \frac{1}{|\mathcal{X}_t|}\sum_{(\mathbf{x}_i, z_i)\in\mathcal{X}_t}\mathcal{L}(\mathbf{w}, b; (\mathbf{x}_i, z_i)), \quad (3)$$

where $\mathcal{X}_t$ is a subset of the original training set, μ and ν are regularization parameters, and $\mathcal{L}$ denotes the hinge-loss.

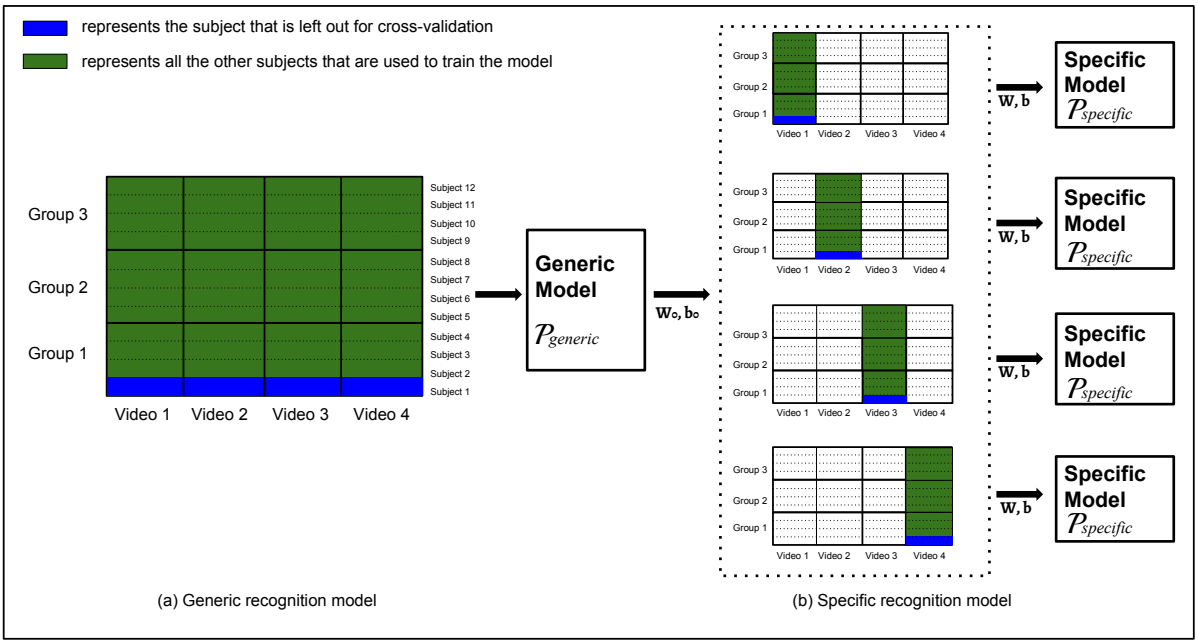


Fig. 1: An illustration of the proposed framework. It is divided into two learning phases, i.e., (a) learning the *generic recognition model* and (b) learning the *specific recognition model*. As we apply *leave-one-subject-out* cross-validation, for *generic recognition model*, we leave all of the samples of one subject (blue) out and train the model with all the other samples (green). For the *specific recognition model*, as we have four different videos, we have $n = 4$ specific problems and optimize them based on the optimized weights obtained from the *generic recognition model*. For the specific model, we also do *leave-one-subject-out* cross-validation.

The term $\frac{\nu}{2}\|\mathbf{w} - \mathbf{w}_0\|^2$ is used to bias $\mathbf{w}$ to be close to $\mathbf{w}_0$. When ν is equal to 0, the model becomes the standard linear SVM, while when ν tends to infinity, $\mathbf{w}$ tends to be equal to $\mathbf{w}_0$. The optimal values for μ, ν are obtained using cross-validation.

For solving $\mathcal{P}_{\text{specific}}$, we use a variant of the Pegasos SGD algorithm. That is, the proposed algorithm receives two parameters as input: (1) the number of iterations, T, and (2) the number of examples to be used for calculating sub-gradients, k. Initially, we set $\mathbf{w}^{(1)}$ to any vector whose norm is at most $1/\sqrt{\nu}$ and $b^{(1)} = 0$. On the t-th iteration, we randomly choose a subset of $\mathcal{X}$, of cardinality k, i.e., $\mathcal{X}_t \subseteq \mathcal{X}$, where $|\mathcal{X}_t| = k$ and set the learning rate to $\eta_t = \frac{1}{\nu_t}$.

We approximate the objective function of $\mathcal{P}_{\text{specific}}$ with

$$\mathcal{P}_{\text{specific}}: \quad \mathcal{J}(\mathbf{w}, b) = \frac{\mu}{2}\|\mathbf{w}\|^2 + \frac{\nu}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 + \frac{1}{k}\sum_{(\mathbf{x}_i, z_i)\in X_t} \mathcal{L}(\mathbf{w}, b; (\mathbf{x}_i, z_i)). \tag{4}$$

The update rules are given as follows

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \frac{\eta_t}{k}\frac{\partial \mathcal{J}}{\partial \mathbf{w}}, \quad b^{(t+1)} \leftarrow b^{(t)} - \frac{\eta_t}{k}\frac{\partial \mathcal{J}}{\partial b},$$

where the first derivatives of $\mathcal{J}$ with respect to $\mathbf{w}$ and b are given respectively as

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}} = \mu\mathbf{w} + \nu(\mathbf{w} - \mathbf{w}_0) + \frac{1}{k}\sum_{(\mathbf{x}_i, z_i)\in X_t} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \tag{5}$$

and

$$\frac{\partial \mathcal{J}}{\partial b} = \frac{1}{k}\sum_{(\mathbf{x}_i, z_i)\in X_t} \frac{\partial \mathcal{L}}{\partial b}. \tag{6}$$

The first derivatives of the hinge loss with respect to $\mathbf{w}$ and b are given respectively as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \begin{cases} -z_i\mathbf{x}_i & \text{if } 1 > z_i(\mathbf{w}^\top\mathbf{x}_i + b), \\ 0 & \text{if } 1 < z_i(\mathbf{w}^\top\mathbf{x}_i + b). \end{cases} \tag{7}$$

and

$$\frac{\partial \mathcal{L}}{\partial b} = \begin{cases} -z_i & \text{if } 1 > z_i(\mathbf{w}^\top\mathbf{x}_i + b), \\ 0 & \text{if } 1 < z_i(\mathbf{w}^\top\mathbf{x}_i + b). \end{cases} \tag{8}$$

Finally, we project $\mathbf{w}^{(t+1)}$ onto the ball of radius $1/\sqrt{\nu}$, i.e., the set $\mathcal{B} = \{\mathbf{w}: \|\mathbf{w}\| \leq 1/\sqrt{\nu}\}$. The output of the algorithm is the pair of $\mathbf{w}^{(T+1)}$, $b^{(T+1)}$.

Once the optimal values of the parameters $\mathbf{w}$ and b are learned, an unseen testing datum, $\mathbf{x}_t$, can be classified to one of the two classes according to the sign of the (signed) distance between $\mathbf{x}_t$ and the separating hyperplane. That is, the predicted label of $\mathbf{x}_t$ is computed as $y_t = \text{sgn}(d_t)$, where $d_t = (\mathbf{w}^\top\mathbf{x}_t + b)/\|\mathbf{w}\|$. The posterior class probability, i.e, a probabilistic degree of confidence that the testing sample belongs to the class to which it has been classified, can be calculated using the Platt scaling algorithm [27] for fitting a sigmoid function, $S(d_t) = 1/(1 + e^{\sigma_A d_t + \sigma_B})$. The scaling parameters σ_A, σ_B are obtained by applying the Platt scaling approach after solving the *generic recognition model*. Platt scaling is a well-known technique that has been shown to be particularly effective for max-margin methods such as SVMs (e.g., see [28]) for evaluating a sample's class membership at the testing phase.

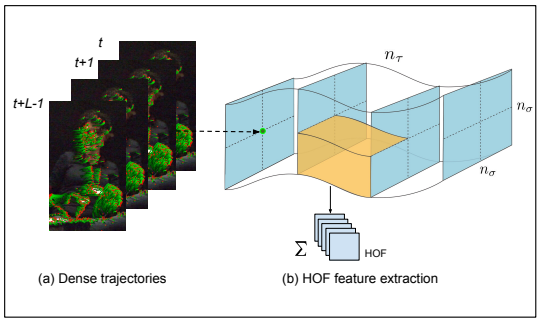


Fig. 2: Illustration of the approach to extract the body HOF feature. (a) Dense trajectory detection results. (b) Dense trajectory is in the spatial scale over L frames. Motion information over a local neighborhood of $N \times N$ pixels along the each trajectory point are extracted. In order to embed the structure information, the local volume is subdivided into a spatio-temporal grid of size $n_\tau \times n_\sigma$. Based on [29], $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$.

C. Feature Extraction

1) Low-level Feature Extraction: Some previous works showed that body features outperform facial features for group membership recognition [8], therefore, we use the body features in this work. In order to extract person-based representations we first need to apply a person detector. In our simplified setting with a fixed number of individuals and a static camera, we use an ad-hoc scheme that divides the frame into equally sized parts. Then, dense trajectories [29] are extracted and, subsequently, HOF descriptors are obtained around each trajectory. HOF descriptors are computed in the spatio-temporal volume aligned with the trajectories as shown in Fig. 2. HOF orientations are quantized into eight bins with full orientations. An additional zero bin is added for pixels with optical flow magnitudes lower than the threshold (i.e., nine bins in total). Thus, the final descriptor size is 108 with the trajectory length $L = 15$ frames. More details on this procedure can be found in [29].

2) Fisher Vector Encoding: Fisher vector (FV) encoding [30] has been widely used in computer vision problems such as action recognition [29] and depression analysis [31], [32]. 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 HOF descriptors. A GMM is then fitted to HOF descriptors. The number of Gaussians is set to $K = 256$ and a subset of 256000 descriptors is randomly sampled to fit a GMM. Subsequently, each clip is represented by a $(2D+1)K$ dimensional Fisher Vector, where D is the dimensionality of the descriptor after performing PCA. We obtain the Fisher Vectors (FVs) from body HOF descriptors.

IV. Experiments and Analysis

A. Data

Experiments are conducted using a database collected to study group analysis from multimodal cues while each group (i.e., four participants) were watching a number of long movie segments [7]. Four long movie segments (duration of each longer than 14 mins and smaller than 24 mins) were used as stimuli, details of which are provided in Table I. Twelve participants (7 females and 5 males), aged between 25 and 38 were recorded while watching these movies. They were arranged into three groups with four participants in each group watching all of the four videos listed in Table I together. Videos were recorded at 1280×720 resolution, 25fps. A representative frame from the database is shown in Fig. 3.

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

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

Fig. 3: A representative frame from the database.

B. Experiments

1) Implementation details: As we have a multi-class ($K = 3$ classes) recognition problem, we follow an "one-against-all" procedure to learn K binary classifiers and apply all classifiers to an unseen sample x to predict the label K for which the corresponding classifier reports the highest confidence score. The confidence score is calculated using the well-known Platt scaling approach [27] for fitting a sigmoid function.

2) Experimental setup: Data from three groups were used in our experiments, namely three groups (twelve subjects) with recordings from four different videos (N1, P1, B1 and U1 movies). As a result, there were twelve subjects from twelve recordings in total. During each recording, each group watched one movie. From each recording, we used 10-seconds clips extracted every 2 minutes. The number of short clips from each recording varies with the length of the movies, i.e., 12 clips for N1 and B1, 9 clips for P1 and 7 clips for U1. Therefore, the total number of clips we used in the experiments is $(12 \times 4 \times 3) + (12 \times 4 \times 3) + (9 \times 4 \times 3) + (7 \times 4 \times 3) = 480$.

We compared the proposed *specific recognition model* with two other models, (1) the *generic recognition model* that trained across all different videos as illustrated in (a) of Fig. 1 and (2) the *independent recognition model* that trained directly in each specific video as illustrated in Fig. 4.

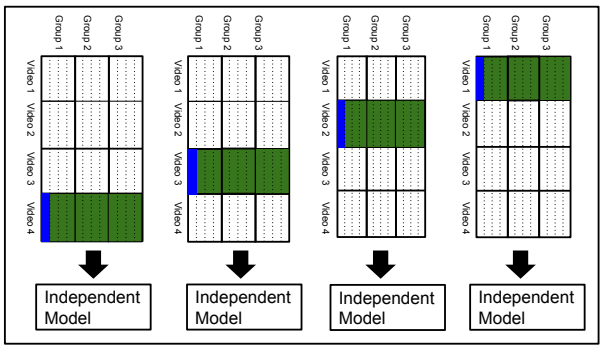


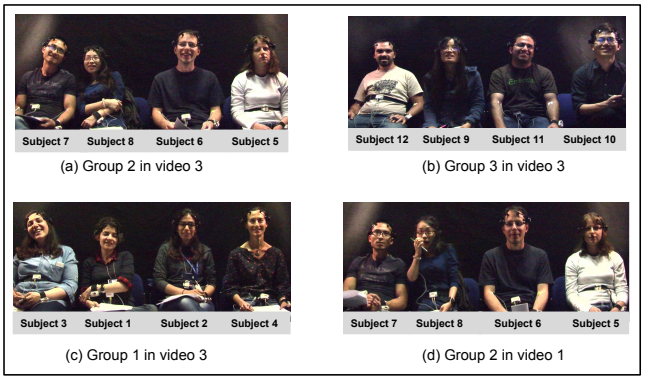
Fig. 4: An illustration of the independent recognition model.

In order to avoid subject-dependency problem, group membership recognition models were trained by applying *leave-one-subject-out* cross-validation. Each time the parameters of the model were optimized over the training-validation samples. *Leave-one-subject-out* refers to, in each fold, using eleven subjects for training-validation and the remaining one subject for testing. For each subject, the learning process is divided into two phases, *generic recognition model* learning and *specific recognition model* learning. The experimental results of the membership recognition were evaluated by the recognition accuracy. In addition, we test the results for the statistical significance.

3) Experimental results and analysis: The recognition results in terms of recognition accuracy by applying *leave-one-subject-out* cross-validation are shown in Table II and Table III. From Table II, we can clearly see that the proposed *specific recognition model* outperforms the other two models in terms of recognition accuracy. 52% is obtained for the *specific recognition model*, while 35% and 33% are obtained from *generic recognition model* and *independent recognition model* respectively. We also perform t-test to see the statistical significance, which is also listed in Table III and shows that the results obtained with the proposed *specific recognition model* are significantly better than chance level, but not for *generic recognition model* and *independent recognition model*. From Table III, we can see that models trained for different subjects have different performances and even for the same subject, the performances in different videos show some differences. Firstly, we can see recognition performs better for group 1 and group 3 than group 2. It is possibly because in group 2, subject 7 and subject 8 have very close relationship (husband and wife), while for group 1 and group 3, group members knew each other at a similar level. From (a) and (d) of Fig. 5, we can see that subject 7 and subject 8 are close to each other, but this is not the case for the other two groups. Therefore, compared to the other two groups, group 2 tends to share less information among all group members. Secondly, from Table III, we can see that group 3 shows very low performance for video 3. It is possibly because subject 5 in (a) of Fig. 5 did not like that movie. As we can see, subject 5 showed a very different behaviour while watching video 3 (Mr. Bean). From (a), (b) and (c) of Fig. 5, we can see that all of the participants seem to be very happy or excited and tend to move a lot, but not subject 5. Thus, in this case, it is difficult to recognize the

TABLE II: Group membership recognition results obtained using different models, the proposed *specific recognition model*, *generic recognition model* and *independent recognition model*. Here are the average recognition accuracy of all subjects obtained from *leave-one-subject-out* cross-validation and statistical significance test (p-value) obtained for comparisons with chance level (0.33).

Different Models	Average Accuracy (p-value)
Specific recognition model	52% (p<0.01)
Generic recognition model($\nu \to \infty$)	35% (p=0.41)
Independent recognition model($\nu = 0$)	33% (p=0.42)



(a) Group 2 in video 3 (b) Group 3 in video 3

(c) Group 1 in video 3 (d) Group 2 in video 1

Fig. 5: Four illustrative frames from different groups and different videos.

group membership of subject 5, which also causes difficulties in membership recognition of the other group members. For video 1 of group 2, as they are watching a relatively scary movie (Descend) and subject 8 (in (a) and (d) of Fig. 5) seems more scared than others, and she behaved differently from the other group members. This situation also adds more difficulties for the membership recognition in this group.

V. Conclusions and Future Work

In this paper, we proposed a novel two-phase Support Vector Machine (SVM) based *specific recognition model* that is learned using an optimized *generic recognition model*. We conducted a set of experiments on a database that includes three groups with four participants watching affective stimuli. Our experimental results show that the proposed approach outperformed the other standard methods. As group membership can be recognized using non-verbal behaviors (i.e., body behaviors), our results indicate that individuals influence each others behaviours within a group and their nonverbal behaviors share commonalities. Our results also show that capitalizing on shared information in a generic recognition problem is important for learning the specific problem at hand, and this optimization approach can be possibly transferred to other recognition domains.

Despite the promising results obtained in the experiments, analysis of group membership remains a challenging problem. As future work, we will use more data for training / testing while also utilizing other feature representations. In addition, we will also apply this two-phase learning approach to other recognition problems.

TABLE III: Group membership recognition results of each subject in each video obtained using the proposed *specific recognition model*. Here are the recognition accuracy for each subject obtained from *leave-one-subject-out* cross-validation.

Groups	Subjects	Videos	*specific recognition model*	Average
Group 1	Subject 1	video 1	0.33	0.39
		video 2	0.58	
		video 3	0.22	
		video 4	0.43	
	Subject 2	video 1	0.50	0.64
		video 2	0.92	
		video 3	0.56	
		video 4	0.57	
	Subject 3	video 1	0.17	0.44
		video 2	0.58	
		video 3	0.44	
		video 4	0.57	
	Subject 4	video 1	0.92	0.76
		video 2	0.83	
		video 3	0.56	
		video 4	0.71	
Group 2	Subject 5	video 1	0.00	0.46
		video 2	0.83	
		video 3	0.00	
		video 4	1.00	
	Subject 6	video 1	0.50	0.27
		video 2	0.25	
		video 3	0.33	
		video 4	0.00	
	Subject 7	video 1	0.08	0.25
		video 2	0.92	
		video 3	0.00	
		video 4	0.00	
	Subject 8	video 1	0.25	0.39
		video 2	0.83	
		video 3	0.33	
		video 4	0.14	
Group 3	Subject 9	video 1	0.67	0.77
		video 2	0.92	
		video 3	0.78	
		video 4	0.71	
	Subject 10	video 1	0.83	0.89
		video 2	1.00	
		video 3	1.00	
		video 4	0.71	
	Subject 11	video 1	0.17	0.24
		video 2	0.58	
		video 3	0.22	
		video 4	0.00	
	Subject 12	video 1	0.92	0.74
		video 2	0.75	
		video 3	0.89	
		video 4	0.43	

Video 1: in video 1 participants were watching movie N1; Video 2: in video 2 participants were watching movie B1; Video 3: in video 3 participants were watching movie P1; Video 4: in video 4 participants were watching movie U1.

VI. ACKNOWLEDGMENTS

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). The work of Christos Tzelepis and Vasileios Mezaris is supported by EU's Horizon 2020 programme under grant agreement H2020-693092 MOVING.

REFERENCES

[1] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *CVPR*, 2009.

[2] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," *arXiv preprint arXiv:1511.06040*, 2015.

[3] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A comparative analysis of facial affect recognition," in *ACMMM*, 2016.

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

[5] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *ACM/IEEE HRI*, 2015.

[6] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *Proc. of IEEE Int. Conf. on Social Computing*, 2011.

[7] J. Abdon Miranda-Correa, M. Khomami Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Mood, Personality and Affect Research on Individuals and Groups," *ArXiv e-prints*, 2017.

[8] W. Mou, H. Gunes, and I. Patras, "Automatic recognition of emotions and membership in group videos," in *CVPRW*, 2016.

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

[10] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *CVPR*, 2013.

[11] L. Goette, D. Huffman, and S. Meier, "The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups," 2006.

[12] M. Williams, "In whom we trust: Group membership as an affective context for trust development," *Academy of management review*, 2001.

[13] E. R. Smith, C. R. Seger, and D. M. Mackie, "Can emotions be truly group level? evidence regarding four conceptual criteria." *Journal of personality and social psychology*, 2007.

[14] E. Salas, R. Grossman, A. M. Hughes, and C. W. Coultas, "Measuring team cohesion observations from the science," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2015.

[15] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding happiest moments in a social context," in *ACCV*, 2012.

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

[17] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *FG*, 2015.

[18] X. Huang, A. Dhall, G. Zhao, R. Goecke, and M. Pietikäinen, "Riesz-based volume local binary pattern and a novel group expression model for group happiness intensity analysis." in *BMVC*, 2015.

[19] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity recognition," in *CVPR*, 2012.

[20] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 2012.

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

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

[23] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Tran. on Multimedia*, 2010.

[24] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for SVM," in *ICML*, 2007.

[25] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for SVM," *Mathematical programming*, 2011.

[26] S. M. Kakade and A. Tewari, "On the generalization ability of online strongly convex programming algorithms," in *Advances in Neural Information Processing Systems*, 2009.

[27] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, 1999.

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

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

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

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

[32] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *ACII*, 2015.