The More the Merrier: Analysing the Affect of a Group of People in Images

Abhinav Dhall[1,2], Jyoti Joshi[1], Karan Sikka[3], Roland Goecke[1,2] and Nicu Sebe[4]
[1] HCC Lab, Vision & Sensing Group, University of Canberra, Australia
[2] IHCC Group, RSCS, Australian National University, Australia
[3] University of California San Diego, USA
[4] University of Trento, Italy
abhinav.dhall@anu.edu.au, Jyoti.joshi@canberra.edu.au, karan.sikka@ucsd.edu, roland.goecke@ieee.org, sebe@disi.unitn.it

***Abstract*— The recent advancement of social media has given users a platform to socially engage and interact with a global population. With millions of images being uploaded onto social media platforms, there is an increasing interest in inferring the emotion and mood display of a group of people in images. Automatic affect analysis research has come a long way but has traditionally focussed on a single subject in a scene. In this paper, we study the problem of inferring the emotion of a group of people in an image. This group affect has wide applications in retrieval, advertisement, content recommendation and security. The contributions of the paper are: 1) a novel emotion labelled database of groups of people in images; 2) a Multiple Kernel Learning based hybrid affect inference model; 3) a scene context based affect inference model; 4) a user survey to better understand the attributes that affect the perception of affect of a group of people in an image. The detailed experimentation validation provides a rich baseline for the proposed database.**

I. INTRODUCTION

Groups are emotional entities and a rich source of varied manifestations of affect. The literature in social psychology suggests that group emotion can be conceptualised in different ways and is best represented by pairing the top-down approach (emotion emerging at the group level and followed by individual participants of the group) and bottom-up approach (overall emotion of group constructed by uniqueness of individual members' emotion expression) [1], [2]. This paper follows the group-as-a-whole perspective to capture elusive emotions arising from a group, broadly focussing on positive and negative affect in images. Automatic analysis of a group of people is an important problem as it has a wide variety of applications such as image retrieval, early event prediction, surveillance, image set visualisation, among others. The model proposed in this paper encompasses both scene information to determine the effect of the top-down approach and individuals' facial features to confirm the bottom-up method.

Affective computing has seen much progress in recent years, especially in automatic emotion analysis and understanding via verbal and non-verbal cues of an individual [3]. However, until recently, relatively little research has examined 'Group' emotion in images. To advance affective computing research, it is indeed of interest to understand and model the (perceived) affect exhibited by a group of people in images. The initial impediment in pursuing this research is obtaining the data, which should contain multiple participants exhibiting diverse emotions in real-world situations.

Fig. 1. Images of a group of people in a social event. The upper, middle and lower rows contain images, where the group displays a positive, neutral and negative affect, respectively.

Furthermore, it is required to create a framework, which can model the perceived affect of group of people in an image.

Advanced and inexpensive sensor technology has resulted in exponential growth in the number of images and videos being uploaded on the internet. This large pool of data enables us to explore the images containing multiple participants (e.g. Figure 1). Consider an image from a social event, such as a birthday party or images of a group of people watching a football game. For automatically organising, visualising and retrieval of these images, there are various cues such as colour, faces, and meta-data information that can be used. The presence of a group of people, posing for or being captured in a photograph, provides an opportunity of analysing the affect, which is perceived by a viewer. This paper describes a novel database and framework for affect classification of a group of people in an image. Note that, in this paper, we are interested in inferring the group affect perceived by the viewer of the image. The images in the proposed database have been collected from the WWW. No self evaluation of affect is available from the members of the group in the images. The terms 'affect' and 'emotion' are used interchangeably throughout this paper as

both are semantically similar terms for general representation of individual's feeling response [1].

In affect analysis, many interesting approaches [4], [5], [6], [7], [8] have been proposed for different problems ranging from expression / emotion inference, such as [9], [5], to medical applications such as pain detection [10] and entertainment [6]. However, apart from [8], other methods only deal with a single person. While prior work exists in analysing the non-verbal behaviour of a group of people in videos in social scenarios such as meetings [11], containing multiple persons. However, the problem, which this paper tries to tackle is different. We are interested at looking at images of people in social events; on the contrary, [11] look at group of people in videos for the understanding of their interaction.

The **key contributions** of this study are:

1) A novel hybrid framework for affect inference of a group of people in an image.
2) A labelled database containing images of groups of people in a wide variety of social events.
3) Multiple Kernel Learning (MKL) based fusion for adding local and scene context.
4) A user survey to understand the factors that affect the perception of affect displayed by a group.

II. Related Work

Recently, the study of a group of people in an image or a video has received much attention in the domain of computer vision. Generally, these methods can be broadly divided into two categories: a) **Bottom-up** methods: The subject's attributes are used to infer information at the group level [12], [13], [14]; b) **Top-down** methods: The group / sub-group information is used as a prior for inference of subject level attributes [15], [16], [17].

First, we look at the bottom-up methods. Recent work in crowd tracking [12] is based on trajectories constructed from the movement of people. Ge *et al.* [12] propose a hierarchical clustering algorithm, which detects sub-groups in crowd video clips. Cameras were installed at the MIT campus for inferring the mood of the passerby [13]. In [13], the scene level happiness is an average over the smiles of multiple people. However, in reality, group mood is not an averaging model [2].

In [8], [18], we proposed models and a database (HAPPEI) for inferring the happiness mood intensities of a group of people in images. The hypothesis is that there are certain attributes, which affect the perception of happiness of a group of people in an image. We proposed group expression models based on topic modelling and manually defined attributes. Inspired by Gallahger *et al.* [15], we represented a group as a min-span tree, in which the faces are the vertices and the weights of the edge define the distance between two faces. The context is based on a survey conducted [18]. Further analysis and details of the survey are in Section IV in this paper. Here, we extend the data and method from positive affect only [18] to a wider gamut of emotion (*Positive-Neutral-Negative*) of group of people. Furthermore, as compared to [18], where only faces were analysed, in this work, the effect of background/scene is analysed as well in a fusion framework.

In another interesting bottom-up method, Murillo *et al.* [19] proposed group classification for recognising urban tribes ('informal club', 'beach party' and 'hipsters' etc.). They used low-level features such as colour histogram and high-level features such as person attributes to learn a Bag of Words (BoW) based classifier. Hipster War [20] proposes a method based on analysing the clothes members of the group for their social membership type.

Gallahger *et al.* [15] proposed a top-down approach based on group derived contextual features age and gender inference of group members. Images were downloaded from the internet [15] and the performance in the experiments showed the positive effect of using group information. In another top-down approach, Wang *et al.* [21] model the social relationship between people standing together in a group for aiding recognition. The social relationships are inferred in unseen images by learning them from weakly labelled images. The authors learn a graphical model based on social-relationships such as '*father-child*', '*mother-child*' etc. and social-relationship features such as relative height, height difference and face ratio. Lu *et al.* [16] proposed a metric learning based method inferring the kinship relationship in images. In object detection and recognition work by Torralba and Sinha [22], context information of the scene and its relationship with the objects is described. Stone *et al.* [17] proposed conditional random fields based social relationship modelling between Facebook contacts for the problem of face recognition.

Inspired by the works mentioned above, we propose a hybrid approach, which combines top-down and bottom-up components. The rest of the paper is arranged as follows: the group database collection process is detailed in Section III. A survey to understand the attributes is described in Section IV. Section V discusses the group affect inference pipeline. Action Unit based representation is discussed in Section V-A. Low-level feature based face representation is represented in Section V-B. Scene Analysis for adding context to the pipeline is described in Section V-C. Experiments and implementation details are described in Section VI. Conclusion and future work discussion is presented in Section VII.

III. The Group Affect database

Over the years, several affect databases have been released. The earlier databases, such as Cohn-Kanade [23] or MMI [24], have led to significant contributions to the field. Each has a single subject posing a specific facial expression in lab-controlled settings (e.g. plain background, controlled illumination, no occlusion). Lately, databases capturing situations arising in the real-world environment (GENKI [5], Gallagher database [15], Acted Facial Expressions in the Wild (AFEW) [14], AM-FED [25]) have been released. GENKI [5] contains smiling/non-smiling pictures of celebrities collected from the internet. The 'in the wild' databases, Gallahger database [15] and AFEW [14], contain images/videos collected from

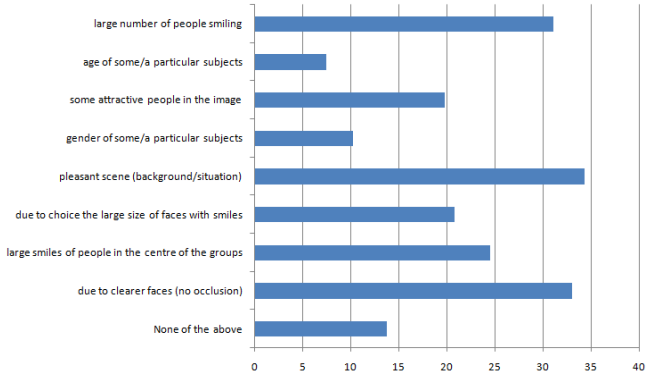

Fig. 2. Summary of responses in the survey.

Internet and movies, respectively. The AFEW database does contain a few videos clip samples of multiple subjects. The Gallagher database contains images of groups of people, which have been labelled for age and gender attributes. In [8], we proposed the HAPPEI database, which contains images of groups of people displaying happy expression only. To tackle the current problem, we collected a new database that was labelled for the perceived affect of a group of people in an image.

The proposed database was acquired by first searching Flickr.com and Google Images for images related to keywords[1], which describe groups and events. Some positive affect images were taken from our HAPPEI database [8]. Faces were then detected [26] on the downloaded images and the images containing fewer than two faces were rejected. Given the nature of the images in the database, the subjects of a group and their expressions can be quite heterogenous. Ideally, one may like to use automatic / semi-automatic methods such as topic discovery algorithms or parsing of related text for extracting labels. However, given the high intra-class variance due to different scenes and subjects, human annotator labels are being used for the database. This is hence a weakly-labelled problem. We would like to mention that for the AFEW database, where the initial labels were generated by parsing subtitles, a final pruning step was performed by the human labelers, while for the HAPPEI database, the labels were manually annotated by the labellers. In this work, the emotion of a group in an image is labelled as *Positive*, *Negative* or *Neutral*. These labels closely resemble the valence axis only ion the Valence-Arousal emotion scale. We used three human annotators for labelling the dataset. Any images without label consensus were removed. The database contains 504 images. Sample images from the proposed *Group Affect Database* are shown in Figure 1.

[1]The sample keywords were: *Tahrir Square*, *London Protest*, *Brazil Football Fans*, *Excited People*, *Happy People*, *Humanitarian Aid*, *Delhi Protest*, *Gaza Protest*, *Party Friends*, *Police Brutality*, *Celebration* etc.

IV. Survey

In order to understand the attributes, which affect the perception of affective state of a group, we conducted a user study [18]. Two sets of surveys were developed. In the first part, subjects were asked to compare two images for their apparent affect and rate the one with a higher positive affect. Further, they were asked various questions about the attibutes / reasons, which made them choose a specific image / group out of the two images / groups. In the second set, only a single image is shown and questions were asked about it.

A total of 149 subjects participated in this survey (94 males and 55 female subjects). Various questions were asked for e.g. *'How would you describe the expression of the group?'*, *'Was your choice motivated by: (multiple answers acceptable)'*, *'How would you describe the MOOD of the group?'*, *'What are the dominating attribute/characteristic of the leader(s) in the group that effect the group's mood?'*, *'Any other reason for your choice?'*. Figure 2 shows the average of the responses to the questions. It is evident that the location of the subjects, the number of smiles etc. play a dominating role in the perception of mood of a group. Figure 3 describes the dominating words in the response to the 'What is the dominating attribute / characteristic of the leader(s) in the group that affect your perception of the group's mood?' and 'Any other reason for your choice?'. Dominating words about the salient participants in the group

(a) Any other reason

(b) Subject attribute

Fig. 3. Most frequently occurring responses in the survey regarding any reason other than the questions asked (a) and the most salient subject (b).

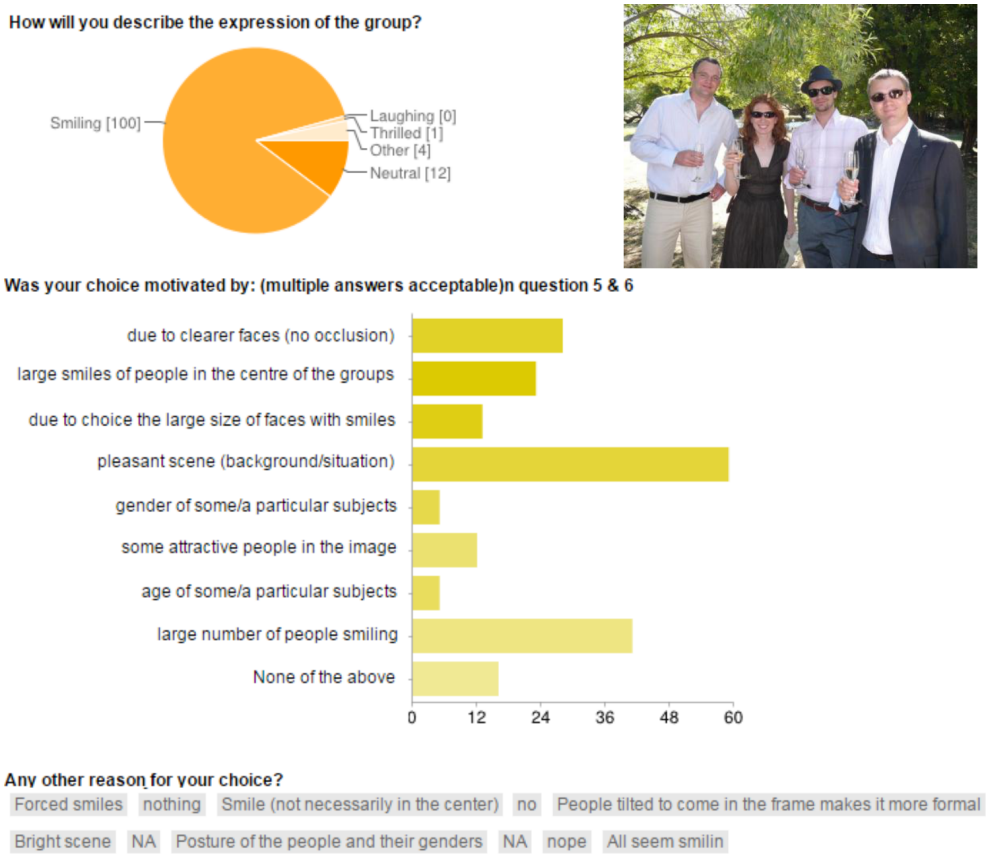

Fig. 4. Analysis of the subject responses of one of the survey images.

are smiling, smiles, attractive, eyes, beauty etc.

In Figure 4, 'Any other reason for your choice?', 50% of the subjects mentioned the reason for affect rating as the pleasant scene. In other examples, subjects mentioned attributes such as age, gender and attractiveness as attributes that affect their perception of the affect of a group of people in an image. Based on these observations, we propose a hybrid framework, which models the local features (face analysis Section V-A and Section V-B) and global features (scene descriptor V-C).

V. Pipeline

The proposed framework for inferring affect is based on multi-modal fusion using MKL. Figure 5 describes the flow of the proposed method. Below, we discuss each sub-system and its contribution to the overall framework.

A. Action Unit Based Face Representation

Facial Action Units (AU) describe the activation of facial muscles, when there is a change in the facial expression. Facial AU are one of the leading and most widely used representations for facial expression analysis. AU have been extensively used in inferring affect. Given an aligned face of a subject in a group, we compute AU features using the CERT toolbox [27]. CERT is based on computation of Gabor filter based features and SVM based classification. CERT is extensively used by the face analysis community for its real-time and stable performance. Furthermore, to model a group, we learn a BoW representation, in which each member's face is represented as a word and the group is represented as a document. The BoW representation has been extensively used in computer vision (e.g. [19]). It represents a document as a histogram of unordered frequency of words. We refer to the BoW formulation, where AU features are the words BoW_{AU}. AU here represent the attributes mentioned in the survey such as smiling, happy etc.

B. Low-level Features

Automatic AU detection is an open problem. The presence of varied backgrounds, head pose / movements, occlusion etc. in challenging conditions make the task even more difficult. Based on the survey (Section IV), along with facial expressions, there are several other facial attributes such as age, attractiveness, gender, presence of occlusion (e.g. beard, glasses etc.) that affect the perception of the affect of a group. Based on these two arguments, we perform feature augmentation by computing low-level features on the aligned faces. Pyramid of Histogram of Gradients (PHOG) [28] and Local Phase Quantization (LPQ) [29] descriptors are computed over an aligned face. PHOG is computed

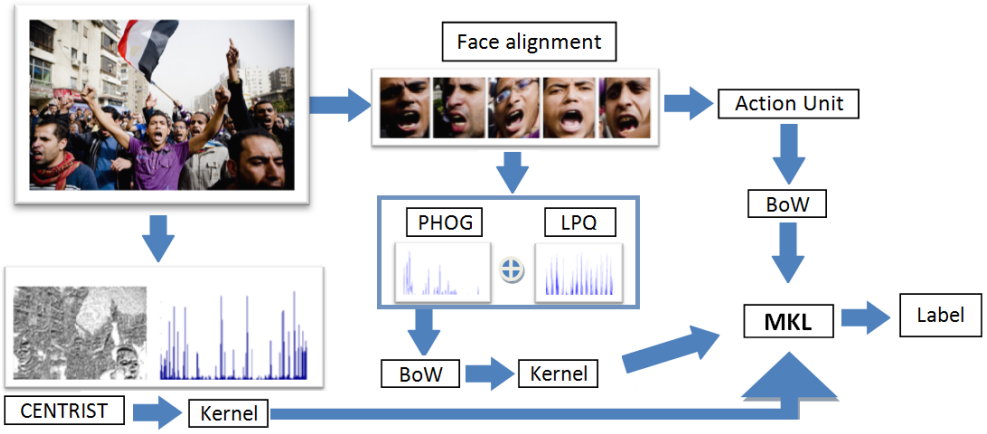

Fig. 5. The proposed group affect inference pipeline.

by applying edge detection on an image, followed by a histogram computation in a pyramid fashion. For computing LPQ, local binary patterns are computed over the coefficient extracted by applying short Fourier transform on an image. The feature combination of these two features is robust to scale and illumination changes [30].

Similar to BoW_{AU}, a BoW representation is learnt on the low-level features. This feature is referred to as BoW_{LL}. Furthermore, feature fusion is performed (details in Section VI) between BoW_{AU} and BoW_{LL}. This fusion based representation can also be viewed as adding implicit information (low-level based attributes) to manually-defined user attributes (AU based representation). This is similar to the automatic action recognition approach defined in [31].

C. Scene Context

Deduced from the survey performed, scene information plays a vital role in the perception of the group affect. We investigate the usefulness of scene analysis features for adding global information to our model. Two widely used scene analysis descriptors are compared – GIST [32] and CENsus TRansform hISTogram (CENTRIST) [33] – w.r.t. the problem of affect analysis of a group of people. GIST and CENTRIST model the scene at a holistic level. We refer to GIST and CENTRIST descriptor representation as $Scene_{GIST}$ and $Scene_{CENTRIST}$, respectively. The scene feature computes statistics at a global level. It takes into consideration not only the background but information that may define the situation such as clothes.

Fusion is performed between BoW_{AU}, BoW_{LL} and scene features. This results in a hybrid framework, where the face analysis subsystem represents a bottom-up approach and scene context analysis represents a top-down approach. This model, in fact, takes inspiration from Moshe Bar's scene context model [34], where two concurrent streams are used for scene processing: a low-resolution holistic representation and a detailed object level representation. In this paper, the low-resolution representation is similar to the scene descriptor analysis. The scene descriptors are computed quickly and give a sense of the context. The same is also confirmed by the human studies conducted by Li *et al.* [35]. The second stream deals with salient objects, which in our problem is the face analysis framework.

D. Multiple Kernel Learning for Fusion

We wish to fuse: BoW_{AU}, BoW_{LL}, $Scene_{GIST}$ and $Scene_{CENTRIST}$. A trivial way is to perform feature fusion. However, it is not guaranteed that the complimentary information will be captured. There may be no increase in the performance of a multi-modal system due to increase in the feature dimension. One obvious option is to apply some feature selection / dimension reduction method. Decision level fusion could also be more promising. Recently, [10] have successfully used MKL for audio-video emotion recognition. MKL computes a linear combination of base kernels. We use the MKL method of [36], as it poses the MKL problem as a convex optimisation problem, which guarantees an optimal solution.

We represent each modality using a single kernel. The notations below are similar to [10]. Given a training set with N group images, let us denote the M feature sets (modalities) as X_m, $m \in \{1, 2, .., M\}$. The labels are represented as $y_i \in \{-1, 1\}$ for $i = 1, ..., N$. For each modality, we compute a Histogram Intersection Kernel (HIK), which can be defined as $k(\mathbf{x_i}, \mathbf{x_j}) = \sum_{i=1}^{n} \min(x_i, x_j)$. The dual formulation of SVM optimisation problem can be written as:

$$\max_{\alpha,\beta} \left[\sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathrm{K}(x_i, x_j) \right] \tag{1}$$

$$\sum_{i}^{N} \alpha_i y_i = 0; 0 \leq \alpha_i \leq C$$

$$\mathrm{K} \succ 0$$

Feature	Positive	Neutral	Negative	Final
BoW_AU	70.93	33.33	37.93	50.43
BoW_LL	76.74	56.66	06.90	50.98
Scene_GIST	52.32	38.33	31.03	42.16
Scene_CENTRIST	50.00	45.00	39.65	45.58

TABLE I

CLASSIFICATION ACCURACIES (IN %) ON THE THREE-CLASS TASK FOR INDIVIDUAL MODALITIES.

Feature	Positive	Neutral	Negative	Final
BoW_LL + BoW_AU + Scene_GIST	63.95	38.33	46.55	51.47
BoW_LL + BoW_AU	86.04	31.66	20.68	51.47
BoW_LL + BoW_AU+ Scene_CENTRIST	51.12	48.33	44.82	48.52
MKL - BoW_LL + BoW_AU + Scene_GIST	82.55 (0.0083)	78.33 (0.7993)	50.00 (0.1924)	67.15
MKL - BoW_LL + BoW_AU + Scene_CENTRIST	83.72 (0.0085)	80.00 (0.7976)	31.03 (0.1938)	67.64

TABLE II

CLASSIFICATION ACCURACIES (IN %) WHEN FUSING THE BEST PERFORMING INDIVIDUAL MODALITIES. FOR MKL, THE VALUES INSIDE THE BRACKET ARE THE LEARNT KERNEL WEIGHTS.

where K is defined as the convex combination of all feature kernels:

$$\mathrm{K} = \sum_{m}^{M} \beta_m \mathrm{k}_m \tag{2}$$

$$\sum_{m}^{M} \beta_m = 1$$

$$\beta_m \geq 0 \ \forall m$$

Grid search is used to find the parameter C. α and β are automatically learnt by MKL [36].

VI. EXPERIMENTAL ANALYSIS

Face processing pipeline: Face and fiducial points are detected using the publicly available[2] MoPS library. A 146-parts model is used for fitting. A total of 1756 faces were detected in 417 images. The data are divided into two sets. Affine warp is computed on the detected faces. Aligned face are downsized to 128×128 size. For the PHOG descriptor[3], bin size = 16, orientation range [0-180], pyramid level = 3. For the LPQ descriptor, rotation invariant LPQ[4] is computed. Feature fusion is performed using LPQ and PHOG. To learn dictionaries for BoW_{LL} and BoW_{AU}, a wide range of dictionary sizes are experimented with. The range of dictionary size is [8-128]. The final dictionary size for BoW_{AU} and BoW_{LL} are 64 and 128, respectively.

Scene descriptors: The publicly available GIST implementation[5] is used with its default parameters: Orientations per scale = [8 8 8 8]; number of blocks = 4. Similarly for the CENTRIST descriptor, publicly available implementation[6] is used.

The LibSVM library [37] with HIK is used for learning classifiers for individual modalities. The cost parameters are set empirically using grid search. Table I shows the classification accuracy for the individual descriptors: BoW_{AU}, BoW_{LL}, $Scene_{GIST}$, $Scene_{CENTRIST}$. It is interesting to see that high-level features based on AU perform similar to the low-level feature combination. The classification accuracy for the *Negative* class is lower for the other two classes. On analysing the confusion matrices, it was found that the number of *Negative* samples getting incorrectly classified as *Positive* or *Neutral* is almost the same. Further investigation revealed one of the possible reasons. For collecting the *Negative* images from the internet, keywords such as 'protest', 'violence', 'unhappy' etc. were used. Generally, in scenarios like this, subjects are not directly posing in front of the camera. Their body gesture may also be intruding with face visibility. For e.g. in protest images, members of a group, generally, raise their arms and may hold placards. This occludes the face at times. We observed that there are more non-frontal faces in the *Negative* class. This leads to error in the face alignment step and the error propagates through the system. Chew *et al.* [38] argue that a small error in face fitting can be compensated with good descriptors. Probably, that is the reason why BoW_{AU}'s *Negative* class performs better than that of BoW_{LL}. CERT has individual detectors for each AU.

For the fusion of the individual modalities, we performed feature fusion and fusion using MKL. The MKL implementation is publicly available[7]. Table II shows the classification accuracy output. As hypothesised, MKL performs better than feature fusion. The MKL based fusion of BoW_{AU}, BoW_{LL} and $Scene_{CENTRIST}$ performs the best out of all the tested configurations. MKL based on BoW_{AU}, BoW_{LL} performs less than MKL based fusion of BoW_{AU}, BoW_{LL} and $Scene_{CENTRIST}$, it is to be understood here, that given the 'in the wild nature of the images, face detection may fail, so we need to add complementary information, which we obtain from the scene level descriptors. The weights learnt for the three kernels are mentioned in Table II.

VII. CONCLUSIONS

In this paper, a novel framework for analysing affect of a group of people in an image has been presented. An 'in the wild' image based database labelled with perceived affect of group of people is collected from the Internet based on keywords search. In order to understand the factors that effect the perception the of emotion in a group, a user study is consulted. A Multiple Kernel Learning based hybrid affect inference method is proposed. Top-down information is extracted using scene descriptors. Bottom-up information is extracted by analysing the face of the members of a

[2] http://www.ics.uci.edu/~xzhu/face/

[3] http://www.robots.ox.ac.uk/~vgg/research/caltech/phog.html

[4] http://www.cse.oulu.fi/CMV/Downloads/LPQMatlab

[5] http://people.csail.mit.edu/torralba/code/spatialenvelope/

[6] https://github.com/sometimesfood/spact-matlab

[7] http://www.cse.msu.edu/~bucakser/software.html

group. Facial Action Unit based Bag of Words features are augmented with low-level based Bag of Words features. The MKL framework considers each feature modality as a separate Histogram Intersection Kernel and performs better than feature fusion methods. This supports the hypothesis of the paper that facial information (bottom-up) and scene information (top-down) together help in inferring the affect conveyed by a group. The framework is backed by the model proposed by Bar [34] and the survey (Section IV). To the best of our knowledge, this work is the first to analyse both positive and negative affect of groups of people in images at social events.

Affect analysis of groups of people is a non-trivial problem. Large heterogeneity due to uniqueness of individuals' expression of emotion poses one of the biggest challenges in assessing the group emotion. Recently, body pose analysis for inferring affect [39] has got much attention. An interesting extension is to fuse body pose information. Along with providing new information about the affect, this may help in scenarios where the face detection is not accurate due to pose, occlusion or blur. To deal with non-frontal faces, head pose normalisation methods such as the ones based on MoPS [40] (as in Section VI face and fiducial points are detected using MoPS) can be used. Explicit modelling of attributes such as clothes style and colours and kinship relationship can also give important information about the social event and hence, can aid in inferring the affect of a group of people in an image. Currently, the affect classes in the group affect database are positive, neutral and negative. In future, the database will be extended and finer emotion labels and intensities will be added.

VIII. Acknowledgement

Nicu Sebe has been supported by the MIUR FIRB project S-PATTERNS and by the MIUR Cluster project Active Ageing at Home.

References

[1] S. G. Barsäde and D. E. Gibson, "Group emotion: A view from top and bottom," *Deborah Gruenfeld, Margaret Neale, and Elizabeth Mannix (Eds.),* Research on Managing in Groups and Teams, vol. 1, pp. 81–102, 1998.

[2] J. R. Kelly and S. G. Barsade, "Mood and emotions in small groups and work teams," *Organizational behavior and human decision processes*, vol. 86, no. 1, pp. 99–130, 2001.

[3] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[4] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *Proccedings of the Conference on Computer Vision and Pattern Recognition-Workshops (CVPRW)*, pp. 76–76, 2005.

[5] J. Whitehill, G. Littlewort, I. R. Fasel, M. S. Bartlett, and J. R. Movellan, "Toward Practical Smile Detection," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2106–2111, 2009.

[6] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 505–523, 2011.

[7] T. Gehrig and H. K. Ekenel, "Facial action unit detection using kernel partial least squares," in *Proceedings of the IEEE International Confernece on Computer Vision and Workshops (ICCW)*, pp. 2092–2099, 2011.

[8] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding Happiest Moments in a Social Context," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 613–626, 2012.

[9] F. Zhou, F. de la Torre, and J. F. Cohn, "Unsupervised Discovery of Facial Events," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2581, 2010.

[10] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pp. 517–524, 2013.

[11] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[12] W. Ge, R. T. Collins, and B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Transaction on Pattern Analysis & Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, 2012.

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

[14] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, p. 0034, 2012.

[15] A. C. Gallagher and T. Chen, "Understanding Images of Groups of People," in *Proceedings of the IEEE Confernece on Computer Vision and Pattern Recognition (CVPR)*, pp. 256–263, 2009.

[16] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.-P. Tan, and G. Wang, "Neighborhood repulsed metric learning for kinship verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2594–2601, IEEE, 2012.

[17] Z. Stone, T. Zickler, and T. Darell, "Autotagging facebook: Social network context improves photo annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

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

[19] A. C. Murillo, I. S. Kwak, L. Bourdev, D. J. Kriegman, and S. Belongie, "Urban tribes: Analyzing group photos from a social perspective," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition and Workshops (CVPRW)*, pp. 28–35, 2012.

[20] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 472–488, Springer, 2014.

[21] G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth, "Seeing people in social context: Recognizing people and social relationships," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 169–182, 2010.

[22] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 763–770, 2001.

[23] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 46–53, 2000.

[24] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 5–pp, 2005.

[25] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected" in-the-wild"," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, pp. 881–888, IEEE, 2013.

[26] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2886, 2012.

[27] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in

Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 298–305, 2011.

[28] A. Bosch, A. Zisserman, and X. Munoz, "Representing Shape with a Spatial Pyramid Kernel," in *Proceedings of the ACM international conference on Image and video retrieval (CIVR)*, pp. 401–408, 2007.

[29] V. Ojansivu and J. Heikkilä, "Blur Insensitive Texture Classification Using Local Phase Quantization," in *Proceedings of the Image and Signal Processing (ICISP)*, pp. 236–243, 2008.

[30] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proceedings of the IEEE Conference Automatic Faces & Gesture Recognition and Workshops (FERA)*, pp. 878–883, 2011.

[31] G. Tsai, C. Xu, J. Liu, and B. Kuipers, "Real-time indoor scene understanding using bayesian filtering with motion cues," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 121–128, 2011.

[32] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[33] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[34] M. Bar, "Visual objects in context," *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[35] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *Journal of vision*, vol. 7, no. 1, p. 10, 2007.

[36] S. Bucak, R. Jin, and A. K. Jain, "Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition," in *Advances in Neural Information Processing Systems*, pp. 325–333, 2010.

[37] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[38] S. W. Chew, P. Lucey, S. Lucey, J. M. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan, "In the pursuit of effective affective computing: The relationship between features and registration," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 4, pp. 1006–1016, 2012.

[39] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: a survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[40] A. Dhall, K. Sikka, G. Littlewort, R. Goecke, and M. Bartlett, "A Discriminative Parts Based Model Approach for Fiducial Points Free and Shape Constrained Head Pose Normalisation In The Wild," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 1–8, 2014.