Automatic Prediction of Group Cohesiveness in Images

Shreya Ghosh, *Student Member, IEEE*, Abhinav Dhall, *Member, IEEE*, Nicu Sebe, *Senior Member, IEEE*, and Tom Gedeon, *Senior Member, IEEE*

Abstract—This paper discusses the prediction of cohesiveness of a group of people in images. The cohesiveness of a group is an essential indicator of the emotional state, structure and success of the group. We study the factors that influence the perception of group-level cohesion and propose methods for estimating the human-perceived cohesion on the group cohesiveness scale. To identify the visual cues (attributes) for cohesion, we conducted a user survey. Image analysis is performed at a group-level via a multi-task convolutional neural network. A capsule network is explored for analyzing the contribution of facial expressions of the group members on predicting the Group Cohesion Score (GCS). We add GCS to the Group Affect database and propose the 'GAF-Cohesion database'. The proposed model performs well on the database and achieves near human-level performance in predicting a group's cohesion score. It is interesting to note that group cohesion as an attribute, when jointly trained for group-level emotion prediction, helps in increasing the performance for the later task. This suggests that group-level emotion and cohesion are correlated. Further, we investigate the effect of face-level similarity, body pose and subset of a group on the task of automatic cohesion perception.

Index Terms—Group-level emotion, Cohesion estimation.

1 Introduction

The concept of 'teamwork' is defined as the collaborative effort of a group of people to accomplish a common goal in the most well-organized way [54]. One of the most important requirements for effective teamwork is cohesion. Group cohesiveness can be defined as a bonding which affects the membership of an individual in a group. The main motivation behind group's cohesiveness is the interpersonal attraction between the group members, group's pride, commitment to the task of the group etc. Cohesiveness is the most important attribute of a successful group [21]. The positive consequences of group cohesiveness are more participation, more conformity, high productivity, more success and more personal level satisfaction [67] etc. The main motivation of our work is understanding the human perception of Group Cohesiveness Score (GCS) [66] from images and mapping the attributes to an Automatic Group Cohesion (AGC) pipeline. Group cohesiveness is defined as the measure of bonding between group members. Higher cohesiveness implies stronger group-level bonding. According to psychological studies, group cohesion depends on several factors such as members' similarity [63], group size [9], group success [73], external competition and threats [65], [52]. The reason behind a strong group bonding can be positive (e.g. group success) or negative (e.g. threats). Cohesion plays an important role in group-level success [4] and it affects the group-level performance. Beal et al. [4] argue that group cohesion plays the most important role in group performance. Similarly, group members' satisfaction [32] also plays an important role in deciding the cohesiveness of a group.

Fig. 1: The group of people in the left and the right images have high and low perceived group cohesion scores, respectively.

Hackman et al. [32] state that members belonging to a cohesive group have more satisfaction as compared to a non-cohesive group. Myers [50] indicates that people belonging to a cohesive group are less prone to anxiety and tension. Lott et al. [45] found that group cohesion helps improve individual members' learning processes. In group dynamics, synchronization of group members' mentality is the stepping stone of group formation [56]. In the next step, group members may realize if the emotional ties are strong enough to hold them together then it will influence the group's performance. This emotional bonding is called *cohesion* of a group. Mainly, group affect can be manifested as the convergence in individual group members' emotional state and behaviour [18]. In the existing cognitive science literature [56], different phases of a group have been proposed. In other words, phases of a group are an affective experience which is shared or held in common, by the members of a group. Inspired by the aforementioned studies, in this work we are interested in investigating the following research questions:

- *How useful are holistic, facial, group structure and body pose information for predicting cohesion in a group?*
- *What are the factors that affect the perception of the cohesiveness in a group?*

- *S. Ghosh is with Monash University. (E-mail: shreya.ghosh@monash.edu).*
- *A. Dhall is with Monash University and Indian Institute of Technology Ropar, India. (E-mail: abhinav.dhall@monash.edu)*
- *N. Sebe is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (E-mail: niculae.sebe@unitn.it).*
- *T. Gedeon is with the Department of Computer Science, Australian National University, Canberra, Australia (E-mail: tom@cs.anu.edu.au).*

Project Page: https://sites.google.com/view/grouplevelaffect/home
Manuscript submitted September 11, 2019; accepted September 6, 2020.

Fig. 2: Group cohesion scale as defined by Treadwell et al. [66].

- *What is the usefulness of cohesiveness as an attribute for tasks such as group emotion prediction?*

In this work, we investigate AGC from an early prediction perspective. This can also be viewed as a first impression of a group's cohesion, similar to the early personality assessment [51] problem in affective computing. This manuscript subsumes Ghosh et al. [27]. The major changes are as follows: a) We study the effect of face level similarity (Section 6.2.2), group structure (Section 6.4), body pose (Section 6.5) and subset of faces (Section 6.6); b) We elaborate the challenges and discussions in Section 3 and add data statistics (Section 5.1); c) We discuss applications relevant to this topic (Section 8). The main contributions of this paper are as follows:

1) *To the best of our knowledge, this is the first study proposing AGC prediction in images;*
2) *We compare several cohesion models, representing scene (holistic), face-level information, group structure and body pose respectively, and show that the former (holistic) contributes more to the perception of cohesion;*
3) *We label and extend the Group Affect Database [16] with group cohesion labels and propose the* ***GAF Cohesion database*** *(sample images from the database are shown in Figure 1);*
4) *From our experimental results, we observe that the perceived group emotion is related to group cohesiveness (Section 7).*

The rest of the paper is structured as follows: Section 2 describes the prior works on Group cohesion. Section 3 and 4 explain the challenges involved in predicting the GCS task and the procedure of our survey. Section 5 discusses the data and labeling process. The details of the proposed methods are described in Section 6. Experiments are discussed in Section 7. Section 8 describes the possible applications of this work. Conclusion, limitations and future research directions are discussed in Section 9.

2 Prior Work

2.1 Group-level Cohesion

2.1.1 Psychological Aspects

According to Barsade et al. [3], several factors impact the perception of a group's cohesion and emotion. The authors [3] argued that social norms and constraints (i.e. interpersonal bonding and individual emotional responses) are important cues for group emotion and cohesion. Gallagher et al. [23] modelled the group as a min span tree based on facial locations and inferred the gender and age of group members using the group-level contextual information. Tajfel et al. [63] stated that one of the main factors affecting a group's cohesiveness is its group members' similarity. Here, similarity can be measured in terms of their occupation, ethnicity, age, relationship etc. This may also imply that due to these factors group members may have a similar point of view about certain issues, which may cause strong bonding between them. Another interesting study by Carron et al. [9] suggested that a small group implies strong cohesion. The reason behind this is that as the number of group members increases, their opinions may vary. This may lead to weaker cohesiveness as compared to small groups. Zaccaro et al. [73] argued that group-level success (towards a task) is another factor, influencing cohesiveness, along with the group's size and its members' similarity. Apart from the positive factors, some negative factors may also influence a group's cohesiveness. Several studies [65], [52] revealed that threats to a group and competition with another group may also increase a group's cohesiveness.

In a seminal work, Hung et al. [38] studied group cohesion in a constrained environment using group meeting data. Several audio and video features were extracted to test their importance on group cohesion. For audio analysis, pauses between individual turns, pauses between floor exchanges, turn lengths, overlapping speech, prosodic cues etc. are taken into consideration. Similarly, video features include pauses between individual turns, pauses between floor exchanges, motion turn lengths, overlapping visual activity, visual energy cues, ranking participants' features, group distribution features etc. Further, a support vector machine is trained for predicting the overall cohesion score. Sharma et al. [55] proposed VGAF dataset for group-level emotion and cohesion prediction in videos. In this paper, we are interested in exploring different dimensions of group cohesion including context, facial emotion and attributes, body pose and group structure in images.

2.1.2 EmotiW 2019 Group-level Cohesion Challenge

This challenge has been organized since 2013 in ACM International Conference on Multimodal Interaction (ACM ICMI) challenge track. The main focus of this challenge is spontaneous affect analysis in varied conditions mainly in real-time. In 2019, it included group-level cohesion as a sub-challenge[1] [14]. Guo et al. [29] predict group cohesion on the basis of face, body and global image features. Xuan et al. [72] propose a hybrid deep learning network via scene, skeleton, UV coordinates and facial image features. Similarly, [75] used face, skeleton and scene features to predict group-level cohesion. From these papers, we can conclude that face, scene and body pose play an important role in group-level cohesion prediction.

1. https://sites.google.com/view/emotiw2019

2.2 Study of 'Group of People'

In recent years, computer vision researchers have analyzed 'group of people' for different tasks. In an interesting work, Chang et al. [11] predicted group-level activity via hierarchical agglomerative as well as the divisive clustering algorithm. To track the group-level activity, multiple cameras are placed in different environments (e.g. in an abandoned prison yard) which detects first group related information such as group formation, dispersion and distinct groups. Further, motion patterns (*Loitering, Fast Moving, Approaching, Following*) and behaviour (e.g. *Flanking, Agitation, Aggression*) were investigated. Wang et al. [69] proposed a method to infer the relationship between group members via geometric structure and appearance-based features of the group. The AMIGOS database [46] has been recently proposed to study different aspects of affect in a group. Similarly, Alameda et al. [1] propose the SALSA database to study group-level personality, emotion and affect in real word settings. In summary, these studies [1], [49], [36], [69], [23], [29] motivate us to use facial, body pose, group structure features for group cohesion.

2.3 Group Emotion

One of the earlier group emotion analyses was proposed by Dhall et al. [15]. They proposed the Group Expression Model (GEM) to predict happiness intensity of a group of people in images. Several other studies [34], [37], [41], [61], [70], [30], [26] mainly extracted scene, face and pose features to predict group emotion. Singh et al. [57] studied the effect of a group on a person's smile. They evaluated the usefulness of visual features in predicting the task. Similar to [15], automatic group-level emotion analysis approaches can be divided into three broad categories: bottom-up approach, top-down approach and hybrid approach.

2.3.1 Bottom-Up Approaches

The bottom-up approaches analyze the group members individually and then assess the contribution of these members towards the overall group's mood. Ge et al. [24] used the bottom-up hierarchical clustering algorithm to track a small group of people. However, their motivation is spreading situation awareness in real-time to help people. Hernandez et al. [34] conducted an interesting experiment, wherein the facial expression of the people passing through the corridor was analyzed for the presence of a smile. The number of smiles was averaged at a given point in time to decide the group-level mood. Vonikakis et al. [68] extracted face level geometric features based on the location of the facial part to infer the expression intensity. All these motivate us to analyze individual-level facial expression first and then pool it at the group-level for cohesion prediction.

2.3.2 Top-Down Approaches

The main motivation behind this set of approaches is to determine global factors and how these impacts the perception of a group's emotion. Dhall et al. [17] computed a scene level descriptor to encode the background information along with the facial and body cues. Huang et al. [37] modelled the group using a conditional random field and represented faces with a local binary pattern variant. Based on these works, we are using the holistic level features (scene features) to get some overall important information.

What do you think about group cohesion by looking at the picture(Image 1)? *

1. strongly agree (bonding is very strong)
2. agree (bonding is strong)
3. disagree (bonding is weak)
4. strongly disagree (bonding is very weak)

Any reason for your choice? *

Short answer text

Fig. 3: Screen shot of the user survey for understanding the factors affecting the perception of a group's cohesiveness.

2.3.3 Hybrid Approaches

Hybrid approaches use both holistic level and individual level information. Mou et al. [48] performed an interesting study of human affect on individual and group scenarios. They created three models: 1) first trained with an individual level database. 2) second trained on a database containing a group of people and 3) third a hybrid fused model of the above two. Smith et al. [58] argued that group-level emotion is different from individual emotion. To predict an individual's role in the overall group emotion, the main question is that 'how much a person is involved in the group?' i.e. what is his/her cohesiveness or bonding with the other group members?

Li et al. [41] used Long Short-Term Memory (LSTM) to encode happiness intensity. Both facial features and scene features are extracted via a deep neural network, which is further input into a LSTM for prediction. Sun et al. [61] also proposed a LSTM network for training a regression model, which achieved good results on the HAPPEI database [15]. Tan et al. [64] extracted facial expression information along with global scene information and pooled it at a global level to predict group emotion. Guo et al. [30] used scene, facial and pose information to encode group emotion. Similarly, Wei et al. [70] also used deep facial and scene features to decode group-level information.

3 Challenges

This section describes the challenges involved in designing a AGC prediction network. To design an automated system for AGC prediction, we wish to examine the factors which affect the perception of cohesion of a group. In the existing literature, the perception of members' similarity [63] is claimed to be a vital visual cue; however, the first perception after viewing an image differs considerably from person to person. As human perception of a group is very subjective and culturally biased, people generally perceive a group-level inference either top-down or bottom-up approach [40]. The top-down approach includes global context, i.e., group history, background, social event etc. All of the aforementioned attributes have an effect on the group

Fig. 4: Survey results: The first column is the image. The second column represents the word cloud of keyword responses (responses against the reason field as shown in survey form Figure 3) and the third column consists of participant responses for a group's cohesion score. (Colour code for 3^{rd} column: green= strongly agree, blue= agree, yellow= disagree, and red= strongly disagree) The fourth column shows the model prediction along with ground truth label for these images. For the 4^{th} column blue= face level prediction, red= image level prediction and orange= ground truth label). Prediction results are in the range [0-3]. In the results, the face level network predicts the level of cohesion on the basis of emotion intensity similarity (e.g. it detects smile faces across image 2 and thus it predicts it as high cohesion). Similarly, it can not predict correctly in case of 2^{nd} and 4^{th} image. [Best viewed in colour]

members. This happens because peoples' interaction may differ in different social events. For example, a person in a family environment (e.g. in a family reunion, wedding, family dinner) have a different attitude than the one in an office-level gathering. Thus, the background information can be used for the analysis of AGC. Top-down attributes mainly contain scene information, social event information, neighbour's proximity information (who is standing with whom), and relative positions (where are people standing?). On the other hand, the bottom-up component deals with individual attributes rather than the overall one. For example, individual attributes, cover an individuals mood/emotion, facial occlusion, relative face size, age, race, head pose and eye blink.

The presence of many people in the scene may lead to challenges such as more than one group formation, face tracking problems due to occlusion, illumination variations, and background variation. Another challenge is the video segment division (for temporal data) because the temporal duration of audio-video samples may vary a lot. Proper dataset collection is also another challenge. This study is an attempt to answer a non-trivial question of group-level cohesion with the help of deep learning and computer vision techniques for affective computing. For better inference of a groups cohesiveness, one can use information related to the group traits, which includes context information, facial information, body pose and other features. The AGC detection takes both top-down and bottom-up scenarios into account at the same time as they are helpful for the analysis of group dynamics and performance in real-world conditions. To further understand these attributes, we conducted a survey discussed below.

4 Survey

In our survey there are a total of 102 participants: 59 male and 43 female belonging to an age group of 22-54 years to understand the important visual cues. The participants were from different backgrounds like student, businessman, corporate employee etc. The form consists of 24 images (as shown in Figure 3) of groups of people in different contexts and having different GCS values

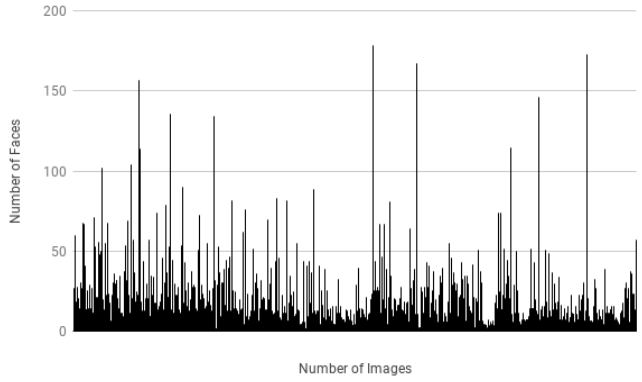

Fig. 5: This is the distribution of the number of faces with the number of images. Here, X-axis represents the number of images and Y-axis represents the number of faces in one image.

(6 images for each GCS value). Based on Treadwell et al. [66], we used four levels of cohesion. Before filling in the form, the participants were first familiarized with the concept of group cohesion labels [66] with images. The participants had to select one of the four cohesion levels for each image and they had to provide reasons behind their choice. Thus, participants provided few keywords related to the AGC score and corresponding image.

After analyzing the responses, we got the statistics as shown in Fig. 4. From the word clouds of Figure 4, we can see that 'team', 'bonding' and 'together' are the most frequent keywords indicating that we are dealing with group-level effects. Further, 'winning', 'trophy', 'work', 'scolding', 'fight' etc. reflect some holistic level features which motivate us to study image-level analysis. Similarly, some keywords such as 'happy', 'cheering', 'angry', 'violence' etc. tell about the mood of the individuals as well as the group. Thus, the survey motivates us to utilize both image-level features and face-level emotion features of an image. Our experiments are based on the understandings from the survey.

5 Database

To create the database, we used and extended the images from the GAF 3.0 database [16]. GAF 3.0 has been created via web crawling based on various keywords related to social events (e.g., *world cup winners, wedding, family, laughing club, birthday party, siblings, riot, protest*, *violence* etc.).

5.1 Data Statistics

We relabeled GAF 3.0 [19] to get a total of 17,175 images. We split the data into three parts: 9,815 images for training, 4,349 images for validation and 3,011 images for testing purposes. Further, we sorted the images with creative commons license and these were used in the EmotiW 2019 group-level cohesion sub-challenge. The updated data splits are 9,300 images for training, 4,244 images for validation and 2,899 images for testing purposes. In Figure 5, the distribution of the number of faces with the number of images is shown. The number of images is plotted along the X-axis and the number of faces corresponding to one image is plotted along Y-axis. The average number of faces in an image is 8. According to [22], small groups are more cohesive than large groups. Figure 6 shows the distribution of the number of faces with the cohesion score. Here, X-axis represents the number of faces and Y-axis represents the cohesion score.

5.2 Data Labeling

The GAF 3.0 database was labelled by 5 annotators (3 females and 2 males) of age group 21-30 years. We label each image for its cohesiveness in the range [0-3] [66]. Treadwell et al. [66] argued that it is better to have these four 'anchor points' (i.e., *strongly agree, agree, disagree* and *strongly disagree*) for annotation instead of having low to high scores. The low to high score scaling may vary perception-wise from person to person. Thus, these soft scaled 'anchor points' are reliable. We adapted this concept and label cohesion score in the range [0-3] where 0: very weak cohesion, 1: weak cohesion, 2: strong cohesion, and 3: very strong cohesion. Along with GCS, GAF 3.0 database is also labelled with three group emotions (*positive, negative and neutral*) across the valence axis. Before the annotation, the annotators have been familiarized with the concepts of GCS labels [66] with corresponding images. First of all, we conducted a tutorial with the annotators regarding concepts of cohesion. Further, we asked the annotators to label the images on the basis of a list of questions which include both social and task cohesion. The questions are a subset of the 27 questions mentioned in Hung et al. [38].

5.3 Annotation Statistics

We further investigated the agreement between the annotators. The average variance and standard deviation between the annotators was 0.31 and 0.54, respectively. Further, we computed principal component analysis on the annotations as shown in Figure 7. It is evident that approx. 86% of the distribution lies in the first component, which suggests that there is a strong agreement between the annotators. Since the annotations were based on a 'mutually exclusive category', we also measured the weighted generalized Cohen's kappa coefficient [31] to determine the inter-rater agreement. The mean of the kappa coefficients value is 0.51. This also indicates a high inter-rater agreeableness.

6 Proposed Method

We use two networks, the first examines the image as a whole and the second examines the facial expression of the group members. The details are discussed below:

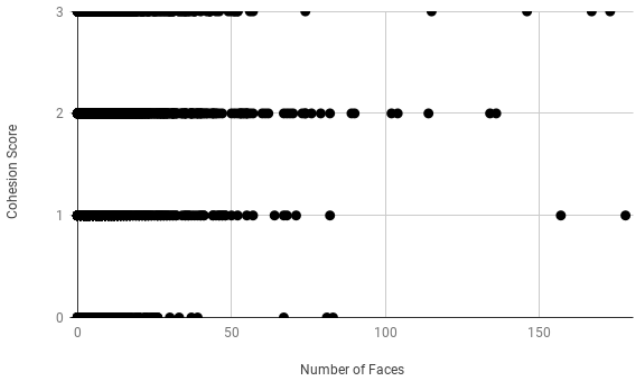

Fig. 6: This is the distribution of the number of faces with the cohesion score. Here, X-axis represents the number of faces and Y-axis represents cohesion score.

GAF 3.0	Ours	VGG-16	AlexNet
MSE, GCS	0.8181	0.8967	1.0375
Accuracy(%) Group Emotion	85.58	40.26	72.21

TABLE 1: GCS and emotion recognition results comparison.

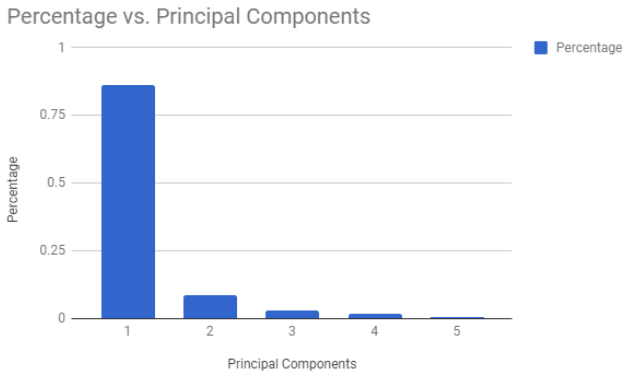

Fig. 7: The Figure shows the Eigen values for the 5 principal components inter rater variance. It is evident that the first principal component consists of 86% of the distribution.

6.1 Image-level Analysis

The motivation for this part is to collectively analyze the group and its surroundings. This is also meant to provide contextual information about the group i.e. where the group is and what type of event they are participating in. We use the Inception V3 [62] to train our model for predicting GCS. The main reason behind choosing inception V3 is that it provides a good trade-off between the number of parameters and accuracy in the case of the ImageNet challenge [62]. We have also conducted experiments on several deep Convolutional Neural Networks (CNNs), and results are shown in TABLE 1. Our inception V3 network is similar to the one in [62] which was proposed for the classification on the ImageNet task except for the last few dense layers including the regression layer. Details of the layers are shown in TABLE 2.

Based on the word cloud from the survey, we note that the participants mentioned some group-level emotion-related keywords e.g. 'violence', 'happy', 'angry' etc. Thus, we perform experiments with joint training for GCS and group emotion (three classes positive, neutral and negative [16]). The motivation is to explore the usefulness of GCS of a group as an attribute for group emotion prediction. The network structure used is shown in TABLE 2 except for the last layer, which predicts three group emotion probabilities and one GCS.

6.2 Face-Level Analysis

6.2.1 Face-Level Emotion Analysis

Motivated by the result of joint training of the AGC and group emotion and survey results, we analyse performance of GCS based on face-level affect. We used the recently proposed CapsNet [53] architecture as shown in Figure 8. In order to overcome the drawbacks of traditional CNNs, Sabour et al. [53] proposed a new CNN like architecture **Capsule Network** (CapsNet), which keeps the spatial orientation related information along with deep features. Here, capsules are a group of neurons which include the instantiation parameters of a certain object. For example, a face has eyes, nose, lips with certain constraints. The main difference between a CNN and a CapsNet is that the latter stores the state of the feature (neuron output) in the form of a vector instead of a scalar. Another salient property of CapsNet is routing by agreement, which means activated capsules follow a hierarchy. Higher level capsules become activated if and only if lower level capsule outputs agree with it. As per [53], CapsNet is invariant to rotation and it can model a spatial hierarchy via dynamic routing and reconstruction regularization. Thus, the network can learn the pattern of viewpoint invariance between the object part and the whole object. From TABLE 3, we can observe that CapsNet performs better than the other state-of-the-art networks. CapsNet can explicitly model the pose and illumination of an object. Inspired by this argument, we choose to train CapsNet. We slightly modified the proposed architecture of CapsNet [53] used for digit classification. CapsNet takes cropped face as input and predicts the seven basic emotions (i.e. *happy*, *neutral*, *sad*, *angry*, *surprise*, *disgust* and *fear*) as an output. Thus, we get emotion probability predictions for each of the faces present in a group image. Further, we pool the predicted emotion labels by computing the average, maximum and minimum (get $batchsize \times 3 \times 7$ dimensional output). This small feature is then fed to two dense layers of 16 and 32 nodes respectively before predicting the GCS. The network structure is mentioned in Table 4.

Layers	Input	Output	Layer Details
Inception V3	b,224,224,3	b,2048	similar to [62]
Dense	b,2048	b,4096	4096
Activation	b,4096	b,4096	ReLU/Swish
Dense	b,4096	b,4096	4096
Activation	b,4096	b,4096	ReLU/Swish
Dense	b,4096	b,4096	4096
Activation	b,4096	b,4096	ReLU/Swish
Cohesion (Sigmoid)	b,4096	b,1	1

TABLE 2: Image-level network architecture. Here, *b* and *BN* refer to the batch size and batch normalization respectively.

6.2.2 Face-Level similarity Analysis

Group members' similarity has different influences on group-level cohesion. Lott et al. [45] conducted experiments on this bottom-up approach and found that individual's attribute similarities (for example, race, ethnicity, occupation, age, attitudes, values and personality traits) correlate with group cohesiveness. Similarly, from the social attraction perspective, Hogg et al. [35] discussed 'similarity among group members' in the context of group cohesion. According to their study [35], similarity is the main criteria for an individual to categorize others into a group. Inspired by this argument, we perform the following experiment:

- Pass an input image of size $100 \times 100 \times 3$ to the VGG-16 network and extracted the FC6 layer feature (4096 dimensional).
- Let n be number of faces in a group and the VGG-16 FC6 layer output be d-dimensional. Calculate cosine similarity across each dimension of the faces. Thus, the resultant cosine similarity of face $f_1(1 \times d)$ and $f_2(1 \times d)$ is $cs_{1,2}(1 \times d)$. cs is defined as

$$cs_{i,j} = cosine\ similarity(f_i^k, f_j^k)$$

  here i and j represents i^{th} and j^{th} face of an image and k lies between $[1 - d]$. Cosine similarity is defined as

$$cosine\ similarity(a, b) = \frac{a.b}{\parallel a \parallel . \parallel b \parallel}$$

- With the cosine similarity matrix corresponding to an image, extract statistical features (maximum, minimum

RAF-DB	Ours	Alexnet	mSVM [42]	DLPCNN [42]
Accuracy(%)	77.48	76.27	65.12	74.20

TABLE 3: Comparison of the performance of CapsNet on RAF-DB.

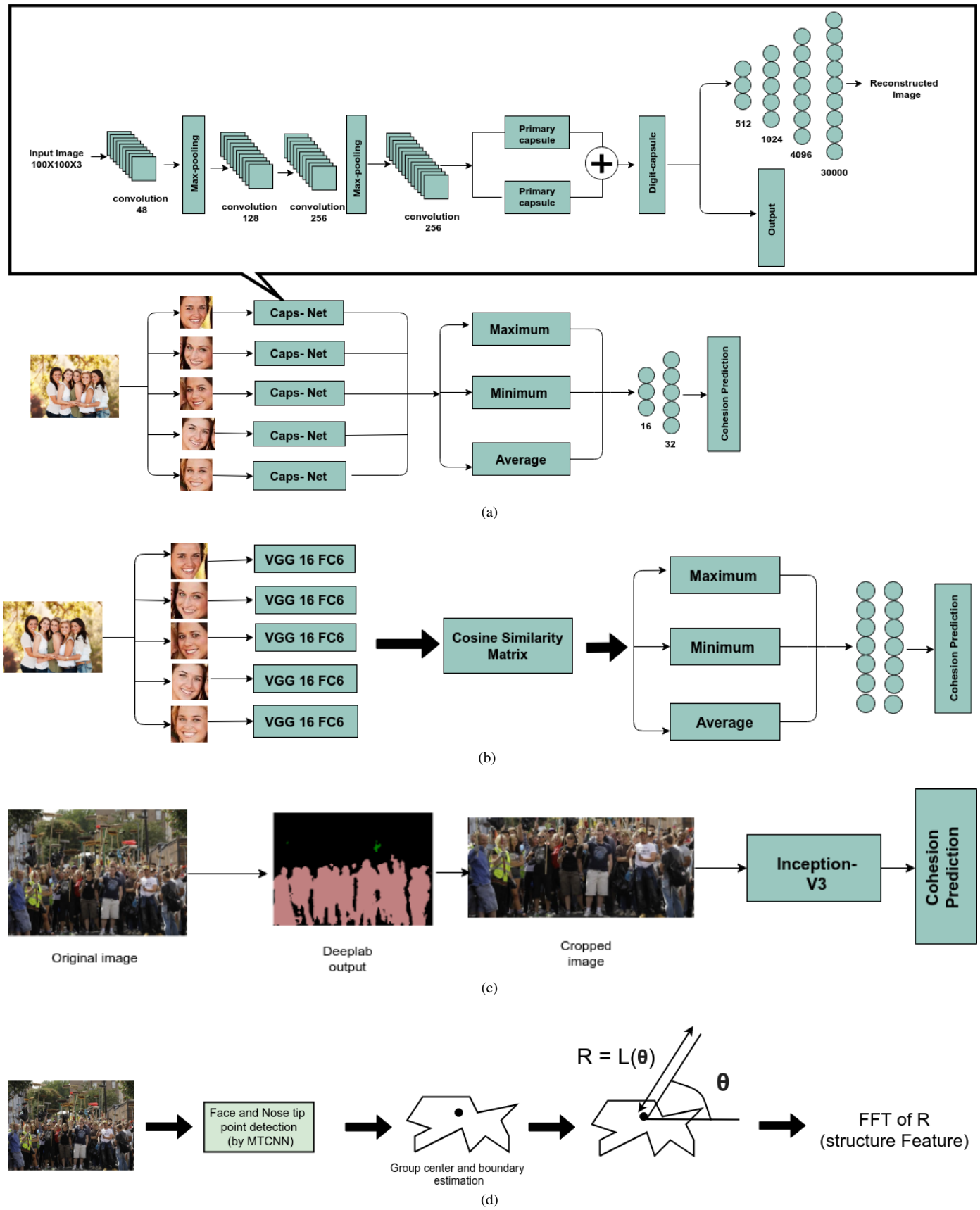

Fig. 8: (a) CapsNet structure for face-level expression analysis. The prediction from this network is further pooled to predict the GCS. The face-level part first predicts expression (as shown in this Figure) and then computes average, minimum and maximum. The details can be found in Section 6.2. (b) Pipeline for cohesion prediction via facial similarity matrix. (c) Pipeline for the analysis of the background level importance using group segmentation. We crop the group [12] before inputting to the network for GCS prediction. (d) Network to encode structure information.

Layers	Input	Output	Layer Details
Dense	b,3,7	b,3,16	16
BN and Activation	b,3,16	b,3,16	ReLU/Swish
Dense	b,3,16	b,3,32	32
BN and Activation	b,3,32	b,3,32	ReLU/Swish
Max Pooling	b,3,32	b,1,32	3(1-D)
Flatten	b,1,32	b,32	-
Cohesion (Sigmoid)	b,32	b,1	1

TABLE 4: Face-level emotion network architecture. Here, b and BN refer to the batch size and batch normalization, respectively.

and average) from this matrix and input to DNN with two dense layers (512, 64) for cohesion score prediction. The details are shown in Table 5 and Figure 8.

6.3 Effect of Background

We investigate how an image's background effects AGC? We used the Deeplab V3plus [12] segmentation network to crop people from a group in an image. We consider an area-wise threshold to chose an image for further analysis. If the segmented area is less than 50% of the total area of an image then this image is considered (Figure 8). We observed that when we use the segmented image for training, then there is a drop (around 0.103 MSE decreased) in performance. It indicates that the image background affects the perception of a group's cohesiveness. The background may reflect something about the social event in which the group is participating and is important for the prediction.

6.4 Effect of Group Structure

We also explore the relationship between group structure and group-level cohesion. The relative distance between people during interaction has been widely studied in social sciences [2] and anthropology [33]. The relative distance between group members is dependent on their relationship, social context and culture. Gallagher et al. [23] also included group structure related contextual information in a group setting. Wang et al. [69] used structure information to distinguish between family and non-family images. It is argued that sometimes group structure leads to perception of bonding. For e.g. group structure among friends and office colleagues differs. It is possible that if group members are friends, they may stand closer and the area inside the polygon structure could be less as compared to that for office colleagues. In order to encode a group's structure information, we follow the slightly modified version of [69], which is mentioned below:

- Detect faces using MTCNN [74] and use the tip of the nose as a face's location.
- Compute k-means clustering to find centroid of the group.

Layers	Input	Output	Layer Details
Dense	b,3,4096	b,3,512	512
BN and Activation	b,3,512	b,3,512	ReLU/Swish
Dense	b,3,512	b,3,64	64
BN and Activation	b,3,64	b,3,64	ReLU/Swish
Max Pooling	b,3,64	b,1,64	3(1-D)
Flatten	b,1,64	b,64	-
Cohesion (Sigmoid)	b,64	b,1	1

TABLE 5: Face-level similarity network architecture. Here, b and BN refer to the batch size and batch normalization, respectively.

Layer	Input	Output	Layer details
Input layer	b, 3, 4096	b, 3, 4096	
Dense	b, 3, 4096	b, 3, 1024	1024 nodes with ReLU activation
Dense	b, 3, 1024	b, 3, 512	512 nodes with ReLU activation
Dropout	b, 3, 512	b, 3, 512	0.5
Dense	b, 3, 512	b, 3, 128	128 nodes with ReLU activation
Dropout	b, 3, 128	b, 3, 128	0.3
Dense	b, 3, 128	b, 3, 3	3 nodes with softmax activation
Max-pooling	b, 3, 3	b, 1, 3	(3, 3)
Flatten	b, 1, 3	b, 3	
Dense	b, 3	b, 3	3 nodes with softmax activation

TABLE 6: This table describes the architecture of the network described in Section 6.6. Here, b refers to the batch size.

- Detect the boundary of the group structure as shown in Figure 8 (we use single region alpha shape [20] with 0.5 shrink factor for this purpose).
- Divide the angle around the centroid into 64 folds:

$$\theta_j = \frac{2\pi}{64} * j, where j \in \{1, 2..., 64\}.$$

- From the centroid to the boundary, take 64 radii (R) of this polygon $R = L(\theta_j)$. Here, R is a vector of length = 64.
- Compute Fast Fourier Transform on R and extract feature along the amplitude spectrum.
- Train DNN of FC layers (64, 32, 1) for prediction.

6.5 Effect of Body Pose

In the survey (Section 4), the participants mentioned keywords related to body pose (e.g. cheering, hugging etc.). Furthermore, prior works [47], [36], [1], [49] in the domain of group-level emotion and personality estimation also used body level features. To understand the effect of body pose based feature analysis on group-level cohesion, we conduct the following experiment:

- Detect human 2D pose and extract part based heat-maps to use as a feature [7].
- Add global average pooling layer followed by a DNN of FC layers (128, 64, 1) for prediction.

6.6 Contribution of Subset of Faces

We also analysed the impact of using a 'subset of faces' for group-level cohesion. This experiment is motivated by the concept of saliency in images. It is possible that few faces in a group are more dominant than others and may affect the annotators' perception of cohesion. Several prior works proposed predicting importance of objects in an image based on factors, which effect the human perception [60], [5], [39]. These factors include compositions (i.e. size and location of objects), semantics (i.e. object type, scene type along with its description strength) and context information. Few works [59], [43], [25] also proposed method for finding the important person in a group image. Although there are several

Network	Accuracy (%)	MSE
Inception V3 (emotion and cohesion prediction)	85.58	0.8181
Inception V3 (emotion prediction)	65.41	NA
Inception V3 (cohesion prediction)	NA	0.8537

TABLE 7: The results of image-level group emotion (classification accuracy) and cohesion (MSE) analysis.

Network Details	Image-Level	Face-level Emotion	Face-level Similarity	Group Structure	Body Pose	EmotiW Baseline
GCS (MSE on Val. set)	0.85	1.11	0.96	1.08	1.88	0.84
GCS (MSE on Test set)	0.53	0.91	0.82	0.98	1.90	0.50

TABLE 8: Comparisons of GCS prediction using the image-level and face-level networks.

Cross validation	MSE (lr=0.001)	MSE (lr=0.01)
1^{st}	0.63958	0.65662
2^{nd}	1.10628	1.06666
3^{rd}	0.70162	0.67964
4^{th}	0.60604	0.76320
5^{th}	0.93969	0.89159
Average	**0.79864**	**0.81155**

TABLE 9: 5-fold cross-validation results of the GAF cohesion database. Here lr = learning rate.

Cross validation	MSE (top-3 faces)	MSE (all faces)
1^{st}	1.20	1.08
2^{nd}	1.68	1.25
3^{rd}	1.21	1.15
4^{th}	1.19	1.12
5^{th}	1.09	1.01
Average	**1.274**	**1.122**

TABLE 11: 5-fold cross-validation on the GAF cohesion database with area wise top-3 and all faces.

factors which influence the perception of cohesion annotation, our study includes only one factor i.e. area of the face.

After face detection [74], we choose the three largest faces. We extract the VGG-16 FC6 layer feature (4096 dimensional). Further, we extract statistical features (maximum, minimum and average) from these features and train a DNN (network architecture - Table 6). For this study, we choose 500 images from the proposed dataset and perform the above-mentioned experiment.

7 Experimental Details and Results

In this section, we discuss the experimental settings and results. First, we treat cohesion prediction as a regression problem (as defined in [66]) and the group emotion prediction as a classification problem (as defined in [16]). We use the Keras [13] deep learning library for the code implementation.

7.1 Experimental Setup

Following experiments are conducted to explore different dimensions of group cohesion: 1) Image-Level analysis, 2) Face-Level emotion analysis, 3) Face-Level similarity analysis, 4) Effect of group structure, 5) Effect of body pose structure, and 6) Contribution of group cohesion on a subset of faces.

7.2 Evaluation Metrics and Implementation Details

Mean squared error (MSE) and overall accuracy are used as evaluation matrices for group-level cohesion and emotion prediction respectively. The implementation detail are discussed below:

Image-Level Analysis. We train Inception V3 network for image-level analysis. We initialize the network with ImageNet pre-trained weights and fine-tune the network with SGD optimizer with a learning rate of 0.001 and momentum 0.9 without learning rate decay. With similar hyperparameters, we jointly train Inception V3 network for both emotion and cohesion prediction too.

Face-Level Emotion Analysis. To predict GCS, we pre-trained a CapsNet on RAF-DB [42]. RAF-DB [42] is a facial expression database containing 30K images with each containing a single subject. The labels are the seven universal basic emotions and the twelve compound emotions. We use the basic emotions (i.e. *happy*, *neutral*, *sad*, *angry*, *surprise*, *disgust* and *fear*) to train our CapsNet. From group images, we extracted faces via MTCNN [74]. After training on RAF-DB, we extract the output probability vector for each face in the group image. Further, we compute statistics over these emotion probabilities and pass it through two more dense layers before the final cohesion score prediction. Our statistics include maximum, average and minimum of respective emotion probabilities. The motivation behind this is that we need to conclude over a group. Hence, these three intensity level analyses perform better for group-level tasks. We train a CapsNet with hyperparameters from the original paper (Adam optimizer with default settings in the Keras library and learning rate decay of 0.001 in every 10^{th} epoch to avoid local minima). We train the rest of the network via SGD optimizer with learning rate 0.01 and without any learning rate decay.

Face-Level Similarity Analysis. We extract the VGG-16 FC6 layer feature[2] for each face in a group. We compute the cosine similarity as described in Section 6.2.2. Further, we compute statistics over these features and pass it through two more dense layers before the final cohesion score prediction. Our statistics include maximum, average and minimum of respective VGG face features. The motivation behind this is that we need to conclude over a group. Hence, these three intensity level analyses perform better for group-level tasks. We train the rest of the network via SGD optimizer with learning rate 0.01 and without any learning rate decay.

Effect of Group Structure and Body Pose Results. Similar as above, we train the DNNs using SGD optimizer with learning rate 0.01 and without any learning rate decay.

Contribution of AGC on a Subset of Faces. We use SGD optimizer with its default parameters in Keras and 'categorical cross-entropy' as a loss function and evaluate with a 5-fold cross-validation protocol.

7.3 Results

The **image-level** based analysis results in TABLE 7 show an interesting pattern. When the inception V3 is individually used for group emotion and cohesion prediction, its performance is lower than the joint training. Thus, it suggests that the network learns more relevant representations of group emotion. We can

2. http://www.vlfeat.org/matconvnet/pretrained/

GAF 2.0	Ours	[64]	[30]	[70]	[16]
Accuracy(%)	85.67	83.90	80.05	77.92	52.97

TABLE 10: Group emotion performance comparison on GAF.

Network	Total Parameters	Time per Epoch
Inception V3	21,802,784	92 sec
CapsNet (encoder only)	7,290,080	10 sec
DNN	approx. 4,000,000	3-4 sec

TABLE 12: Number of parameters and training time comparison.

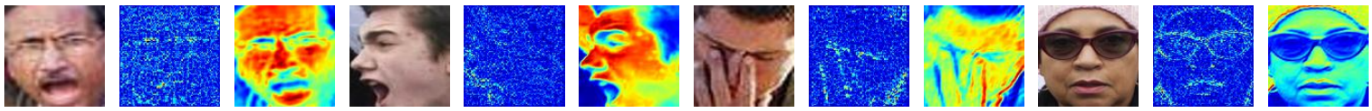

Fig. 9: Visualization of facial emotion based AGC network - Each set of three images shows the original image, saliency map and class activation map (CapsNet) respectively. Here, the red coloured region indicates activated regions. It is visible that the CapsNet can handle non-frontal, occluded, scaled and rotated images properly. [Best viewed in colour]

Fig. 10: Visualization of image-level cohesion. Each row consists of the original image, saliency map, and class activation map. The first row focuses on background features, the second row focuses on foreground features, the third row focuses on context level feature and last row on facial regions. [Best viewed in colour]

conclude that the emotion and cohesion at group-level are inter-related terms. The human perception behind group emotion and cohesion has some sort of similarity. To verify this hypothesis, we calculate Spearman's Rank-Order Correlation between cohesion and emotion. Without joint training, the Spearman's correlation coefficient ρ is 0.67. On the other hand, Spearman's ρ increases to 0.72 with joint training. This validates our hypothesis above that there exists a correlation between emotion and cohesion.

This is in accord with psychology studies [3]. It is also interesting to note that the effect of joint training is opposite to the GCS prediction as to the prediction error increases. One possible reason is that GCS and group emotion features contradict each other. Let us consider an example of a sobbing family, which has high GCS and negative group emotion and compare that with a celebrating sports team, which will also have high GCS. In the later, the group emotion will be positive. Scenarios like this may lead to ambiguity during the joint training from the GCS prediction perspective.

Our image-level, face-level, group structure and body pose experimental results on GAF Cohesion database are shown in TABLE 8. Overall, the image-level network performs better than face-level, group structure and body pose attributes on both validation and test set. The results indicate that context information is richer than others. Thus, the validation and test set MSE (0.85 and 0.53 respectively) are lower as compared to the other attributes. Further, the contribution of face level information is more relevant as compared to the group structure and body pose. We also conduct experiments corresponding to the facial emotion and face-level similarity of the group members. From the MSE score, we can conclude that face-level similarity is more relevant on group cohesion context. As there is a high probability that group members with similar age, race, ethnicity, occupation etc. will be same in a group [45]. On the other hand, the group structure and body pose vary a lot in a different context. Due to this reason, the group structure and group members' body pose does not contribute much towards the inference of group cohesion.

TABLE 9 describes 5-fold cross-validation results on the GAF cohesion database. In TABLE 10, group emotion is predicted, when AGC information is used for joint training. The results on the GAF 2.0 show better performance than the other state-of-the-art methods. This shows that cohesiveness information is useful for group emotion prediction.

The results regarding the contribution of group cohesion on a subset of faces are shown in Table 11. The results indicate that faces with more area contribute more to the group-level emotion prediction as compared to the small faces. The salient regions (here, area-wise large faces) of the images could influence annotator's perception. Generally, large objects catch the viewer's attention as compared to the smaller objects. From Table 11, we can infer that the main expression cue lies in area-wise top few faces and other faces contribute less towards group-level cohesion.

From Table 12, it is observed that Inception V3 requires more time and space. However, CapsNet and DNN are relatively lighter networks and require comparatively less time for training.

7.4 Visualization (Saliency vs Class Activation)

In this section, we discuss visual attributes which our network learnt. We visualize the class activation map and discuss its comparison with the saliency. From Figure 9, we can observe that in spite of non-frontal, rotated, occlusion, blurred faces, CapsNet can handle each case efficiently. Especially, it deals with the rotation and scaling of different objects in an image individually and shows better performance over both occluded and partially occluded images which is beneficial for our problem. Moreover, it did not require data augmentation while training and thus it is efficient regarding time complexity. Similarly, for image-level analysis, (as shown in Figure 10) the top row activates the background, the second row activates the foreground, the third row activates the subject and the last row activates both the front person and background. In the case of the top row, it activates the background,

Fig. 11: This figure shows frame-wise cohesion score in 'team sports' context. From left to right, predicted cohesion scores indicate a smooth transition from strong to very strong cohesion.

as the group takes up a small space as compared to the visible background. Similarly, in the second case, the foreground is more dominant as compared to the background. In the third row, the main features of the protests that are activated are the banners. In the last picture, it activates both foreground and background, especially the facial region.

From Figures 9 and 10, it is observed that salient image regions and class activation regions are not highly correlated. Although, there are some regions in images, which are common in both cases as sometimes the human visual system is influenced by salient features of the images.

8 Applications

According to Hung et al. [38], group-level cohesion is applicable to various contexts such as team sports [8], group psychotherapy [6] and military training [28]. As group-level cohesion influences a friendly and cooperative working environment, it is an important factor in group-level performance. The main consequence of high group-level cohesion is members' collaborative approach for group tasks. The performance-enhancing effects of cohesion motivate each individual to perform better in the group that is committed towards the group's outcome [10]. Additionally, cohesion influences the four stages of group dynamics (Forming, Storming, Norming and Performing) [6]. In the forming stage, the group cohesion is important as it influences similar minded people to join in a task specific group. It is the beginning of the group's formation. In the recruitment process of an important group, the cohesion based judgement can be made as it is important for collaborative performance. In the storming process, group cohesion prediction is also relevant as one can observe the pattern of how group cohesion leads to addition and deletion of group members. Norms are the informal rules that group members are adopted to regulate members' behaviour. In the norming process too, cohesion comes into play. As these norms are formed by people, cohesion influences these studies. Finally, the performing stage is the result of the previous stages where the group performed task. The group-level cohesion can be deployed in this context as the groups performance ideally depends on group members contributions. Thus by analysing these stages, certain group-level behaviours (aggressive, passive etc.) can be automatically predicted. In this context, we compute frame-wise cohesion score of a group in YouTube video (Fig. 11). The smooth change in cohesion score from strong to very strong indicates the possibility of video level cohesion prediction in real-world settings such as for studying group dynamics (mentioned above).

Further in the context of group performance as well as leader selection, one can study the verbal as well as non-verbal communication patterns of the team members to get more clues regarding group dynamics and decide who will be the ideal candidate to become a group leader [44]. For improving team performance, a leader is required who can handle internal team conflict and encourage greater team cohesiveness. By analyzing a long video, one can automatically predict the probable leader of the team by judging the cohesiveness of a group.

9 Conclusion, Limitation and Future Work

The main motivation behind our work is understanding factors, which affect cohesion perception and mapping them to a prediction network. Our model performs comparatively well and achieves near the human-level performance. From our experimental results, it can be deduced that AGC and emotion are interrelated. In this work, we studied different dimensions of the group's cohesiveness from computer vision perspective. From our experimental results, we observed that the usefulness of body pose and group structure is relatively less as compared to the face and the holistic features. We also observed that the CapsNet [53] also performs well on facial expression recognition without data augmentation. Although the faces in a group image vary largely, i.e. the face can be occluded, blurred or non-frontal and others. Via visualization, we learnt that the scene information encodes the background, clothes and various objects in an image. This information is also known as the top-down contextual information.

The main limitation of our work is the cultural influence in data annotation as it is related to the perception of cohesion. A potential future direction for our work is to investigate how facial attributes affect AGC prediction. It will be interesting to analyze the role of the group members' behavioural signals along with the face. Although the image-level network does encode a few of them, however, its complete contribution requires further investigation. It will be of interest to analyze the fashion quotient of the group by parsing the clothes for patterns and themes, which correspond to specific social events, although, some patterns are already encoded in our scene level analysis. Furthermore, another possible direction is to include kinship related information in the network because irrespective of visual expression, sometimes kinship indicates strong cohesion [71]. Additionally, it would be interesting to study different forms of group cohesion i.e. task and social cohesion. Our study is limited to the spatial analysis of the image. Temporal data can provide more information regarding group dynamics which can be useful in a few task relevant scenarios.

Acknowledgment

We are grateful to all the brave frontline workers who are working hard during this difficult COVID19 situation. Tom Gedeon and

Abhinav Dhall's research is partially supported by the Australian Research Council's grant DP190102919.

REFERENCES

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2015.

[2] I. Altman. The environment and social behavior: Privacy, personal space, territory, and crowding. *ERIC*, 1975.

[3] S. G. Barsade and D. E. Gibson. Group emotion: A view from top and bottom. *Composition.*, 1998.

[4] D. J. Beal, R. R. Cohen, M. J. Burke, and C. L. McLendon. Cohesion and performance in groups: a meta-analytic clarification of construct relations. *Journal of Applied Psychology*, 88(6):989, 2003.

[5] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3562–3569. IEEE, 2012.

[6] L. J. Braaten. Group cohesion: A new multidimensional model. *Group*, 15(1):39–55, 1991.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[8] A. V. Carron, L. R. Brawley, and W. N. Widmeyer. The measurement of cohesiveness in sport groups. *Advances in sport and exercise psychology measurement*, 23(7):213–226, 1998.

[9] A. V. Carron and K. S. Spink. The group size-cohesion relationship in minimal groups. *Small group research*, 26:86–105, 1995.

[10] A. Chang and P. Bordia. A multidimensional approach to the group cohesion-group performance relationship. *Small Group Research*, 32(4):379–405, 2001.

[11] M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu. Group level activity recognition in crowded environments across multiple cameras. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 56–63. IEEE, 2010.

[12] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018.

[13] F. Chollet. keras. *GitHub Repository,2015*, 2015.

[14] A. Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pages 546–550, 2019.

[15] A. Dhall, R. Goecke, and T. Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, 2015.

[16] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *ACM International Conference on Multimodal Interaction*, pages 524–528, 2017.

[17] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 427–432, 2016.

[18] 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 *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.

[19] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 653–656. ACM, 2018.

[20] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983.

[21] C. R. Evans and K. L. Dion. Group cohesion and performance: A meta-analysis. *Small group research*, 22(2):175–186, 1991.

[22] D. Feltz. Understanding motivation in sport: A self-efficacy perspective. *Motivation in sport and exercise*, pages 93–105, 1992.

[23] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *IEEE Computer Vision and Pattern Recognition*, 2009.

[24] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1003–1016, 2012.

[25] S. Ghosh and A. Dhall. Role of group level affect to find the most influential person in images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[26] S. Ghosh, A. Dhall, and N. Sebe. Automatic group affect analysis in images via visual attribute and feature networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1967–1971. IEEE, 2018.

[27] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon. Predicting group cohesiveness in images. In *IEEE Joint Conference on Neural Networks*, 2018.

[28] J. Griffith. Further considerations concerning the cohesion-performance relation in military settings. *Armed Forces & Society*, 34(1):138–147, 2007.

[29] D. Guo, K. Wang, J. Yang, K. Zhang, X. Peng, and Y. Qiao. Exploring regularizations with face, body and image cues for group cohesion prediction. In *2019 International Conference on Multimodal Interaction*, pages 557–561, 2019.

[30] X. Guo, L. F. Polanía, and K. E. Barner. Group-level emotion recognition using deep models on image scene, faces, and skeletons. In *ACM International Conference on Multimodal Interaction*, pages 603–608, 2017.

[31] K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 2008.

[32] J. R. Hackman. *Group influences on individuals in organizations.* Consulting Psychologists Press, 1992.

[33] E. T. Hall. A system for the notation of proxemic behavior. *American anthropologist*, 65(5):1003–1026, 1963.

[34] J. Hernandez, M. Hoque, W. Drevo, and R. W. Picard. Mood meter: counting smiles in the wild. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 301–310, 2012.

[35] M. A. Hogg. Group cohesiveness: A critical review and some new directions. *European review of social psychology*, 4(1):85–111, 1993.

[36] X. Huang, A. Dhall, R. Goecke, M. Pietikäinen, and G. Zhao. Multimodal framework for analyzing the affect of a group of people. *IEEE Transactions on Multimedia*, 20(10):2706–2721, 2018.

[37] 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 *British Machine Vision Conference*, pages 34–1, 2015.

[38] H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia,2015*, 1(6):563–575, 2015.

[39] S. J. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International journal of computer vision*, 100(2):134–153, 2012.

[40] J. R. Kelly and S. G. Barsade. Mood and emotions in small groups and work teams. *Organizational behavior and human decision processes*, 86(1):99–130, 2001.

[41] J. Li, S. Roy, J. Feng, and T. Sim. Happiness level prediction with sequential inputs via multiple regressions. In *ACM International Conference on Multimodal Interaction*, pages 487–493, 2016.

[42] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Computer Vision and Pattern Recognition*, 2017.

[43] W.-H. Li, B. Li, and W.-S. Zheng. Personrank: Detecting important people in images. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 234–241. IEEE, 2018.

[44] C. G.-G. López, F. M. Alonso, M. M. Morales, and J. A. M. León. Authentic leadership, group cohesion and group identification in security and emergency teams. *Psicothema*, 27(1):59–64, 2015.

[45] A. J. Lott and B. E. Lott. Group cohesiveness as interpersonal attraction: A review of relationships with antecedent and consequent variables. *Psychological Bulletin*, 1965.

[46] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras. Amigos: A dataset for mood, personality and affect research on individuals and groups. *IEEE Transactions on Affective Computing*, 2019.

[47] W. Mou, O. Celiktutan, and H. Gunes. Group-level arousal and valence recognition in static images: Face, body and context. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 5, pages 1–6. IEEE, 2015.

[48] W. Mou, H. Gunes, and I. Patras. Alone versus in-a-group: A comparative analysis of facial affect recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 521–525, 2016.

[49] W. Mou, C. Tzelepis, V. Mezaris, H. Gunes, and I. Patras. A deep generic to specific recognition model for group membership analysis using non-verbal cues. *Image and Vision Computing*, 81:42–50, 2019.

[50] A. E. Myers. Team competition, success, and the adjustment of group members. Technical report, Illinois University Urbana Group Effectiveness Research Lab, 1961.

[51] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European*

Conference on Computer Vision, 2016.
[52] M. W. Rempel and R. J. Fisher. Perceived threat, cohesion, and group problem solving in intergroup conflict. *International Journal of Conflict Management*, 8(3):216–234, 1997.
[53] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.
[54] E. Salas, N. J. Cooke, and M. A. Rosen. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*, 50(3):540–547, 2008.
[55] G. Sharma, S. Ghosh, and A. Dhall. Automatic group level affect and cohesion prediction in videos. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 161–167. IEEE, 2019.
[56] M. E. Shaw. *Group dynamics: The psychology of small group behavior*. McGraw Hill, 1971.
[57] V. K. Singh, A. Atrey, and S. Hegde. Do individuals smile more in diverse social company?: Studying smiles and diversity via social media photos. In *Proceedings of the ACM on Multimedia Conference*, 2017.
[58] 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.
[59] C. Solomon Mathialagan, A. C. Gallagher, and D. Batra. Vip: Finding important people in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4858–4866, 2015.
[60] M. Spain and P. Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 91(1):59–76, 2011.
[61] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu. Lstm for dynamic emotion and group emotion recognition in the wild. In *ACM International Conference on Multimodal Interaction*, pages 451–457, 2016.
[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
[63] H. Tajfel. *Social identity and intergroup relations*. Cambridge University Press, 2010.
[64] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao. Group emotion recognition with individual facial emotion cnns and global image based cnns. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 549–552, 2017.
[65] W. R. Thompson and D. P. Rapkin. Collaboration, consensus, and detente: The external threat-bloc cohesion hypothesis. *Journal of Conflict Resolution*, pages 615–637, 1981.
[66] T. Treadwell, N. Lavertue, V. Kumar, and V. Veeraraghavan. The group cohesion scale-revised: reliability and validity. *Journal of Group Psychotherapy, Psychodrama and Sociometry*, 54(1):3, 2001.
[67] J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell. *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, 1987.
[68] V. Vonikakis, Y. Yazici, V. D. Nguyen, and S. Winkler. Group happiness assessment using geometric features and dataset balancing. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 479–486, 2016.
[69] X. Wang, G. Guo, M. Rohith, and C. Kambhamettu. Leveraging geometry and appearance cues for recognizing family photos. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.
[70] Q. Wei, Y. Zhao, Q. Xu, L. Li, J. He, L. Yu, and B. Sun. A new deep-learning framework for group emotion recognition. In *ACM International Conference on Multimodal Interaction*, pages 587–592, 2017.
[71] D. R. White. Sage handbook of social network analysis, 2011 edited by john scott and peter carrington kinship, class, and community. *SAGE publications*, 2011.
[72] T. Xuan Dang, S.-H. Kim, H.-J. Yang, G.-S. Lee, and T.-H. Vo. Group-level cohesion prediction using deep learning models with a multi-stream hybrid network. In *2019 International Conference on Multimodal Interaction*, pages 572–576, 2019.
[73] S. J. Zaccaro and M. C. McCoy. The effects of task and interpersonal cohesiveness on performance of a disjunctive group task. *Journal of Applied Social Psychology*, 18(10):837–851, 1988.
[74] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
[75] B. Zhu, X. Guo, K. Barner, and C. Boncelet. Automatic group cohesiveness detection with multi-modal features. In *2019 International Conference on Multimodal Interaction*, pages 577–581, 2019.

Shreya Ghosh is currently pursuing her PhD at Monash University, Australia. She received MS(R) degree in the Computer Science and Engineering from the Indian Institute of Technology Ropar, India. She received the B.Tech. in CSE from the Govt. College of Engineering and Textile Technology (Serampore, India). Her research interests include affective computing, computer vision, deep learning. She is a student member of the IEEE.

Abhinav Dhall received the PhD degree from the Australian National University in 2014. He is currently a Lecturer with Monash University, Australia and an Assistant Professor (on leave) with Indian Institute of Technology Ropar, India. He was awarded the Best Doctoral Paper Award at ACM ICMR 2013, Best Student Paper Honourable mention at IEEE AFGR 2013 and Best Paper Nomination at IEEE ICME 2012. His research interests are in computer vision and affective computing. He is a member of the IEEE.

Nicu Sebe received the PhD degree from Leiden University, The Netherlands, in 2001. He is currently a Professor with the University of Trento, Italy, leading the research in the areas of multimedia information retrieval and human behaviour understanding. He was the General Co-Chair of the IEEE FG Conference 2008 and the ACM Multimedia 2013, and the Program Chair of the International Conference on Image and Video Retrieval in 2007 and 2010 and the ACM Multimedia 2007 and 2011. He was the Program Chair of ICCV 2017 and ECCV 2016 and the General Chair of ACM ICMR 2017. He is a fellow of the International Association for Pattern Recognition.

Tom Gedeon received the BSc and PhD degrees from the University of Western Australia, Crawley, Australia. He is currently Professor of Computer Science and Head of the Human-Centred Computing (Hcc) Research Area of the Research School of Computer Science in the College of Engineering and Computer Science at The Australian National University in Canberra. He is a former president of the Asia-Pacific Neural Network Assembly and a former President of the Computing Research and Education Association of Australasia. He serves on journal advisory boards as member or editor. He is a senior member of the IEEE.