A Deep Generic to Specific Recognition Model for Group Membership Analysis using Non-verbal Cues

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 the non-verbal behaviors and how this can be utilized for analysis of group membership, i.e., recognizing which group each individual is part of. This paper presents a novel Support Vector Machine (SVM) based Deep *Specific Recognition Model (DeepSRM)* that is learned based on a *generic recognition model*. The *generic recognition model* refers to the model trained with data across different conditions, i.e., when people are watching movies of different types. Although the *generic recognition model* can provide a baseline for the recognition model trained for each specific condition, the different behaviors people exhibit in different conditions limit the recognition performance of the generic model. Therefore, the *specific recognition model* is proposed for each condition separately and built on top of the *generic recognition model*. A number of experiments are conducted using a database aiming to study group analysis while each group (i.e., four participants together) were watching a number of long movie segments. Our experimental results show that the proposed *deep specific recognition model* (44%) outperforms the *generic recognition model* (26%). The recognition of group membership also indicates that the non-verbal behaviors of individuals within a group share commonalities.

Keywords: Non-verbal behavior analysis, Group membership, Automatic group analysis, Deep learning

1. INTRODUCTION

Automatic analysis of a group of people has received much attention in computer vision community for different research purposes. Gallagher and Chen (2009) propose a framework to predict ages and genders of individuals in group images. Ibrahim et al. (2015) 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 Mou et al. (2016a, 2015). Research works focusing on the analysis of social dimensions, such as engagement and rapport in group settings have also been reported in Leite et al. (2015) and Hagad et al. (2011). Zhang and Hung (2016) and Vascon et al. (2016) propose frameworks for F-formation detection in group conversations. Most of the aforementioned works analyze what is happening within the group. Only recently, works on automatic analysis of the relationship between the members of different groups have emerged. Abdon Miranda-Correa et al. (2017) predict whether a person is being alone or in a group utilizing neuro-physiological signals.

In this paper we investigate the prediction of group membership for each individual, using non-verbal behaviors, when they are part of a group of four participants sitting together and watching four movies. We form four groups, each of which contains four participants, with no overlaps between the group members (sixteen participants in total). 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 research in cognitive and behavioral science Barsade (2002), individuals in a particular group tend to affect the behaviors of each other, i.e., mimic one another or exhibit similarities in non-verbal behaviors. Such shared behaviors within the group, and possible differences between different groups, allow the automatic recognition of group membership Mou et al. (2016b).

Towards this direction, we propose a novel approach to the group membership recognition problem by introducing a novel *specific recognition model* that is built on the top of a *generic recognition model*. In the proposed framework the data at hand consists of recordings (videos) of different groups watching different types of movies. We define four different conditions as people are watching four different types of movies, namely, “horror”, “comedy”, “action”, and “adventure”, as shown in Table 1. The *generic recogni-*

Table 1: The stimuli video clips extracted from different films that are used in the paper. The video IDs are stated in parentheses and are used to refer to the videos in the rest of the paper; the corresponding conditions and video durations (in minutes) are also listed.

Movie	Condition	Duration
Descent (N1)	Horror	23:35
Mr. Bean (P1)	Comedy	18:43
Batman the Dark Knight (B1)	Action	23:30
Up (U1)	Adventure	14:06

tion model, that was proposed in our previous work Mou et al. (2017), allows the group membership recognition across all different conditions. However, since group members may behave distinctly in different conditions (e.g., while watching horror movies vs. comedies), the performance of *generic recognition model* may be significantly limited. Addressing the membership recognition problem with an *independent recognition model*, i.e., using solely the data from the same condition, becomes very challenging due to the limited number of samples available from each video. Moreover, when the group members are in different conditions, 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 allows them to share some common behavioral characteristics. In light of these, we hypothesize that the *generic recognition model* can provide a useful baseline for the optimization of the *specific recognition model*. Therefore, we propose a *specific recognition model* for each condition specifically, but we learn it on the top of the *generic recognition model*.

This paper is an extended version of our previous work Mou et al. (2017). In Mou et al. (2017), we proposed a two-phase learning framework to solve the group membership recognition problem, where we first trained a *generic recognition model* using all videos across all conditions and, then optimized the *specific recognition model* for each specific condition based on the optimization results obtained from the *generic recognition model*. Different from the aforementioned paper, in this work we unify the *generic recognition model* and the *specific recognition model* under a single deep framework. Specifically, in this work we optimize the *generic recognition model* and the *specific* *recognition model* jointly. In this way, the framework is converted to an end-to-end structure, which is easier for both training and testing. Furthermore, we conduct new experiments with a larger dataset. In the rest of the paper, we refer to the *specific recognition model* (as presented in our previous work Mou et al. (2017)) as the two-phase *Specific Recognition Model* (SRM) and to the proposed *Deep Specific Recognition Model* as the DeepSRM.

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

2. Related Work

Analysis of group-related phenomena has been studied for a long time across various disciplines, such as psychology and computer science Goette et al. (2006); Smith et al. (2007); Allen et al. (2017); Sanchez-Cortes et al. (2012); Girard et al. (2017). It has applications in very diverse areas, such as human-robot interaction Leite et al. (2015), security Saxena et al. (2008), and marketing analysis Eberl (2010). In these works group is defined as consisting of at least two members. However, at times dyads is separated as one category as dyads often form and dissolve more easily than groups, and people show different behaviors and experience different emotions in dyads than in groups Reiter-Palmon et al. (2017). Group dynamics encompasses those behaviors and psychological processes that occur within a group (intragroup dynamics) or between groups (intergroup dynamics) Lehmann-Willenbrock et al. (2017). Therefore, analysis in group settings is more difficult than that in individual settings due to the complex dynamics.

We will review the literature of group analysis from data, features and methodologies. In addition, representative works on group analysis using non-verbal cues are illustrated in Table 2.

Data for group analysis. Sanchez-Cortes et al. (2012) collected the Emergent LEAder corpus (ELEA) database to analyze the emergent leadership phenomenon, where the participants were asked to participate in a winter survival task. The dataset consists of 40 meetings (i.e., 28 four-person meetings and 12 three-person meetings). All groups were newly formed, namely, composed of unacquainted people in each group. Leite et al. (2015) collected a dataset for engagement analysis in group settings, where three

Table 2: Representative works on group analysis.

Reference	Analyzed Phenomena	Individual or group level	Number of images or frames	Data Type	Data Source	Features
Celiktutan et al. (2017)	Engagement & personality	Individual level	15,300 seconds	Dynamic videos	Experiments	Nonverbal-audio & visual & physiological signals
Girard et al. (2017)	Facial action unit	Individual level	172,800	Dynamic videos	Experiments	Face features
Mou et al. (2016b,a)	Emotion	Individual level	144,000	Dynamic videos	Experiments	Face & body
Dhall et al. (2015a), Huang et al. (2015)	Emotion	Group level	3,134	Static images	Web	Face & scene
Dhall et al. (2015b)	Emotion	Group level	504	Static images	Web	Face & scene
Mou et al. (2015)	Emotion	Group level	250	Static images	Web	Face, body & context
Leite et al. (2015)	Engagement	Individual level	6,348 seconds	Dynamic videos	Experiments	Audio, face, body & context
Gallagher and Chen (2009)	Age & gender	Individual level	5,080	Static images	Web	Context
Hung and Gatica-Perez (2010)	Group cohesion	Group level	14,400 seconds	Dynamic videos	Experiments	Audio & visual activity
Our work	Group Membership	Group level	1,792,575 (71,703 seconds)	Dynamic videos	Experiments	Body features

children were interacting with two robots. Seven groups of data were collected in total and the average interaction time of each group is 4 minutes and 36 seconds. Celiktutan et al. (2017) collected a Multimodal Human-Human-Robot Interactions (MHHRI) dataset for engagement and personality investigation in human-human dyadic interaction and human-robot interactions (two people interact with a humanoid robot). 18 people participated in the recording and 12 independent interaction sessions were recorded with different sensors, e.g., first-vision cameras, Kinet depth sensors and physiological sensors, which resulted in around 4 hours 15 minutes recordings. For automatic group emotion analysis, the first dataset was HAPPEI collected by Dhall et al. (2015a) from Flickr by using key words that describe groups and events. 2,886 images were collected and all images were annotated with a group-level happy intensity. In addition, 8,500 faces were manually annotated for face level happiness intensity by 4 human annotators. After that, Dhall et al. (2015b) collected a GAFF database, which extended the HAPPEI database from positive affect only Dhall et al. (2015a) to other emotion categories (i.e., positive, neutral and negative) of a group of people. In a further step, Mou et al. (2015) collected a new database, i.e., MultiEmoVA, which were annotated along both arousal (i.e., high, medium and low) and valence (i.e., positive, neutral and negative) dimension.

Features for group analysis. Non-verbal behaviors are very important cues for group analysis Barsade (2002). The most frequently used non-verbal behaviors include gaze patterns, body motion, head movements, and facial expressions Sanchez-Cortes et al. (2012); Avci and Aran (2014). Sanchez-Cortes et al. (2012) used nonverbal behaviors (both audio and visual modalities) to automatically identify emergent leaders in small group scenarios. Avci and Aran (2014) studied the relationship of a group's performance with the interaction between group members and the individuals' personality traits using the audio and visual nonverbal behaviors. Hung and Gatica-Perez (2010) did group cohesion estimation by utilizing audio, visual, and audio-visual cues, such as activity of each person and motion information. Dhall et al. (2015b) and Mou et al. (2015) utilized nonverbal features including face, body and context features to analyze the group-level affect displayed by a group of people on the image. Mou et al. (2016b) analyzed the affect of individuals and group membership by using the face and body cues and reported that body behaviors showed better performance for group membership recognition. Consequently, in this work we focus on using body behaviors for group membership analysis and recognition.

Methodologies for group analysis. Group analysis can be reviewed under two different ways, i.e., individual-level analysis and group-level analysis. Individual-level analysis refers to behavior analysis of each individual member of the group, while group-level analysis refers to the collective analysis of the whole group. An important issue for analysis in group settings is whether the analysis should be at group-level or individual level Reiter-Palmon et al. (2017). Trust is typically in individual-level Reiter-Palmon et al. (2017), while cohesion is in group-level. However, emotion analysis can be both in individual-level Sariyanidi et al. (2015) and group-level Barsade and Gibson (1998). In terms of individual-level analysis, Gallagher and Chen (2009) proposed a framework to recognize the attributes of individuals from group images, i.e., age and gender. Mou et al. (2016b) proposed a framework for individual affect analysis in group videos along arousal and valence. Leite et al. (2015) studied the individual engagement estimation in group settings in the context of human-robot interaction. Celiktutan et al. (2017) investigated the individual personality and engagement in dyadic interaction and human-robot interactions. Hagad et al. (2011) automatically predicted the individual rapport in dyadic interactions based on posture and congruence. In light of group-level analysis, a large number of works focus on group-level analysis. For example, Salas et al. (2015) studied group cohesion and Smith et al. (2007) studied group emotion from a social psychological perspective. Pioneering works on the study of automatic group-level emotion analyzed the overall affect displayed by the whole group Dhall et al. (2012, 2015a,b); Mou et al. (2015); Huang et al. (2015). In addition, some previous works on group-level analysis focused on group activity recognition Lan et al. (2012a,b). However, to the best of our knowledge, there are no existing works focusing on automatic recognition of group membership.

3. Proposed framework

In this work, we propose a novel framework for the recognition of group membership in group videos by analyzing body behaviors. The framework is illustrated in Fig. 1. We present a novel deep learning based *specific recognition model* built upon a *generic recognition model*. The Deep Specific

Recognition Model (DeepSRM) is shown in Eq. 1.

$$\begin{aligned} \mathcal{P}_{\text{deep-specific}}: \quad & \min_{\substack{\mathbf{w}_0, b_0 \\ \mathbf{w}_j, b_j, j=1,\ldots,4}} \frac{\lambda}{2}\|\mathbf{w}_0\|^2 + + \sum_{j=1}^{4} \left(\frac{\mu_j}{2}\|\mathbf{w}_j\|^2 + \frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2 \right) \\ & + \frac{1}{\ell}\sum_{i=1}^{\ell} \mathcal{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i)) + \frac{1}{\ell_t}\sum_{j=1}^{4} \left(\sum_{(\mathbf{x}_{it}, z_{it}) \in \mathcal{X}_t} \mathcal{L}(\mathbf{w}_j, b_j; (\mathbf{x}_{it}, z_{it})) \right), \end{aligned} \tag{1}$$

we denote this training set as $\mathcal{X} = \{(\mathbf{x}_i, z_i), i = 1, \ldots, \ell\}$, where $\mathbf{x}_i$ denotes the feature representation of the i-th training sample and z_i, the corresponding ground truth label, being equal to $+1$ if the sample belongs to the respective class, or -1 otherwise. Where $\mathcal{X}_t = \{(\mathbf{x}_{it}, z_{it}), it = 1, \ldots, \ell_t\}$ is a subset of the original training set, $\mathbf{w}_0$, b_0, $\mathbf{w}_j$, b_j are the optimization parameters (for the generic and the j-th specific model, respectively), λ, μ_j, and ν_j, $j = 1, \ldots, 4$ are regularization hyper-parameters, and $\mathcal{L}$ denotes the hinge-loss.

The *generic recognition model* uses all data across all conditions (i.e., "horror", "comedy", "action", and "adventure"), shown as the generic SVM layer in Fig. 1. The details of the *generic recognition model* is illustrated in section 3.1. The *specific recognition model* is specific to one specific condition, i.e., "horror", "comedy", "action", or "adventure", which (1) utilizes data from only one specific condition, (2) is built based on the *generic recognition model* and (3) is trained jointly with the *generic recognition model*. The details of the *specific recognition model* is illustrated in section 3.2. As the data across different conditions 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*.

In our previous work Mou et al. (2017), the *generic recognition model* and the various *specific recognition models* were trained separately. Specifically, we first trained a *generic recognition model*, obtaining an optimal value of the parameter $\mathbf{w}_0$, and then we trained a set of *specific recognition models* based on the optimized *generic recognition model*. In this paper, we propose a novel end-to-end approach to train the *generic recognition model* and all the *specific recognition models* simultaneously, simplifying the whole procedure significantly.

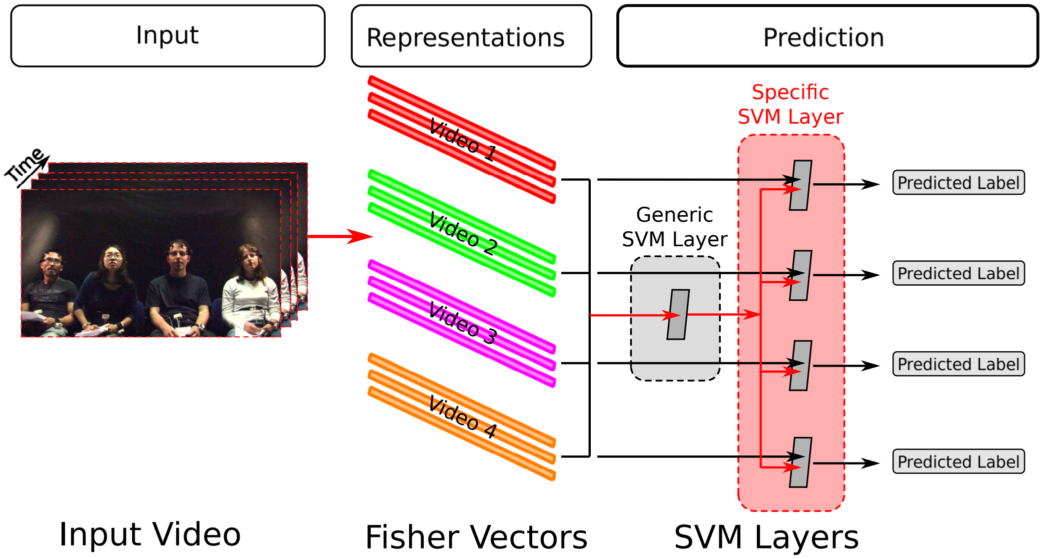


Figure 1: An illustration of the proposed framework. It consists of three parts, i.e., input, representations, and prediction. The prediction part contains SVM layers, both generic SVM layer and the specific SVM layers. In this way, we learn the *generic recognition model* in generic SVM layer and learn the *specific recognition model* in specific SVM layer. For the *specific recognition model*, as we have four different conditions, we have $n = 4$ specific problems and optimize them based on the optimized weight, w_0, obtained from the *generic recognition model*. More details of the computation of the loss can refer to Fig. 2.

3.1. The Generic Recognition Model

In our case, a certain condition is a certain movie. The *generic recognition model* is not taking the different conditions into consideration, but use all of the available training samples across all conditions. If we remove the terms related to the j-th specific condition from Eq. 1, it becomes the *generic recognition model*, illustrated in Eq. 2. The generic optimization problem, which we denote as $\mathcal{P}_{generic}$, can be cast as follows:

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

For solving the above optimization problem we use a SGD algorithm and we arrive at the optimal solution $(\mathbf{w}_0, b_0)$, which describes the separating hyperplane $\mathcal{H}_0\colon {\mathbf{w}_0}^\top\mathbf{x} + b_0 = 0$. Then, we use the optimal $\mathbf{w}_0$ to construct the set of *specific recognition models*, as described below.

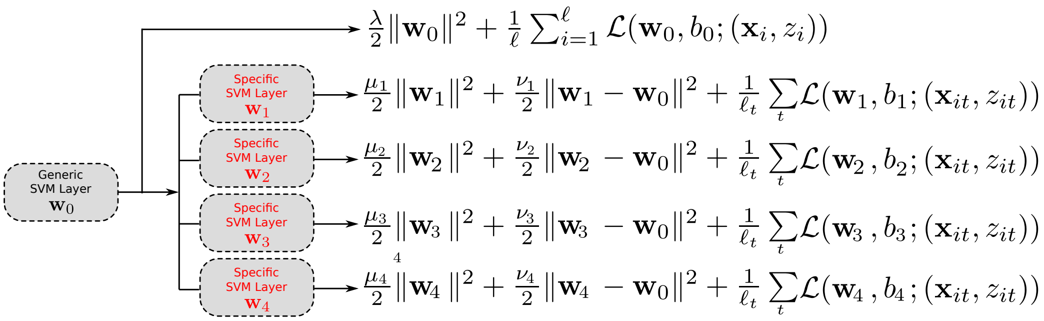

$\frac{\lambda}{2}\|\mathbf{w}_0\|^2 + \frac{1}{\ell}\sum_{i=1}^{\ell}\mathcal{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i))$
Specific SVM Layer $\mathbf{w}_1$
$\frac{\mu_1}{2}\|\mathbf{w}_1\|^2 + \frac{\nu_1}{2}\|\mathbf{w}_1 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t\mathcal{L}(\mathbf{w}_1, b_1; (\mathbf{x}_{it}, z_{it}))$
Specific SVM Layer $\mathbf{w}_2$
$\frac{\mu_2}{2}\|\mathbf{w}_2\|^2 + \frac{\nu_2}{2}\|\mathbf{w}_2 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t\mathcal{L}(\mathbf{w}_2, b_2; (\mathbf{x}_{it}, z_{it}))$
Generic SVM Layer $\mathbf{w}_0$
Specific SVM Layer $\mathbf{w}_3$
$\frac{\mu_3}{2}\|\mathbf{w}_3\|^2 + \frac{\nu_3}{2}\|\mathbf{w}_3 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t\mathcal{L}(\mathbf{w}_3, b_3; (\mathbf{x}_{it}, z_{it}))$
4
Specific SVM Layer $\mathbf{w}_4$
$\frac{\mu_4}{2}\|\mathbf{w}_4\|^2 + \frac{\nu_4}{2}\|\mathbf{w}_4 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t\mathcal{L}(\mathbf{w}_4, b_4; (\mathbf{x}_{it}, z_{it}))$

Figure 2: Illustration of the computation of the loss.

3.2. The Specific Recognition Model

A *specific recognition model* is specific to a certain condition, i.e., "horror", "comedy", "action", or "adventure", which is denoted by j in equation 3. Each *specific recognition model* is built using the outputs obtained from the *generic recognition model*. That is, we use the value for $\mathbf{w}_0$ from Eq. 2 in order to construct the specific optimization problem. The j-th condition is denoted as $\mathcal{P}^j_{\text{specific}}$ and cast as follows

$$\begin{aligned}\mathcal{P}^j_{\text{specific}}:\ \min_{\mathbf{w}_j, b_j}\ & \frac{\mu_j}{2}\|\mathbf{w}_j\|^2 + \frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2 \\ & + \frac{1}{\ell_t}\sum_{(\mathbf{x}_i, z_i)\in\mathcal{X}_t}\mathcal{L}(\mathbf{w}_j, b_j; (\mathbf{x}_i, z_i)),\ j = 1, \ldots, 4\end{aligned} \tag{3}$$

where $\mathcal{X}_t$ is a subset of the original training set, μ_j and ν_j are regularization parameters, and $\mathcal{L}$ denotes the hinge-loss. The term $\frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2$ is used to bias $\mathbf{w}_j$ to be close to $\mathbf{w}_0$. The method of solving $\mathcal{P}^j_{\text{specific}}$ is proposed in Mou et al. (2017), and attached to the appendix.

As shown in the above optimization problem, besides the standard regularization scheme, where we try to constrain the norms of $\mathbf{w}_j$ and $\mathbf{w}_0$ so as to prevent overfitting, we also add the term $\frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2$ so as to bias $\mathbf{w}_j$ to be close to $\mathbf{w}_0$, for all $j = 1, \ldots, 4$.

3.3. Feature Extraction

3.3.1. Low-level Feature Extraction

Some previous works showed that body features outperform facial features for group membership recognition Mou et al. (2016b); therefore, we use the

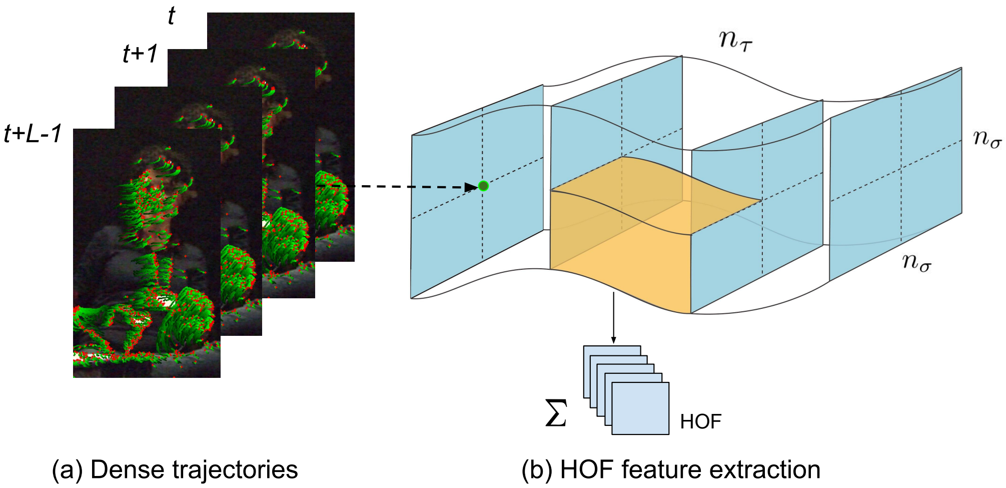
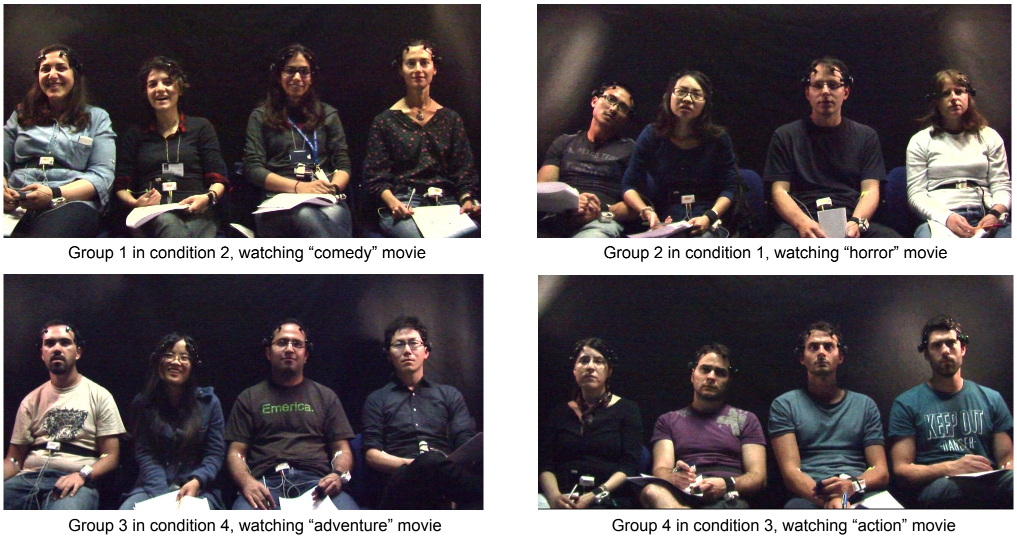


Figure 3: Illustration of the approach to extract the body HOF feature. (a) Dense trajectory detection results. (b) Dense trajectory is in the spatial scale over L frames. Motion information over a local neighborhood of $N \times N$ pixels along the each trajectory point are extracted. In order to embed the structure information, the local volume is subdivided into a spatio-temporal grid of size $n_\tau \times n_\sigma$. Based on Wang et al. (2013), $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$.

body features in this work. In order to extract person-based representations, we first apply a person detector. Constrained by our experimental setups - a fixed number of people in the video and a static camera, we use an ad-hoc scheme that is to equally divide the frame in four parts. 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 Wang et al. (2013) are extracted and, subsequently, Histograms of Optical Flow (HOF) descriptors are obtained around each trajectory. HOF descriptors are computed in the spatio-temporal volume aligned with the trajectories as shown in Fig. 3. 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 Wang et al. (2013).



Group 1 in condition 2, watching “comedy” movie

Group 2 in condition 1, watching “horror” movie

Group 3 in condition 4, watching “adventure” movie

Group 4 in condition 3, watching “action” movie

Figure 4: Representative frames from the database under different conditions.

3.3.2. Fisher Vector Encoding

Fisher Vector (FV) encoding Sánchez et al. (2013) has been widely used in computer vision problems, such as action recognition Wang et al. (2013) and depression analysis Jain et al. (2014); Dhall and Goecke (2015). 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 (FV) from body HOF descriptors.

4. Experiments and analysis

4.1. 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 Abdon Miranda-Correa et al. (2017). They were arranged into four groups with four participants in each group watching all of the four videos listed in Table 1 together. Videos were recorded at 1280 × 720 resolution, 25fps. Four representative frames from the database are shown in Fig. 4. Here we use two sub-datasets from the full database, namely Data-I and Data-II, which are different in terms of the number of samples and the method of getting the small clips from the long videos.

Data-I. It includes data from three groups (eleven subjects) with recordings in four different conditions (N1, P1, B1 and U1 movies, see Table 1). As a result, there were eleven subjects from eleven recordings in total. During each recording, each group watched one movie. From each recording, we used the last 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 $(12\times4\times3)+(12\times4\times3)+(9\times4\times3)+(7\times4\times3) = 480$.

Data-II. This dataset contains data from four groups (with sixteen participants, 8 females and 8 males) with recordings under four different conditions (N1, P1, B1 and U1 movies, see Table 1). As a result, there were sixteen subjects from fifteen recordings in total, that is 2 groups (12 subjects) with recordings from 4 movies (N1, P1, B1 and U1), 1 group (3 subjects) with recordings from 4 movies and 1 group (4 subjects) with recordings from 3 movies (B1, N1 and P1). During each recording, each group watched one movie. Each recording was segmented into 20-seconds clips with no overlap between the clips. Each clip was used as a single sample. The number of short clips from each recording varies with the length of the movies, i.e., 70 clips for N1 and B1, 56 clips for P1, and 42 clips for U1. Therefore, the total number of clips we used in the experiments is $(70 \times 4 \times 4) + (70 \times 4 \times 4) + (56 \times 4 \times 3) + (42 \times 4 \times 4) = 3584$.

4.2. Experiments

4.2.1. Implementation details

The network is implemented using Theano Theano Development Team (2016) and Lasagne Dieleman et al. (2015) libraries. All the parameters of the network, i.e., for the generic SVM layer and the four specific SVM layers (see Fig. 1), are learned using the standard back-propagation technique.

Table 3: **Data-I** tested on both SRM and DeepSRM with group membership recognition results obtained using different models, the proposed *specific recognition model*, *generic recognition model* and *independent recognition model*. 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 = 33% are also provided. ACC refers to recognition accuracy.

Different Models	Acc (p-value) chance level = 33% **SRM**	Acc (p-value) chance level = 33% **DeepSRM**
Generic recognition model ($\nu \to \infty$)	34% (p=0.42)	34% (p=0.58)
Independent recognition model ($\nu = 0$)	33% (p=0.45)	33% (p=0.52)
Specific recognition model	**42% (p<0.05)**	**40% (p<0.05)**

4.2.2. Experimental setup

We used both Data-I and Data-II to test our models. On one hand, we compared the proposed *specific recognition model* with two other models, (1) the *generic recognition model* that trained across all different conditions and (2) the *independent recognition model* that trained directly in each specific condition. We also compared this new framework (DeepRSM) to the framework proposed in our previous work Mou et al. (2017) (we refer this two-phase *specific recognition model* in the rest of the paper as SRM).

In order to avoid subject-dependency problem, group membership recognition models were trained by applying *leave-one-subject-out* cross-validation. *Leave-one-subject-out* refers to, in each fold, using eleven subjects for training-validation and the remaining one subject for testing. Each time the parameters of the model were optimized over the training-validation samples. The experimental results of the membership recognition were evaluated by the recognition accuracy. In addition, we performed statistical significance analysis to see the significance of the results obtained.

4.2.3. Experimental results and analysis

The recognition results in terms of recognition accuracy by applying *leave-one-subject-out* cross-validation are shown in Table 3 and Table 4. From both Table 3 and 4, we can clearly see that the proposed *specific recognition model*

Table 4: **Data-II** tested on both SRM and DeepSRM with group membership recognition results obtained using different models, the proposed *specific recognition model*, *generic recognition model* and *independent recognition model*. 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 = 25% are also provided. ACC refers to recognition accuracy.

Different Models	Acc (p-value) chance level=25% **SRM**	Acc (p-value) chance level=25% **DeepSRM**
Generic recognition model ($\nu \rightarrow \infty$)	26% (p=0.79)	26% (p=0.50)
Independent recognition model ($\nu = 0$)	30% (p=0.09)	30% (p=0.07)
Specific recognition model	**38% (p<0.05)**	**44% (p<0.05)**

outperforms the other two models in terms of recognition accuracy under both SRM and DeepSRM setups. Recognition accuracy of 43% is obtained for the *specific recognition model* with Data-I tested on two-phase SRM, while 34% and 33% are obtained from *generic recognition model* and *independent recognition model* respectively. A recognition accuracy of 40% is obtained for the *specific recognition model* with Data-I tested on DeepSRM, while 34% and 33% are obtained from *generic recognition model* and *independent recognition model* respectively. A recognition accuracy of 38% is obtained for the *specific recognition model* with Data-II tested on two-phase SRM, while 26% and 30% are obtained from *generic recognition model* and *independent recognition model* respectively. A recognition accuracy of 44% is obtained for the *specific recognition model* with Data-I tested on two-phase SRM, while 26% and 30% are obtained from *generic recognition model* and *independent recognition model* respectively. We also perform a t-test to see the statistical significance, which is also listed in Table 3 and 4. The statistical significance tests show 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*.

We also compared the performance obtained with the *specific recognition model* between the two-phase SRM and the DeepSRM. As we tested the models using different data and the chance levels are different, it is difficult to

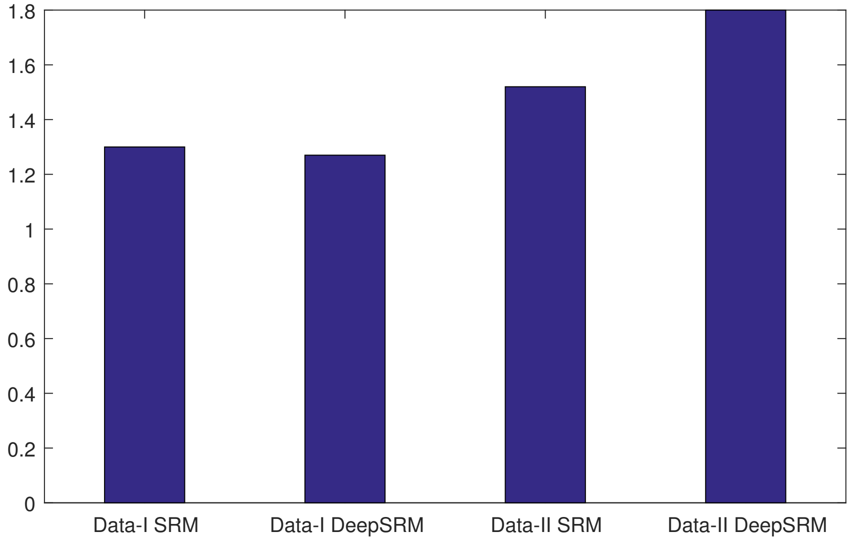


Figure 5: The recognition accuracy obtained from the proposed *specific recognition model* divided by the chance level for all different setups, i.e., Data-I tested on two-phase SRM and DeepSRM models respectively as well as Data-II tested on two-phase SRM and DeepSRM models respectively.

compare them directly. Therefore, we divided the recognition accuracy by the corresponding chance level and the results are illustrated in Fig. 5. From Fig. 5, we can see that the DeepSRM tested with Data-II performs better than two-phase SRM, while for Data-I two-phase SRM performs better than the DeepSRM. It is possibly because that compared to the non-deep framework, i.e., two-phase SRM, more parameters are learned at once while training the deep neural network, therefore, more data is needed to train the DeepSRM. In our experiments, in Data-I, there are 480 samples, while in Data-II, there are 3,584 samples, which is more than 7 times as many as Data-I. We can see that the best performance is obtained from the deep framework tested with Data-II, which outperforms the two-phase framework while tested with both Data-I and Data-II. In addition, the deep framework can be trained more easily compared to the non-deep framework, which needs to be trained by two steps, first *generic recognition model* and then *specific recognition model*. However, the deep framework can be trained in one step, which can simplify the problem in terms of implementation but provide better results. The computational cost for training two-phase SRM and DeepSRM models in terms of time is 28570 seconds and 4050 seconds for Data-II respectively while implementing on a computer with with 32G RAM and Intel Core i7-4790S CPU. Although the cost is much lower for DeepSRM, we have to bear in mind that they are not directly comparable as DeepSRM has been trained in a GPU mode, a Titan X GPU used.

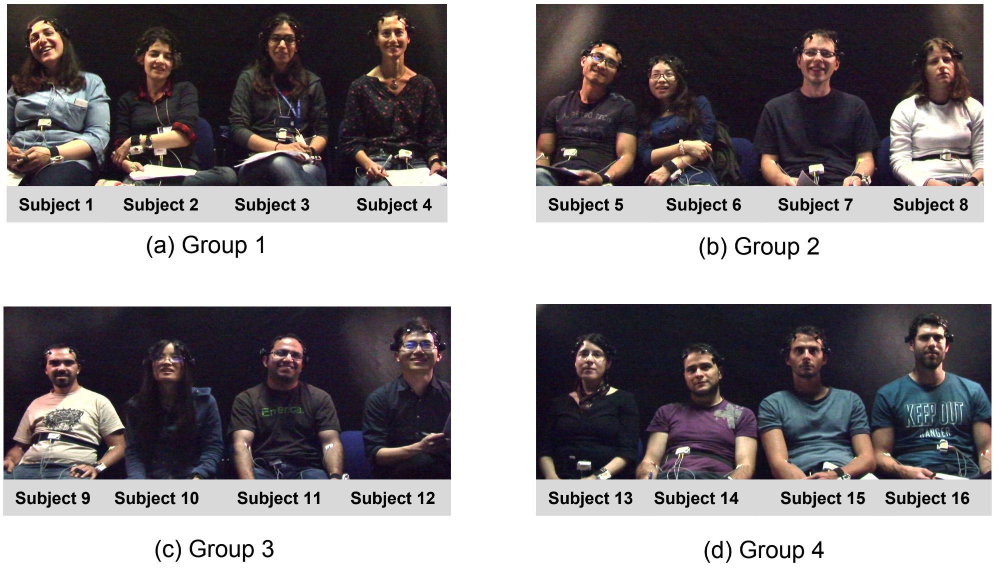
The recognition accuracy of different subjects is illustrated in Table 5. The corresponding subject can be found in Fig. 6 based on the subject ID. From Table 5, we can see that the recognition accuracy of the group membership varies among different subjects. For example, the membership of subjects 1, 3, 4 and 16 is better recognized than that of subjects 8 and 9. For subject 8, we can see from Fig. 6 (b) that subject 8 showed a very different behavior from the other group members. Specifically, we can see that subjects 6, and 7 seemed to be very happy or excited and tend to move a lot, but not subject 8. Thus, in this case, it is difficult to recognize the group membership of subject 8, which also causes difficulties in membership recognition of the other group members. The results could be due to the fact that she did not like this movie. Therefore, in order to improve the recognition accuracy of the group membership, in addition to the performance obtained for each participant in the video, it is also helpful to have some contextual information, such as the movie preference of each subject. In addition, in group 3, subjects 10, 11 and 12 were friends and classmates prior to participating in the experiments. However, subject 9 is new to this group and in this case, he was possibly sharing less non-verbal cues with the other three group members. Considering this, data should be collected with people that are unacquainted prior to attending the recording as has been done in Girard et al. (2017). Table 6 shows the average recognition accuracy under different conditions/movies. From this table, we can see that the performance varies among different movies. The results for Mr. Beans movie are the best and this is possibly due to Mr. Beans being a comedy, i.e., it can easily induce positive affect in different subjects, which is consistent with the findings of Bhullar (2012) that the more positive our mood, the more likely are to be susceptible to the happiness of others. Table 6 shows the average recognition accuracy under different conditions/movies. From this table, we can see that the performance varies among different movies. Mr. Beans presents the best result, it is possibly because that Mr. Beans is a comedy and can easily induce the positive affect for different subjects, which is consistent with the findings of Bhullar (2012) that the more positive our mood, the more likely are we to be susceptible to the happiness of others.

Table 5: The accuracy of group membership recognition of each subject

subject ID	1	2	3	4	5	6	7	8
Accuracy	0.82	0.30	0.71	0.63	0.28	0.30	0.31	0.21
subject ID	9	10	11	12	13	14	15	16
Accuracy	0.16	0.43	0.40	0.52	0.50	0.37	0.35	0.82

Table 6: The average recognition accuracy under different conditions, i.e., while people are watching different movies

Movie	Descent	Mr. Bean	Batman	Up
Accuracy	0.37	0.56	0.41	0.41



(a) Group 1

(b) Group 2

(c) Group 3

(d) Group 4

Figure 6: Four illustrative frames from four groups of data and the ID of each subject.

5. Conclusions and future work

In this paper, we propose a novel *specific recognition model* that is learned jointly with a *generic recognition model* for the problem of group membership recognition, using non-verbal behaviors of each group's member, under different conditions, i.e., when people are watching different types of movies (i.e., "horror", "comedy", "action", and "adventure"). The *generic recognition model* is trained using all data across conditions, which allows for group membership recognition across all different conditions. However, since group members may behave distinctly in different conditions, the performance of *generic recognition model* is limited. To address this, we propose a *specific recognition model* for each specific condition built on the top of the *generic recognition model*, so as to use the *generic recognition model* to provide a baseline. We conduct a set of experiments for group membership recognition on two datasets that include different groups, with each group comprising four participants watching affective stimuli.

The experimental results show that the proposed *specific recognition model* outperforms the compared approaches, i.e., *generic recognition model* and *independent recognition model*, as shown in our previous work Mou et al. (2017). However, compared to Mou et al. (2017), the newly proposed DeepSRM can be trained at once by learning both the *generic recognition model* and all the *specific recognition models* simultaneously, rather than learning them separately. In this way, the framework for DeepSRM is simplified, while at the same time its performance is improved when there is sufficient data. On the other hand, as group membership can be recognized using non-verbal behaviors (i.e., body behaviors), our results indicate that individuals affect each other's behaviors 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 the future work, we plan to experiment with other feature representation. It is also important to use different contextual information to assist the recognition process, such as personality, movie preference, and the personal relationships between group members. In addition, we also plan to apply this learning approach to other recognition problems.

6. 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 EUs Horizon 2020 programme under grant agreement H2020-693092 MOVING.

7. REFERENCES

Abdon Miranda-Correa, J., Khomami Abadi, M., Sebe, N., Patras, I., 2017. AMIGOS: A Dataset for Mood, Personality and Affect Research on Individuals and Groups. ArXiv e-prints. 2, 13

Allen, J. A., Fisher, C., Chetouani, M., Chiu, M. M., Gunes, H., Mehu, M., Hung, H., 2017. Comparing social science and computer science workflow processes for studying group interactions. Small Group Research. 4

Avci, U., Aran, O., 2014. Effect of nonverbal behavioral patterns on the performance of small groups. In: Proc. of ACM workshop on Understanding and Modeling Multiparty, Multimodal Interactions. 6

Barsade, S. G., 2002. The ripple effect: Emotional contagion and its influence on group behavior. Administrative Science Quarterly. 2, 6

Barsade, S. G., Gibson, D. E., 1998. Group emotion: A view from top and bottom. Composition. 7

Bhullar, N., 2012. Relationship between mood and susceptibility to emotional contagion: is positive mood more contagious? North American Journal of Psychology. 17

Celiktutan, O., Skordos, E., Gunes, H., 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. IEEE Tran. on Affective Computing. 5, 6, 7

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Tran. on Intelligent Systems and Technology. 25

Dhall, A., Goecke, R., 2015. A temporally piece-wise fisher vector approach for depression analysis. In: Proc. of Affective Computing and Intelligent Interaction (ACII). 12

Dhall, A., Goecke, R., Gedeon, T., 2015a. Automatic group happiness intensity analysis. IEEE Tran. on Affective Computing. 5, 6, 7

Dhall, A., Joshi, J., Radwan, I., Goecke, R., 2012. Finding happiest moments in a social context. In: Proc. of Asian Conf. on Computer Vision (ACCV). 7

Dhall, A., Joshi, J., Sikka, K., Goecke, R., Sebe, N., 2015b. 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). 5, 6, 7

Dieleman, S., Schlter, J., Raffel, C., Olson, E., Snderby, S. K., Nouri, D., et al., Aug. 2015. Lasagne: First release.
URL http://dx.doi.org/10.5281/zenodo.27878 13

Eberl, M., 2010. An application of pls in multi-group analysis: The need for differentiated corporate-level marketing in the mobile communications industry. Handbook of partial least squares. 4

Gallagher, A. C., Chen, T., 2009. Understanding images of groups of people. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR). 2, 5, 7

Girard, J. M., Chu, W.-S., Jeni, L. A., Cohn, J. F., 2017. Sayette group formation task (gft) spontaneous facial expression database. In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG). 4, 5, 17

Goette, L., Huffman, D., Meier, S., 2006. The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. 4

Hagad, J. L., Legaspi, R., Numao, M., Suarez, M., 2011. Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence. In: Proc. of IEEE Int. Conf. on Social Computing. 2, 7

Huang, X., Dhall, A., Zhao, G., Goecke, R., Pietikäinen, M., 2015. Riesz-based volume local binary pattern and a novel group expression model for group happiness intensity analysis. In: Proc. of British Machine and Vision Conference (BMVC). 5, 7

Hung, H., Gatica-Perez, D., 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. IEEE Tran. on Multimedia. 5, 6

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

Jain, V., Crowley, J. L., Dey, A. K., Lux, A., 2014. Depression estimation using audiovisual features and fisher vector encoding. In: Proc. Int. Workshop Audio/Visual Emotion Challenge. 12

Lan, T., Sigal, L., Mori, G., 2012a. Social roles in hierarchical models for human activity recognition. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR). 7

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., Mori, G., 2012b. Discriminative latent models for recognizing contextual group activities. IEEE Tran. on Pattern Analysis and Machine Intelligence. 7

Lehmann-Willenbrock, N., Hung, H., Keyton, J., 2017. New frontiers in analyzing dynamic group interactions: Bridging social and computer science. Small Group Research. 4

Leite, I., McCoy, M., Ullman, D., Salomons, N., Scassellati, B., 2015. Comparing models of disengagement in individual and group interactions. In: Proc. of ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI). 2, 4, 5, 7

Mou, W., Celiktutan, O., Gunes, H., 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 (FG). 2, 5, 6, 7

Mou, W., Gunes, H., Patras, I., 2016a. Alone versus in-a-group: A comparative analysis of facial affect recognition. In: Proc. of ACM Conf. on Multimedia (ACMMM). 2, 5

Mou, W., Gunes, H., Patras, I., 2016b. Automatic recognition of emotions and membership in group videos. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition and Workshops (CVPRW). 2, 5, 6, 7, 10

Mou, W., Tzelepis, C., Mezaris, V., Gunes, H., Patras, I., 2017. Generic to specific recognition models for membership analysis in group videos. In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG). 3, 4, 8, 10, 14, 19, 24

Platt, J., et al., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 25

Reiter-Palmon, R., Sinha, T., Gevers, J., Odobez, J.-M., Volpe, G., 2017. Theories and models of teams and groups. Small Group Research. 4, 7

Salas, E., Grossman, R., Hughes, A. M., Coultas, C. W., 2015. Measuring team cohesion observations from the science. Human Factors: The Journal of the Human Factors and Ergonomics Society. 7

Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J., 2013. Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision (IJCV). 12

Sanchez-Cortes, D., Aran, O., Mast, M. S., Gatica-Perez, D., 2012. A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Tran. on Multimedia. 4, 6

Sariyanidi, E., Gunes, H., Cavallaro, A., 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Tran. on Pattern Analysis and Machine Intelligence (PAMI). 7

Saxena, S., Brémond, F., Thonnat, M., Ma, R., 2008. Crowd behavior recognition for video surveillance. In: Advanced Concepts for Intelligent Vision Systems. 4

Smith, E. R., Seger, C. R., Mackie, D. M., 2007. Can emotions be truly group level? evidence regarding four conceptual criteria. Journal of personality and social psychology. 4, 7

Theano Development Team, May 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688. URL http://arxiv.org/abs/1605.02688 13

Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M., Murino, V., 2016. Detecting conversational groups in images and sequences: A robust game-theoretic approach. Computer Vision and Image Understanding. 2

Wang, H., Kläser, A., Schmid, C., Liu, C.-L., 2013. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision (IJCV). 11, 12

Zhang, L., Hung, H., 2016. Beyond f-formations: Determining social involvement in free standing conversing groups from static images. In: Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR). 2

8. APPENDIX

For solving $\mathcal{P}^j_{\text{specific}}$, as proposed in Mou et al. (2017), 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 $\mathbf{w}_j^{(1)}$ to any vector whose norm is at most $1/\sqrt{\nu_j}$ and $b_j^{(1)} = 0$. 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 $\eta_t = \frac{1}{\nu_j t}$. We approximate the objective function of $\mathcal{P}^j_{\text{specific}}$ with

$$\begin{aligned}\mathcal{P}^j_{\text{specific}}:\ \mathcal{J}(\mathbf{w}_j, b_j) = \frac{\mu_j}{2}\|\mathbf{w}_j\|^2 + \frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2 \\ + \frac{1}{k}\sum_{(\mathbf{x}_i, z_i)\in X_t} \mathcal{L}(\mathbf{w}_j, b_j; (\mathbf{x}_i, z_i)),\ j = 1, \ldots, 4.\end{aligned} \tag{4}$$

The update rules are given as follows

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

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

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

and

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

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

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

and

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

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

Once the optimal values of the parameters $\mathbf{w}_j$ and b_j 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 $y_t = \operatorname{sgn}(d_t)$, where $d_t = \mathbf{w}_j^\top \mathbf{x}_t + b_j$. 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 Platt et al. (1999) 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 Chang and Lin (2011)) for evaluating a sample's class membership at the testing phase.