Automatic Group Level Affect and Cohesion Prediction in Videos

Conference Paper · September 2019

DOI: 10.1109/ACIIW.2019.8925231

3 authors:

Garima Sharma
Monash University (Australia)
18 PUBLICATIONS **338** CITATIONS
SEE PROFILE

Shreya Ghosh
Curtin University
46 PUBLICATIONS **930** CITATIONS
SEE PROFILE

Abhinav Dhall
Indian Institute of Technology Ropar
179 PUBLICATIONS **6,680** CITATIONS
SEE PROFILE

2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

Automatic Group Level Affect and Cohesion Prediction in Videos

Garima Sharma[1] Shreya Ghosh[1] Abhinav Dhall[1,2]
[1]*Indian Institute of Technology Ropar* [2]*Monash University*
{2017csz0004, shreya.ghosh}@iitrpr.ac.in abhinav.dhall@monash.edu

***Abstract*—This paper proposes a database for group level emotion recognition in videos. The motivation is coming from the large number of information which the users are sharing online. This gives us the opportunity to use this perceived affect for various tasks. Most of the work in this area has been restricted to controlled environments. In this paper, we explore the group level emotion and cohesion in a real-world environment. There are several challenges involved in moving from a controlled environment to real-world scenarios such as face tracking limitations, illumination variations, occlusion and type of gatherings. As an attempt to address these challenges, we propose a 'Video level Group AFfect (VGAF)' database containing 1,004 videos downloaded from the web. The collected videos have a large variations in terms of gender, ethnicity, the type of social event, number of people, pose, etc. We have labelled our database for group level emotion and cohesion tasks and proposed a baseline based on the Inception V3 network on the database.**

***Index Terms*—Group Level Emotion, Group Cohesion, Multimodal affect, Context analysis.**

I. Introduction

Emotion is an essential part of human experience. It impacts a lot on human cognition, perception, learning, communication, decision-making and many other daily life scenarios. Being a social-being, human-mind inclines towards any group environment which further develops several consequences such as inter-personal emotion, bonding, interaction, ethics etc. Over the past few years in affective computing, automatic group affect analysis has been of interest due to its applications in business and organization domains [1]. The main objective of this paper is to explore the techniques for automatic group affect prediction which mainly comprises of *group level emotion and cohesion*. With respect to group of people, emotion can be analyzed at both intra-personal and inter-personal levels. Intra-personal emotion is the awareness of one's own emotional states. When people in a group spend time together due to any reason, gradually each and every person in the group is effected by other's emotional state. In psychological terms, synchronization of group members' mentality is the stepping stone of group formation [2]. Further, 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 [3]. In the existing cognitive science literature [2], 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.

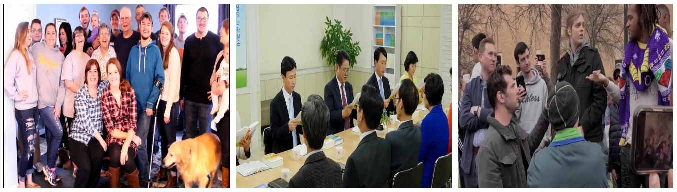


Fig. 1. Frames from VGAF across three emotion categories. The frames have positive, neutral and negative emotion, respectively (left to right).

A group's perceived affect may change over the group's lifetime [2]. The main driving forces behind this process are group members' affective diversity, cohesiveness and emotional culture. For simplicity, the perception of group affect in videos is studied in this paper. In this study, group level emotion is divided into three categories: positive, negative and neutral. Cohesion is measured following the Group Cohesion Scale (GCS) [4]. Group's cohesiveness and group emotion are interrelated terms. Group cohesiveness can be defined as a bonding which effects the membership of an individual in a group. The main motivation behind group's cohesiveness is inter-personal 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 [5]. The positive consequences of group cohesiveness are more participation, more conformity, high productivity, more success and more personal level satisfaction [6].

Our main challenge is to deal with the above-mentioned points from affective computing perspective. For experiments, we curate group videos from the YouTube (having creative common license). All of the videos contain multiple subjects having variations in contexts, lighting conditions and camera quality, etc. Thus, the proposed baseline model is validated in real-world scenarios.

The main contributions of this paper are as follows:

1) We collected an 'in the wild' database containing 1,004 videos of group of people. To the best of our knowledge, this is the largest publicly available video based group affect database.
2) We labelled our data for three group level emotion (positive, negative and neutral) categories and group cohesion scale having [0-3] range (Fig. 1 and 2 shows the sample frames).

978-1-7281-3891-6/19/$31.00 ©2019 IEEE

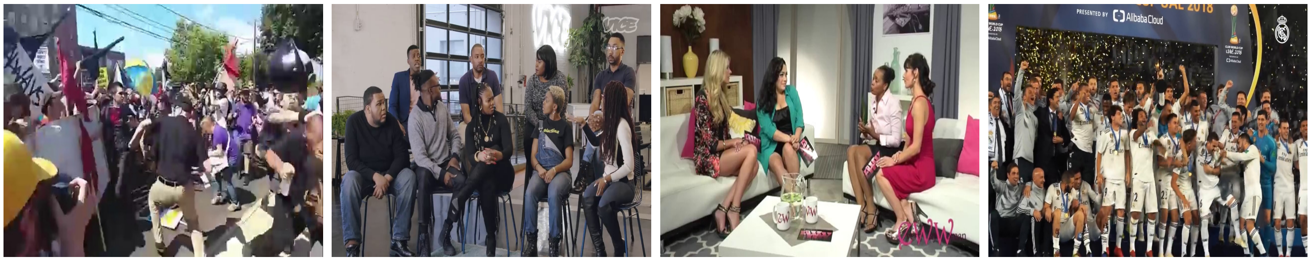


Fig. 2. Frames from VGAF in which cohesion is defined on a scale [0-3], where 0 and 3 represents very weak and very strong cohesion, respectively. The cohesion level in frames varies from 0 to 3 from left to right.

3) Our proposed pipeline predicts group level emotion and cohesion on the basis of visual and audio features.

The rest of the paper is structured as follows: Section II describes the prior work in this area. Section III contains challenges and survey details. Section IV provides the details of the collected database. The architecture used to provide the baseline is discussed in section V. Section VI discusses the experimental results.

II. Related Work

A. Group Level Emotion

The group level emotion analysis started to gain attention from 2012 [7], [8]. One of the earlier work on group level emotion was proposed by Dhall et al. [1], [3]. They proposed a Group Expression Model (GEM) which predict the happiness intensity of a group of people. Generally, two kind of approaches (i.e. top down and bottom up) are used in literature to analyze a group of people. In the case of bottom up approaches, each person's facial expressions are analyzed and combined to predict the overall group level emotion. Vonikakis et al. [9] followed bottom up approach by extracting the geometric features from each person's face and predicted the emotional intensity of the group. On the other hand, top down approach leveraged global features to analyze a group. This approach is mainly based on the theory that the inter-personal relation, the social norm and context are more important factors to predict the emotion of the overall group [10], [11]. Smith et al. [12] also focused on the relevance of inter-personal relationships by stressing on 'how much a person is involved in the group?' as a driving factor for group level emotion. Due to these valid assumptions, the combination of holistic and face level features are widely used for deciding the emotion of a group of people.

The deep neural networks are widely used to extract the features for the prediction of group level emotion. Li et al. [13] used both face and scene level features and used LSTM to learn the group level score through regression. The evaluation was done on HAPPEI database [1]. Guo et al. [14] combined the holistic, face level, pose and hand movement information and performed a decision level fusion for group level information. Similarly, global and local features are fused together by Tan et al. [15] and Mou et al. [16]. These studies evaluate the effect of each attribute independently for group level emotion perspective. Quach et al. [17] performed a group level expression recognition in crowd videos. Their study estimated the expression of large size of group by training an end to end temporal non-volume preserving fusion.

The presence of other persons also effects the emotion of a person. Singh et al. [18] analyzed a group of people over some social interactions. The study found that a person may smiles more in a diverse company of people (like success, family functions etc.). In another interesting study, Mou et al. [19] performed an experiment to study the cross-subject affect analysis. The experiments show that the persons share the emotional and behaviour information in a group. Hence, the affect of one person can be used to predict the affect of others in a group. Apart from audio-visual cues, several physiological based information is also proved to be useful for finding the group related attributes. Correa et al. [20] recently proposed AMIGOS database which consists of EEG, ECG and GSR signals for the prediction of mood, personality and affects in a group setting.

B. Group Level Cohesion

There are several psychological based studies which show that several factors like inter-personal agreement, social norms, emotion of each individual affects the group level affect [10]. The group cohesion defines the inter-personal bonding between group members which results from the persons having the same opinion [21]. The factors affecting the group cohesiveness can be positive (like success, family functions) or negative (like riots or events raising the competition between people) [22], [23].

Zhang et al. [26] studied the behaviour of people in a group by focusing on their interactions and other activities. The study uses the wearable social sensors which record

TABLE I
Comparison of the prior databases in group level analysis domain.

Database	Modality	No. of samples	Recording	Labels
AMIGOS [20]	Video	40	Lab	Continuous emotions
GAF [24]	Image	15K	Web	3 Group emotions
HAPPEI [1]	Image	3K	Web	Happiness intensity
MultiEmoVA [16]	Image	400	Web	Arousal-valence
SALSA [25]	Audio-video	-	Lab	Personality score
VGAF (Ours)	Audio-video	1,004	Web	3 Group emotions

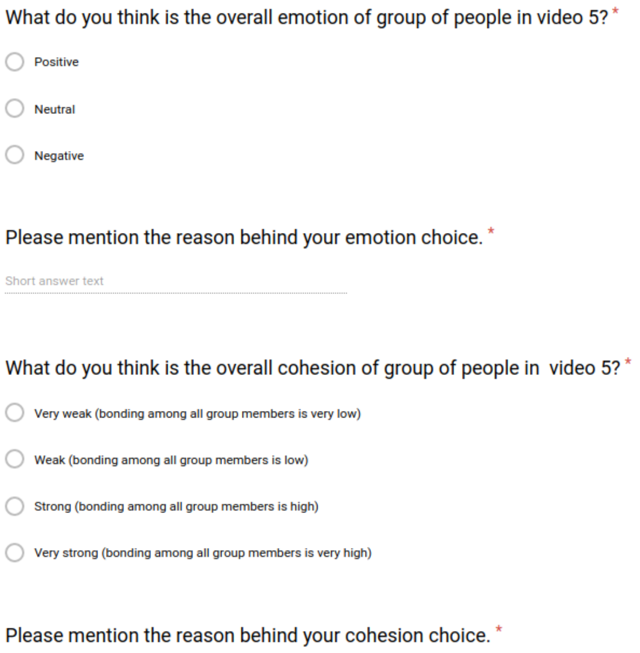
the movements, audio, face to face interactions, and their location. Hung et al. [27] analyzed the behaviour of a group by focusing on the cohesion. The study performed an experiment to analyze the cohesion on audio-visual data collected from a group meeting of four people. Features like pauses between individual turns, floor exchanges, turn length of each individual, overlapping, prosodic cues, etc are used to extract the useful information from the audio signals. Similarly, from visual information, features like motion of a person during overlapping speech, motion when not speaking and inter-personal synchrony, etc are extracted. Naive Bayes classifier and SVM was used in the study to predict cohesion. To the best of our knowledge, this was the first work to investigate the automatic cohesion of a group of people in videos. Ghosh et al. [24] also proposed an image based emotion and cohesion based database. The authors used holistic and face level features for the prediction of emotion and cohesion. Their experiments also showed that the holistic features are more important as compared to the face level to estimate the cohesion. The study proved that the emotion and cohesion are inter-related and can be estimated together from an image.

Table I provides the details of some group level emotion based databases. The MultiEmoVA [16] is an image based database collected from Flicker and Google images. The images are annotated for three categories in arousal and valence space i.e. negative, neutral, positive and low, medium, high, respectively. The database has a large variation in the resolution and number of people present in the group images. The SALSA database [25] is a multimodal database recorded in natural environment. Data is recorded in poster presentation and a cocktail party where some sensors were worn by the participants. The collected data was labelled for five personality trait scores and the position, head and body orientation for every 45 frames.

III. Challenges and Survey

This section describes the challenges involved in designing an automatic group level emotion and cohesion framework. First of all, we wish to know the attributes which greatly affect the perception of group level emotion and cohesion in a group video. In the existing literature [28] various visual cues are mentioned. However, the first perception after watching a video can differ considerably from person to person. Additionally, various challenges are also involved during real-world video analysis.

We conduct a survey to understand the perception of important visual cues for group level analysis. The survey is conducted online over 20 participants via a Google form. The form consists of 10 videos (as shown in Fig. 3) of groups of people in different contexts and have different emotion and cohesion labels. The participants are from various occupational domain like student, businessman, corporate employee, etc. The participants are familiarized with the emotion and cohesion concepts corresponding to few videos. The participants have to select one of the three emotion and four cohesion levels for each video and reasons behind their choice. Thus, we get few keywords corresponding to emotion and cohesion which further leads us to design our network. Fig. 4 refers to the emotion and cohesion related word clusters, corresponding to the keywords obtained. It can be observed from the word clusters that perceived emotion is estimated by *face, expressions, voice, argument,* etc. Whereas, keywords like *people, fighting, tone, listening, discussion,* etc. are important to decide the cohesion of a group of people.



What do you think is the overall emotion of group of people in video 5? *

○ Positive

○ Neutral

○ Negative

Please mention the reason behind your emotion choice. *

Short answer text

What do you think is the overall cohesion of group of people in video 5? *

○ Very weak (bonding among all group members is very low)

○ Weak (bonding among all group members is low)

○ Strong (bonding among all group members is high)

○ Very strong (bonding among all group members is very high)

Please mention the reason behind your cohesion choice. *

Fig. 3. Screenshot of the survey form. The survey is conducted to obtain the viewer's perception on group level emotion and cohesion.

IV. Database Description

We curate Video Group AFect (VGAF) database as databases in this domain are mostly constrained to lab environment (Table I). It can be observed from Table I that most of the databases are either recorded in a constrained environment or are image based only. Also, all of the databases focuses only on group level emotions. Only [24] has cohesion labels along with the emotion labels. Whereas, proposed VGAF is an audio-video based database containing group level emotion as well as cohesion labels. The database contains *1,004* videos having duration of 8-25 sec. To the best of our knowledge, VGAF is first video based database containing labels for emotion and cohesion. The database, frames and related baseline codes will be made publicly available https://sites.google.com/view/grouplevelaffect/home[Link].

A. Data Collection

We collected the videos form the YouTube having creative commons licence. For relevant video collection, we have searched on the basis of group related keywords. The keywords

Fig. 4. The word cluster of survey participant's keyword responses against the reason field as shown in survey form Fig. 3. The left and right one is corresponding to the emotion and cohesion, respectively.

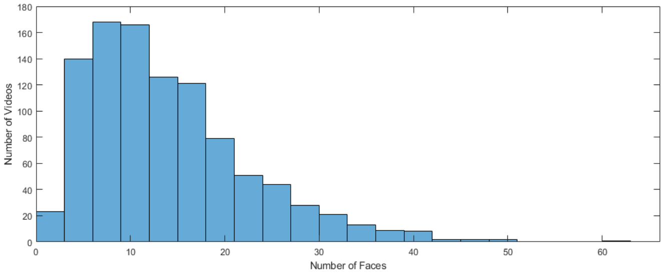


Fig. 5. The distribution of number of people vs. number of videos in VGAF.

are selected corresponding to the range of emotions and cohesion (e.g. *interview, festival, party, silent protest, violence*, etc.). We consider two and more than two people as a group. The collected videos have different resolution to keep the database unconstrained. The videos are then divided into short clips of duration 8 to 25 seconds. Generally, in a group video, the focus of camera varies to different part of a group while recording. Based on the structure of the group, either all the group members are present in a single frame or a frame may have only the part of the group. Hence, the videos are divided such that it includes all the group members over time. The resulting short clips are carefully analyzed and the videos which do not fulfil the given task were discarded. The resulting database consists of total 1,004 videos, where each video has a non-uniform number of clips. The frame length of the clips lies between 137 to 1,205.

To obtain the statistics regarding the number of people in the collected videos, we plot a graph with number of videos along y-axis and number of faces along the x-axis (Fig. 5). For this purpose, we use OpenCV People-Counter[1]. From Fig 5, it is observed that most of the videos contain an approximate of 4-18 number of people. This shows the diversity of the collected database which includes different types of possible groups in real-world scenarios.

The video clips are divided into three parts i.e. *Train* (587 samples), *Validation* (215 samples) and *Test* (202 samples). The division is performed such that all the clips of a video are contained in one split only. The splits also have almost equal distribution of samples across each class of emotion and cohesion.

B. Data Annotation

Each resulting clipping is manually annotated by 3 persons for emotion and cohesion. Each annotator is informed with the basic concepts of emotion and cohesion. Additionally, the observations derived from the conducted user survey is also considered for the generalized human perception. Videos which are not having the mutual consensus are discarded from the database. The assigned labels for group level emotion are along the valance axis of emotion (i.e. positive, neutral and negative) [29]. For group cohesion, annotation is done in the range [0-3], where 0 represents very low cohesion and 3 represents very high cohesion [4]. According to Treadwell et al. [4], the soft scaled anchor points (i.e. strongly agree, agree, disagree and strongly disagree) are reliable as it does not vary from annotator's perception-wise. It is to be noted that the emotion and cohesion is assigned for the whole video sequence for irrespective of the facial expression of one person. Hence, the label signifies the overall emotion and cohesion of the group and not for any individual. Fig. 1 and 2 show few frames from the collected videos with emotion and cohesion labels. It is to be noted that each video is labelled for emotion as well as cohesion. The separate frames are shown in the figure to provide a clear distinction between the two labels.

V. Network Architectures

In this section, we describe our proposed network details. We calculate our baseline by using inception V3 architecture as mentioned in [24]. An image level network is trained on GAF cohesion database [24] which consist of 15K group images with their emotion and cohesion label. The rationale behind choosing this architecture is that it can extract scale invariant contextual information. Additionally, it also provides a good trade-off between the number of hyper-parameters and accuracy (on ImageNet challenge) as compared to other CNN architectures. Thus, we can collectively analyze group and its surroundings. The given network has already obtained decent results for both group level emotion and cohesion task [24].

In this context, we extract the frame level predictions of emotion and cohesion via Inception V3 model which is pre-

TABLE II
The details of the VGAF database.

Total no. of original videos (downloaded from YouTube)	234
Total no. of clips in database	1,004
Duration	8-25 sec
Database split	Train - 587, Val - 215 Test - 202
Emotion labels (with classwise distribution)	Neutral (291), Positive (296) Negative (417)
Cohesion label	[0-3] range 0 - very weak, 3 - very strong

[1] https://www.pyimagesearch.com/2018/08/13/opencv-people-counter/

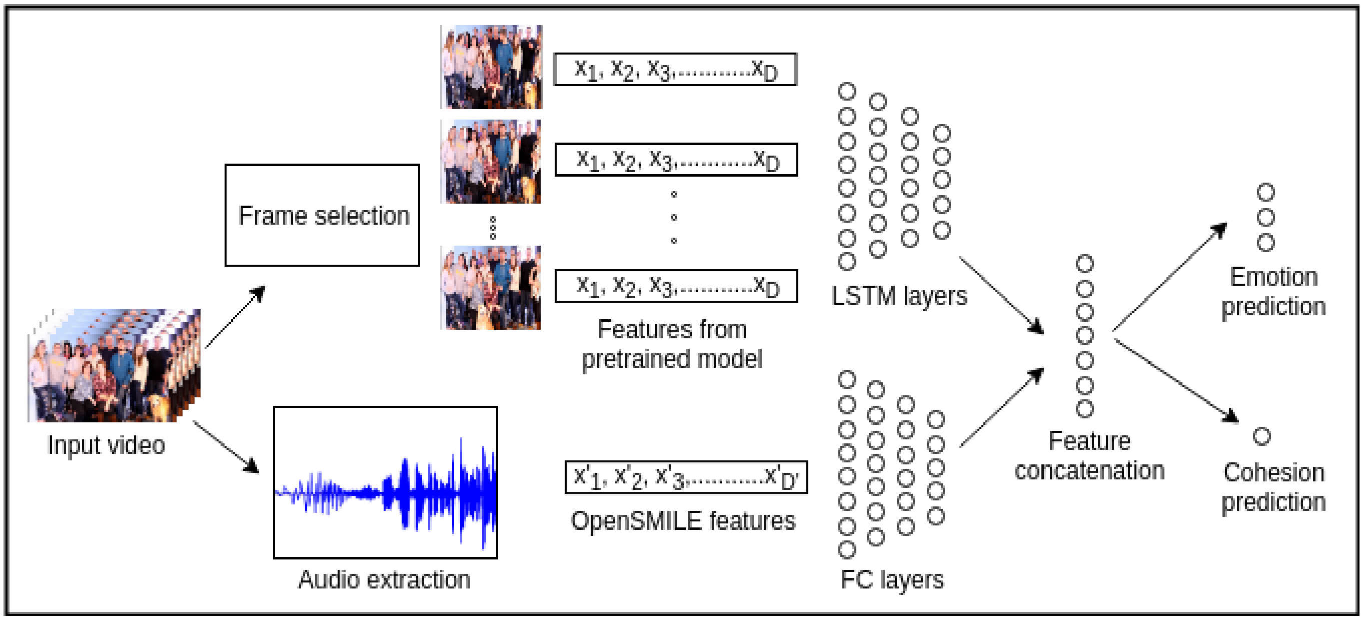

Frame selection
$x_1, x_2, x_3, \dots x_D$
$x_1, x_2, x_3, \dots x_D$
$x_1, x_2, x_3, \dots x_D$
Features from pretrained model
Input video
Audio extraction
$x'_1, x'_2, x'_3, \dots x'_{D'}$
OpenSMILE features
LSTM layers
FC layers
Feature concatenation
Emotion prediction
Cohesion prediction

Fig. 6. The proposed network architecture for emotion and cohesion prediction using audio and visual features.

trained on GAF-cohesion database [24]. The proposed baseline method is divided into three parts - using visual cues, audio and the ensemble of two, which are explained in detail below:

A. Visual model

The appearance-based information present in videos plays an important role to decide the emotion present in them. The variations present in the video across time provides additional information which is to be exploited. While on the other hand, encoding of these variations adds extra complexity to the emotion prediction model. It becomes difficult to train a network to understand the overall affect of the group of people shown across frames. To extract the relevant visual features for the task of group level emotion as well as cohesion, we use the pre-trained network for each frame with different variations. The variants of visual baseline are as follows:

1) We use the pre-trained model to obtain the frame wise emotion and cohesion prediction [24]. This step is used only to identify the performance of a model trained on images on videos. Hence, no learning is done in this step. The maximum of emotion and average of cohesion value is selected as the overall label for a video from the obtained frame wise prediction.
2) Similar to the above experiment, we extract frame wise second last layer features from the pre-trained Inception network and train a 4 layer small LSTM network for video level inference. In order to get uniform number of frames, we consider 672 frames which are the average of the frames across all videos. The four layers have the 256, 512, 1024 and 2048 number of kernels. A softmax and sigmoid activation are used in the prediction layers for classification and regression tasks, respectively. Both emotion and cohesion are trained simultaneously in a single network.

B. Audio model

To analyze the group level affect, audio features also play an important role as the use of only facial expressions may mislead towards the overall group affect estimation. Pitch, speech rate, duration, etc. are proved to be relevant to affect analyses. In group setting, these features are important to distinguish situations like an argument and a discussion where the visual model may fail. To leverage this, we extract INTERSPEECH 2013 ComParE challenge feature set from OpenSMILE [30] toolkit[2], which are widely used in the INTERSPEECH challenge [31]. These set of features include voice related low level descriptors along with energy, spectral, cepstral, logarithmic harmonic-to-noise ratio, spectral harmonicity and psychoacoustic spectral sharpness [31]. 6373 features are extracted for each video and training is done by using 5 Fully Connected (FC) layers of size 128, 256, 512, 1024 and 2048. Similar to the visual model, the network is jointly trained for emotion and cohesion task with softmax and sigmoid activation, respectively.

C. Ensemble of audio and video modalities

To exploit the presence of both audio and visual information in the data, both features are fused together. The fusion is done by concatenating the features obtained from the last layers of LSTM and FC layer, resulting in 4096 features (2048 from each audio and visual model). The direct prediction is obtained after fusing the features by using softmax and sigmoid activation for classification and regression. Fig. 6

[2] https://www.audeering.com/opensmile/

TABLE III
RESULTS OF THE PROPOSED BASELINE METHOD FOR VAL AND TEST SET.

Method	Emotion (%)		Cohesion (MSE)	
	Val	Test	Val	Test
Pre-trained Inception V3 Prediction	46.97	44.5	0.077	0.054
Audio only	50.23	48	1.217	0.840
Video only	52.09	42.00	1.097	0.716
Audio-video	50.23	47.50	1.09	1.738

shows the proposed pipeline which is used as a baseline for VGAF database.

VI. RESULTS AND DISCUSSION

This section will discuss the experimental details for the proposed baseline network explained in earlier section. For all the experiments, we train the same network for the joint prediction of emotion and cohesion. Emotion is considered as three class classification task whereas cohesion is treated as a regression problem similar to the [24]. For all the experiments we have used Keras[3] library with Tensorflow backend.

To train the network for joint prediction, we perform few baseline experiments on VGAF database. Table III shows the results corresponding to these experiments. The Inception V3 prediction represents the results obtained from direct prediction of pre-trained model on each frame. The videos have a large number of variations across time hence, the model trained on images fails to capture those changes for group level emotion. The direct prediction results in 46.97% accuracy and 0.077 MSE value on the val set. However, Inception V3 results in 0.077 and 0.054 MSE for val and test set, respectively.

The use of the audio-video features is found to be more useful to predict the emotion and cohesion of a group of people. From Table III, it is observed that using only the audio features results in 50.23% and 48% val and test accuracy which is greater than the visual appearance. Visual information provides 52.09% and 42% emotion accuracy on val and test set, respectively.

It is noticed in Table III that the emotion and cohesion prediction are almost directly proportional to each other. However, it is contrary to the assumption that the ensemble of audio and visual features will help for better learning. The combination leads to a drop in the performance of emotion and cohesion. The use of audio features is found to be very useful to predict the emotion and cohesion of a group of people. Alone, the audio features result in 48% and 0.840 MSE across test set for emotion and cohesion, respectively. One of the possible reason behind the poor performance of visual features may be that the pre-trained Inception V3 model is trained on images. The images have all the faces together in one frame. However, in the case of videos, the faces of all the people in a group are distributed across multiple frames. The model trained on images fails to capture the changes across frames.

TABLE IV
THE CONFUSION MATRIX (IN %) ACROSS AUDIO-VISUAL MODEL FOR GROUP LEVEL EMOTION ON TEST SET.

	Neutral	Positive	Negative
Neutral	10	33	57
Positive	9	45	46
Negative	8	22	70

In the case of cohesion, the holistic level of visual information is more important than others. Hence, it is evident that both audio and visual information is important for emotion and cohesion, respectively. But, the ensemble of the two modalities fails to incorporate the audio and visual level features together. Table IV shows the confusion matrix of the ensemble model on test set. The distribution is shown corresponding to the percentage of class-wise prediction to the total samples present in the database. From IV it is clear that ensemble of audio-visual model is unable to learn the neutral class for group level emotion.

The VGAF database, which is used to perform all the experiments consist of 1K videos. The videos present in the database have large variations which makes it difficult to train for the emotion and cohesion prediction. The number of faces present in the database also varies at a great extent. The database contains the videos for different situations which make it difficult to learn from a deep neural network which requires a large amount of similar data.

VII. CONCLUSION AND FUTURE WORK

In this study, we mainly explore video-based group level emotion and cohesion prediction task. It is an attempt to investigate the questions regarding group traits with the help of deep learning techniques and affective computing. Similar to Ponce et al. [32], group level emotion and cohesion are inferred from an early prediction perspective that is the first impression of a group. The proposed model primarily encode scene (along with facial expression information) for the prediction of group level emotion and cohesion.

There exist a number of situations where one can analyze the group of people. But the presence of the challenges (e.g. face tracking for all faces in a video, different social scenario, illumination variation, occlusion, different pose, presence of varied data etc.) are limiting its usage. Some of the possible (but not limited) future direction can be:

1) Using face level features along with the holistic features to perform better joint training of emotion and cohesion.
2) The effect of group structure, body pose, fashion quotient, kinship, personal attributes, personality and inter personal distance can also be used along with face level information. These factors can influence the group level traits to some extent.
3) Learning techniques like few shot learning can also be incorporated to extract information from different group settings.

[3] https://keras.io/

4) Further, the network can be optimized to reduce the memory usage and run time complexity making it suitable for real time applications.

Acknowledgment

We acknowledge the support of NVIDIA for providing us TITAN Xp G5X 12GB GPU for the research purpose.

References

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

[2] M. E. Shaw, *Group dynamics: The psychology of small group behavior*. McGraw Hill, 1971.

[3] 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 *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. IEEE, 2015.

[4] T. Treadwell, N. Lavertue, V. Kumar, and V. Veeraraghavan, "The group cohesion scale-revised: reliability and validity," *Journal of Group Psychotherapy, Psychodrama and Sociometry*, 2001.

[5] C. R. Evans and K. L. Dion, "Group cohesion and performance: A meta-analysis," *Small group research*, vol. 22, no. 2, pp. 175–186, 1991.

[6] 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.

[7] 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*, 2012.

[8] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard, "Mood meter: counting smiles in the wild," in *ACM UbiComp*, 2012.

[9] V. Vonikakis, Y. Yazici, V. Dung Nguyen, and S. Winkler, "Group happiness assessment using geometric features and dataset balancing," in *ACM International Conference on Multimodal Interaction*, 2016.

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

[11] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[12] 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*, p. 431, 2007.

[13] 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*, 2016.

[14] 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*, 2017.

[15] 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 *ACM International Conference on Multimodal Interaction*, 2017.

[16] W. Mou, O. Celiktutan, and H. Gunes, "Group-level arousal and valence recognition in static images: Face, body and context," in *IEEE Automatic Face and Gesture Recognition*, 2015.

[17] K. G. Quach, N. Le, K. Luu, C. N. Duong, I. Jalata, and K. Ricanek, "Non-volume preserving-based feature fusion approach to group-level expression recognition on crowd videos," *arXiv preprint arXiv:1811.11849*, 2018.

[18] 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 2017 ACM on Multimedia Conference*, 2017.

[19] W. Mou *et al.*, "Your fellows matter: Affect analysis across subjects in group videos," 2019.

[20] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, 2018.

[21] A. V. Carron and K. S. Spink, "The group size-cohesion relationship in minimal groups," *Small Group Research*.

[22] W. R. Thompson and D. P. Rapkin, "Collaboration, consensus, and detente: The external threat-bloc cohesion hypothesis," *Journal of Conflict Resolution,1981.*

[23] M. W. Rempel and R. J. Fisher, "Perceived threat, cohesion, and group problem solving in intergroup conflict," *International Journal of Conflict Management,1997.*

[24] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon, "Predicting group cohesiveness in images," *CoRR*, vol. abs/1812.11771, 2018. [Online]. Available: http://arxiv.org/abs/1812.11771

[25] 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*, vol. 38, no. 8, pp. 1707–1720, 2015.

[26] Y. Zhang, J. Olenick, C.-H. Chang, S. W. Kozlowski, and H. Hung, "Teamsense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 150, 2018.

[27] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Transactions on Multimedia,2015.*

[28] H. Tajfel, *Social identity and intergroup relations*. Cambridge University Press,2010.

[29] 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*, 2017.

[30] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[31] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[32] 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.