Finding Happiest Moments in a Social Context

Abhinav Dhall[1], Jyoti Joshi[2], Ibrahim Radwan[2] and Roland Goecke[2,1]

[1] IHCC Group, Research School of Computer Science, Australian National University
[2] Vision & Sensing Group, Faculty of ISE, University of Canberra, Australia
abhinav.dhall@anu.edu.au, {jyoti.joshi, ibrahim.radwan}@canberra.edu.au,
roland.goecke@ieee.org

Abstract. We study the problem of expression analysis for a group of people. Automatic facial expression analysis has seen much research in recent times. However, little attention has been given to the estimation of the overall expression theme conveyed by an image of a group of people. Specifically, this work focuses on formulating a framework for happiness intensity estimation for groups based on social context information. The main contributions of this paper are: a) defining automatic frameworks for group expressions; b) social features, which compute weights on expression intensities; c) an automatic face occlusion intensity detection method; and d) an 'in the wild' labelled database containing images having multiple subjects from different scenarios. The experiments show that the global and local contexts provide useful information for theme expression analysis, with results similar to human perception results.

1 Introduction

Expression analysis has been a long studied problem for its importance in human-computer interaction, affective computing and security. Research has been focussing on inferring the emotional state of a single subject only. In a social event, such as a party, a wedding or a graduation ceremony, many pictures are taken, which contain multiple subjects or groups of people. In this paper, we propose a framework for computing the *theme expression*, which a group of people convey, with a particular focus on their level of happiness. Here, we are interested in knowing an individual's intensity of happiness and its contribution to the overall mood of the scene. The contribution towards the theme expression can be affected by the social context. The context can constitute various global and local factors (such as the relative position of the person in the image, their distance from the camera and the level of face occlusion). We model this global and local information based on a group graph (Figure 1). We embed these features in our method and pose the problem in a probabilistic graphical model framework based on a relatively weighted soft-assignment.

Analysing the theme expression conveyed by images of groups of people is an unexplored problem that has many real-world applications: event summarisation and highlight creation, candid photo shot selection, expression apex detection in video, video thumbnail creation, etc. One basic approach is to average the

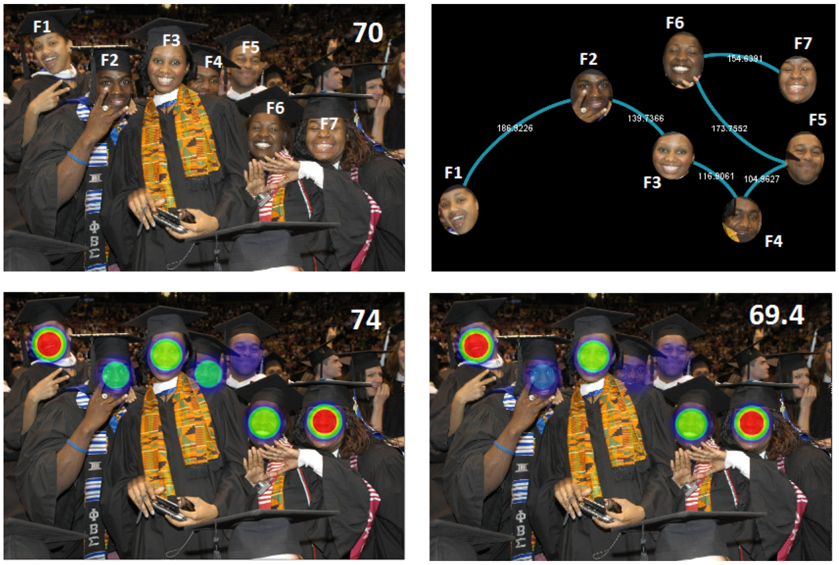


Fig. 1. *Group Expression Model.* The figure describes the effect of social attributes on the expression intensity estimation of the image. *Top Left:* Original image, which a human labeler gave a happiness intensity score of 70. *Top Right:* Min-span tree depicting connection between faces where the edge weight is the distance between the faces. *Bottom Left:* Happiness intensity heat map using averaging; note how some occluded faces are given a high weighting. *Bottom Right:* Happiness intensity heat map with social context, the contribution of the **occluded faces** (**F2** and **F4**) towards the overall intensity of the group is penalised.

happiness intensities of all people in the group. However, group dynamics define where people stand and how much of their face is visible. These social attributes play an important role in defining the overall happiness an image conveys.

Facial expressions result from the movement in facial muscles in response to a person's affective state, intention, or social communication. Much research has been done on facial expression recognition, e.g. [1]. [2] present an interesting approach for smile detection and intensity computation. They proposed a new image based database labelled for smiling and non-smiling images and evaluated several state of the art methods for smile detection. In all these works, the faces are considered independent of each other, i.e. single subject per sample.

Related work

Only recently, social context has started to be explored for face analysis. [3] proposed contextual features based on group structure for computing the age and gender. The global attributes described in our paper are similar to their contextual features of social context, but their problem is inverse to ours. [3] compute contextual features based on the structure of the group for better age and gender recognition. The approach can be termed as a top-down method: from global context to better local decisions. On the other hand, our approach can be seen as bottom-up. Happiness intensities and relative weights based on the global and local context are computed for each face, leading to a better prediction of the overall mood of a group. [4] model the social relationship between people standing together in a group for aiding recognition and the social relationships in unseen images by learning them from weakly labelled images.

In face recognition [5], social context is employed to model relationships betweens people such as friends on Facebook, using a Conditional Random Field. Hoque *et al.* [6] installed cameras at four locations on campus and tried to estimate the mood of people looking into the camera and compute a mood map for the campus. However, the scene level happiness is an averaging of individual person's smile. Recently, Aggarwala *et al.* [7] proposed a framework for selecting candid images from a video of a single person. Their approach is based on various factors such as face clarity, eye blink and expression. The approach is limited to one person only.

To the best of our knowledge, this is the first work covering theme expressions of groups that takes the social context into account. There are several factors which effect group level expression analysis. Local factors (individual subject level): age, gender, face visibility, face pose, eye blink, etc. Global factors: where do people stand, with whom do people stand. In this paper, the focus is on face visibility, smile intensity and relative face size and relative face distance. Further, we require labelled data of image containing groups of people. For getting a variety of data, we use Flickr images.

The key-contributions of the paper are:

1. Frameworks for automatic happiness intensity estimation of a group of people based on social features.
2. Face level global and local attributes are mapped to the attributes space. Manual and data-driven attributes are modelled in a Bayesian framework.
3. An occlusion intensity estimation method, which handles partial occlusion.
4. A labelled 'in the wild' database containing images of groups of people.

The remainder of the paper is organised as follows: Section 2 discusses first a naive group expression model based on averaging. Then, social context is discussed in the form of global and local features. The min-span tree is computed to define the relative neighbourhood in a group for a face. The location and relative size of a face define its global contribution. Local social context is defined by an individual person's happiness and occlusion intensities. The global and local context are applied as weights to the group expression model. In Section 3, we project the social context as attributes. The manual attributes are combined with data-driven attributes in a supervised hierarchical Bayesian framework. Section 4 discusses the new database containing images of groups of people. Section 5 discusses the results of the proposed frameworks, which includes both quantitative and qualitative experiments. The two proposed group expression models, GEM_{LDA} and GEM_w, are compared.

2 Group Expression Model

Given an image $\mathbf{I}$ containing a group of people $\mathcal{G}$ of size $\mathbf{s}$ and their happiness intensity level $\mathcal{I_H}$, a simple *Group Expression Model (GEM)* can be formulated as an average of the happiness intensities of all faces in the group

$$\text{GEM} = \frac{\sum_i \mathcal{I_H}_i}{s} \qquad . \tag{1}$$

In this simple formulation, both global information, e.g. the relative position of the people in the image, and local information, e.g. the level of occlusion of a face, are being ignored. We wish to add these social context features as weights to the process of determining the happiness intensity of a group image. We will show in Section 5 that the simple GEM formulation has a larger error as compared to the GEM that considers the global and local context.

2.1 Global context via group graph

We consider the tip of the nose as the position $\mathbf{p_i}$ of a face f_i in the image. To map the global structure of the group, a fully connected graph $G = (V, E)$ is constructed. Here, $V_i \in \mathcal{G}$ represents a face in the group and each edge represents the link between two faces $(V_i, V_m) \in E$. The weight $w(V_i, V_m)$ is the Euclidean distance between $\mathbf{p_i}$ and $\mathbf{p_m}$. We compute Prim's minimal spanning tree algorithm on $\mathcal{G}$, which provides us with the information about the relative position of people in the group with respect to their neighbours. In Figure 1, the min-span tree of the group graph is displayed.

Once, the location and minimally connected neighbours of a face are known, the relative size of a face f_i with respect to its neighbours is calculated. The size of a face is taken as the distance between the location of the eyes, $d_i = ||\mathbf{l} - \mathbf{r}||$. The relative face size $\theta_\mathbf{i}$ of f_i in region r is then given by

$$\theta_\mathbf{i} = \frac{d_i}{\sum_i d_i / n} \tag{2}$$

where the term $\sum_i d_i/n$ is the mean face size in a region r around face f_i, with r containing a total of n faces including f_i. Generally speaking, the faces which have a larger size in a group photo are of the people who are standing closer to the camera. Here, we assume that the expression intensity of the faces closer to the camera contributes more to the overall group expression intensity as compared to the faces of people standing in the back. Eichner and Ferrari [8] made a similar assumption to find if a person is standing in the foreground or at the back in a multiple people pose detection scenario.

Based on the centre locations $\mathbf{p_i}$ of all faces in a group $\mathcal{G}$, the centroid $\mathbf{c_g}$ of $\mathcal{G}$ is computed. The relative distance $\delta_\mathbf{i}$ of each face f_i is described as

$$\delta_\mathbf{i} = ||\mathbf{p}_i - \mathbf{c}_g|| \qquad . \tag{3}$$

δ_i is further normalised based on the mean relative distance. Faces closer to the centroid are given a higher weighting than the faces further away. Using Equations 2 and 3, we assign a global weight to each face in the group

$$\psi_{\mathbf{i}} = ||1 - \alpha\delta_i|| * \frac{\theta_i}{2^{\beta-1}} \tag{4}$$

where α and β are the parameters, which control the effect of these weighting factors on the global weight. Figure 1 demonstrates the effect of the global context on the overall output of GEM.

2.2 Local context

Above, we have defined the global context features, which compute weights on the basis of two factors: (1) where are people standing in a group and (2) how far are they away from the camera. The local context is described in terms of an individual person's level of face visibility and happiness intensity.

Occlusion Intensity Estimate: Occlusion in faces, whether self-induced (e.g. sunglasses) or due to interaction between people in groups (e.g. one person standing partially in front of another and occluding the face), is a common problem. Lind and Tang [9] introduced an automatic occlusion detection and rectification method for faces via GraphCut-based detection and confident sampling. They also proposed a face quality model based on global correlation and local patterns to derive occlusion detection and rectification.

The presence of occlusion on a face reduces its visibility and, therefore, hampers the clear estimation of facial expressions. It also reduces the face's contribution to the overall expression intensity of a group portrayed in an image. Based on this local face level phenomenon, the happiness intensity level $\mathcal{I}_{\mathcal{H}}$ of a face $\mathbf{f_i}$ in a group is penalised if (at least partially) occluded. Thus, along with an automatic method for occlusion detection, an estimate of the level of occlusion is required. Unlike [9], we propose to learn a mapping model $\mathcal{F} : \mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ are the descriptors calculated on the faces and $\mathbf{Y}$ is the amount of occlusion.

The mapping function $\mathcal{F}$ is learnt using the Kernel Partial Least Squares (KPLS) [10] regression framework. The PLS set of methods have recently become very popular in computer vision [11–13]. Schwartz *et al.* [12, 13] use PLS for dimensionality reduction as a prior step before classification. Guo and Mu [11] use KPLS based regression for simultaneous dimensionality reduction and age estimation. In our problem of estimating the amount of occlusion, the training set $\mathbf{X}$ is a set of input samples x_i of dimension N. $\mathbf{Y}$ is the corresponding set of vectors y_i of dimension M. Then for a given test sample matrix X_{test}, the estimated labels matrix $\hat{Y}$ is given by

$$\hat{\mathbf{Y}} = \mathbf{K}_{test}\mathbf{R} \tag{5}$$

$$\mathbf{R} = \mathbf{U}(\mathbf{T}^T\mathbf{K}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} \tag{6}$$

where $\mathbf{K}_{test} = \mathbf{\Phi_{test}}\mathbf{\Phi}^T$ is the kernel matrix for test samples. $\mathbf{T}$ and $\mathbf{U}$ are the $n \times p$ score matrices of the p extracted latent projections. See Rosipal [10] for details and derivation of KPLS.

The input sample vector x_i is a normalised combination of Hue, Saturation and the Pyramid of Histogram of Gradients (PHOG) [14] for each face. In the training set, $\mathbf{X}$ contains both occluded and non-occluded faces, $\mathbf{Y}$ contains the labels identifying the amount of occlusion (where 0 signifies no occlusion). The labels were manually created during the database creation process (Section 4). The output label y_i is used to compute the local weight λ_i, which will penalise $\mathcal{I}_\mathcal{H}$ for a face f_i in the presence of occlusion. It is defined as

$$\lambda_\mathbf{i} = ||1 - \gamma y_i|| \tag{7}$$

where γ is the parameter, which controls the effect of the local weight.

Happiness Intensity Computation: A regression based mapping function $\mathcal{F}$ is learnt using KPLS for regressing the happiness intensity of a subject's face. The input feature vector is the PHOG descriptor computed over aligned faces. As discussed earlier, the advantage of learning via KPLS is that it performs dimensionality reduction and prediction in one step. Moreover, KPLS based classification has been successfully applied to facial action units [15].

2.3 Weighted GEM

We combine the global and local contexts defined in Eq. 4 and Eq. 7 to formulate the relative weight for each face f_i as

$$\pi_i = \lambda_i \psi_i \tag{8}$$

This relative weight is applied to the $\mathcal{I}_\mathcal{H}$ of each face in the group $\mathcal{G}$ and based on Eq. 1 and Eq. 8, the new weighted GEM is defined as

$$\text{GEM}_w = \frac{\sum_i \mathcal{I}_{\mathcal{H}i}\pi_i}{s} \tag{9}$$

This formulation takes into consideration the structure of the group and the local context of the faces in it. The contribution of each face's f_i $\mathcal{I}_{\mathcal{H}i}$ towards the overall mood of the group is weighted relatively.

3 Social Context as Attributes

The social features described above can also be viewed as manually defined attributes. Lately, attributes have been very popular in the computer vision community (e.g. [16]). Attributes are defined as high-level semantically meaningful representations. They have been, for example, successfully applied to object recognition, scene analysis and face analysis. Based on the regressed happiness intensities, we define the attributes as *Neutral*, *Small Smile*, *Large Smile*, *Small*

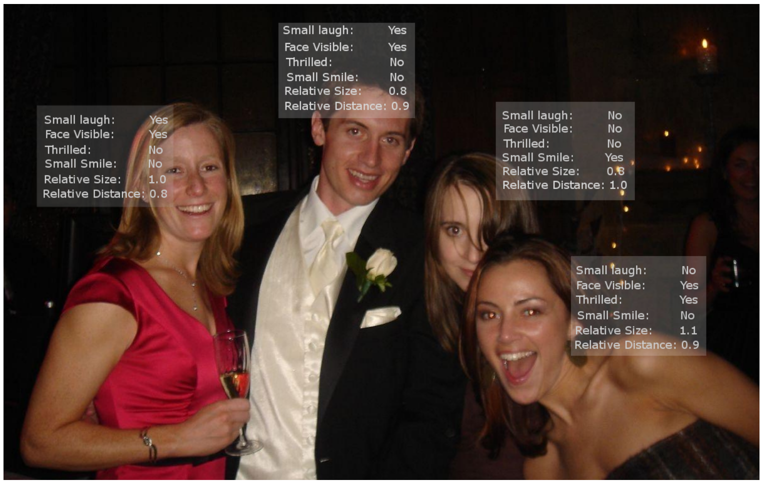


Fig. 2. The picture shows some manually defined attributes for subjects in a group.

Laugh, *Large Laugh*, *Thrilled*, and for occlusion as *Face Visible*, *Partial Occlusion* and *High Occlusion*. These attributes are computed for each face in the group. Attributes based on global context are *Relative Distance* and *Relative Size*. Figure 2 describes the manual attributes for faces in a group.

Defining attributes manually is a subjective task, which can result in many important discriminative attributes being ignored. Inspired by Tsai *et al.* [17], we extract low-level feature based attributes. They propose the use of manually defined attributes along with data-driven attributes. Their experiments show a leap in performance for human action recognition based on combination of manual and data-driven attributes. We compute weighted bag of visual words based on extracting low level features. Furthermore, a topic models is learnt using Latent Dirichlet allocation (LDA) [18]. We join the manually defined and weighted data-driven attributes.

3.1 GEM_{LDA}

Topic models, though originally developed for document analysis domain, have found a lot of attention in computer vision problems. One very popular topic modelling technique is Blei *et al.*'s Latent Dirichlet Allocation (LDA) [18], a hierarchical Bayesian model, where topic proportions for a document are drawn from a Dirichlet distribution and words in the document are repeatedly sampled from a topic, which itself is drawn from those topic proportions.

Weighted Soft Assignment: K-means is applied to the image features for defining the visual words. For creating a histogram, each word of a document is assigned to one or more visual words. If the assignment is limited to one word, it is called hard assignment and if multiple words are considered, it is called soft assignment. The cons of hard assignment are that if a patch (face in a group $\mathcal{G}$) in an image is similar to more than one visual word, the multiple association information is lost. Therefore, Jiang *et al.* [19] defined a soft-weighting assignment to weight the significance of each visual word towards a patch. For a visual vocabulary of K visual words, a K-dimensional vector $T = [t_1...t_K]$ with each component t_k representing the weight of a visual word k in an group $\mathcal{G}$ is defined as

$$\mathbf{t_k} = \sum_{i}^{N} \sum_{j}^{M} \frac{1}{2^{i-1}} sim(j,k), \tag{10}$$

where M_i represents the number of face f_j whose ith nearest neighbour is the visual word k. The measure $sim(j,k)$ represents the similarity between face f_j and the visual word k. It is worth noting that the contribution of each word is dependent to its similarity to a visual word weighted by the factor $\frac{1}{2^{i-1}}$.

Relatively weighted soft-assignment: Along with the contribution of each visual word to a group G, we are also interested in adding the global attributes as weights here. As our final goal is to understand the contribution of each face f_i towards the happiness intensity of its group $\mathcal{G}$, we use the relative weight formulated in Eq. 8 to define a 'relatively weighted' soft-assignment. Eq. 10 can then be modified as

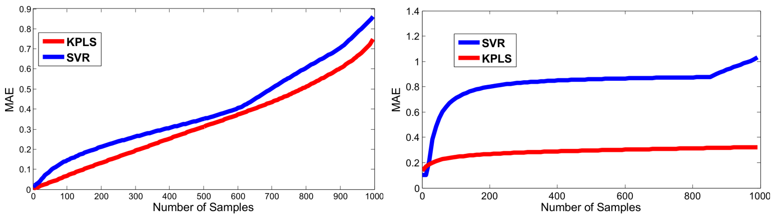
$$\mathbf{t_k} = \sum_{i}^{N} \sum_{j}^{M} \frac{\psi_j}{2^{i-1}} sim(j,k) \qquad . \tag{11}$$

Now, along with weights for each nearest visual word for a patch, another weight term is being induced, which represents the contribution of the patch to the group. These data-driven visual words are appended with the manual attributes. Note that the histogram computed here is influenced by the global attributes of the faces in the group.

The default LDA formulation is an unsupervised Bayesian method. In their recent work, Blei *et al.* [20] proposed the Supervised LDA (SLDA) by adding a response variable for each document. It was shown to perform better for regression and classification task. Since, we have the human annotated labels for the happiness intensities at the image level, using a supervised topic model is a natural choice. The document corpus is the set of groups $\mathcal{G}$. The word here represents each face in $\mathcal{G}$. We apply the Max Entropy Discriminant LDA (MedLDA) [21] for topic model creation and test label inference. We call our LDA formulations for groups GEM_{LDA}. In the results, section we compare GEM, GEM_w and GEM_{LDA}.

4 Images and Labelling

In the past, various databases relevant to our work have been proposed [2, 22]. The GENKI database [2] is an uncontrolled database containing single subject



(a) Happiness Intensity (b) Occlusion Intensity

Fig. 3. The two graphs describe the MAE. a) Happiness intensity comparison. b) Occlusion intensity comparison.

per image smiling and non-smiling. However, it does not suit our purpose as the intensity levels of happiness are not labelled at both image and face level. The Acted Facial Expressions In The Wild database [22] is a dynamic facial expressions database collected from movies. It contains both single and multiple subjects videos. However, there are no intensity labels present.

Web based photo sharing websites such as Flickr and Facebook contain billions of images. From a research perspective, not only are these a huge repository of images but also include rich associated labels, which contain very useful information describing the scene in them. We collected an 'in the wild' database from Flickr. The database contains 4600 images. We used an automatic program to search and download images, which had keywords associated with group of people and events. A total of 40 keywords were used (e.g. 'party+people', 'group+photo', 'graduation+ceremony'). After downloading the images, a Viola-Jones (VJ) object detector trained on different data (the frontal and pose models in OpenCV) was applied to the images and only the images, which contained more than one subject, were kept. Images with false detections were manually removed. We call this labelled image collection: HAPpy PEople Images (HAPPEI).

The labellers annotated all the images with the overall happiness intensity. Also, in 4886 images, 8500 faces were annotated for face level happiness intensity, occlusion intensity and pose by 4 human labellers. For happiness intensity corresponding to six stages of happiness: *Neutral*, *Small Smile*, *Large Smile*, *Small Laugh*, *Large Laugh* and *Thrilled*. The LabelMe [23] based Bonn annotation tool [24] was used for labelling.

Method	GEM	GEM_w	GEM_{LDA}
MAE	0.455	0.434	0.379

Table 1. The table compares the Mean Average Error for the three group expression models: GEM, GEM_w and GEM_{LDA}.

5 Results

Face processing pipeline: Given an image, the VJ object detector [25] models trained on frontal and profile faces are computed. For extracting the fiducial points, [26]'s part based point detector is applied. This gives us nine points, which describe the location of the left and the right corners of both eyes, the centre point of the nose, left and right corners of the nostrils, and the left and right corners of the mouth. For aligning the faces, we applied an affine transform based on these points.

As the images have been collected from Flickr and contain different scenarios and complex backgrounds, classic face detectors, such as the Viola-Jones object detector, give a fairly high false positive rate (13.6%). To minimise this error, we learned a non-linear binary Support Vector Machine (SVM) [27]. The training set contains samples containing faces and non-faces. For face samples, we manually selected all the true positives from the output of VJ detector executed on 1300 images from our database. For non-faces, we made the training set as follows: we manually selected the false positives from the same VJ output on our database. To create a large number of false positives from real world data, we collected an image set containing monuments, mountains and water scenes but no persons facing the camera. To learn the parameters for SVM, we used five-fold cross validation.

Implementation Details: Given a test image I containing group $\mathcal{G}$, the faces in the group are detected and aligned. The faces are cropped to 70×70 size. For happiness intensity detection, PHOG features are extracted from the face. Here, pyramid level $L = 3$, angle range $= [0-360]$ and bin count $= 16$. The number of latent variables were chosen as 18 after empirical validation. PHOG is scale invariant. The choice of using PHOG is motivated from our earlier work [28], where PHOG performed well for facial expression analysis.

The parameters for MedLDA were $\alpha = 0.1$, $k = 25$, for SVM $fold = 5$. 1500 documents were used for training and 500 for testing. The range of label is the group happiness intensity range [0-100] with a step size of 10. For learning the dictionary, k the number of words was empirically set to 60. In Eq. 4 and Eq. 7, the parameters were set as follows: $\alpha = 0.3$, $\beta = 1.1$ and $\gamma = 0.1$. We performed both quantitative and qualitative experiments. 2000 faces were used for training and 1000 for testing of the happiness and occlusion intensity regression models.

Human label comparison: We compare the Mean Average Error (MAE) for GEM, GEM_w and GEM_{LDA}. We compare the performance of our occlusion intensity and happiness intensity estimators, which are based on KPLS with a Support Vector Regression [27] based occlusion intensity detector. Figure 3 displays the comparison based on the MAE scores. The MAE for occlusion intensity are 0.79 for KPLS and 1.03 for SVR. The MAE for happiness intensity estimation for KPLS is 0.798 and for SVR 0.965. Table 1 shows the MAE comparison of GEM, GEM_w and GEM_{LDA}. As hypothesised, the effect of adding social features is evident with the lower MAE in GEM_w and GEM_{LDA}.

User Study We performed a two-part users survey. A total of 15 subjects were asked to a) give happiness intensities to 40 images and b) rate the output of

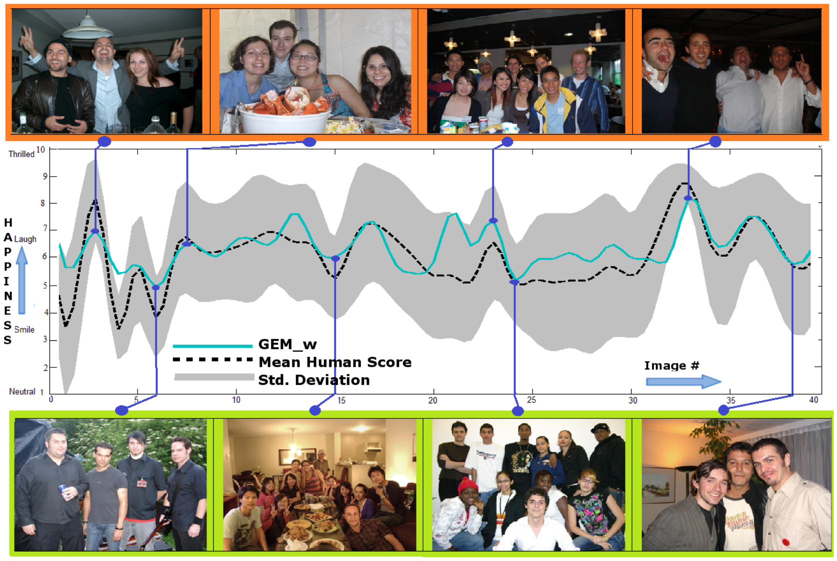


Fig. 4. The graph describes the comparison of the group happiness intensity as calculated by our methods with the results from the user study. The top row shows images with high intensity score and the lower row shows images which are close to neutral. Please note that the images are from different events.

the three methods for their output of the top-5 happiest images from an event. Here, the users were asked to rate a score from the range 0 (not good at all) to 5 (good) for the three methods for three social events each. They did not know which output belonged to which method. For part a), Figure 4 shows the output. Note that the happiness scores computed by the GEM_w are close to the mean human score and are well within the range of the standard deviation of the human labellers' scores. For part b), we performed ANOVA tests with the hypothesis that adding social context to group expression analysis leads to an estimate closer to human perception. For GEM and GEM_w, $p < 0.0006$, which is statistically significant in the one-way ANOVA. For GEM and GEM_{LDA}, $p < 0.0002$, which is also statistically significant.

Image ranking from an event: For comparison of the proposed framework, volunteers were asked to rank a set of images containing a group of people from an event. Now the task is as follows: Given a social event with different or the same people present in the same or different photographs, we wish to find the happiest moment of the event. Therefore, we rank all the images by their decreasing amount of happiness intensity. Figure 5 is a snapshot for an event ranking experiment. In the first row, the images are arranged based on their timestamp, i.e. when they were shot. The second row shows the ranking

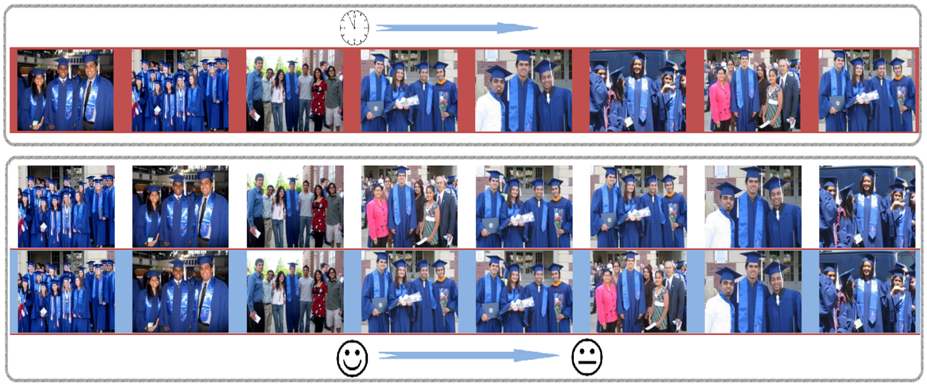


Fig. 5. This figure shows a graduation ceremony. The top row (red background) are the images from a graduation ceremony organised by timestamps. The second row (white background) are the images ranked by their decrease in intensity of happiness (from left to right) by human annotators. The third row (blue) are the images ranked by their decrease in intensity of happiness (from left to right) by our method $\mathbf{GEM}_w$. A higher resolution figure can be found in the supplementary material.

by human labellers. The highest happiness intensity image is on the left and decreases from left to right. Now, the output of the $\mathbf{GEM}_w$ is in row 3, where our method ranked the images in order of their decreasing happiness intensity.

Candid Group Shot Selection: There are situations in social gatherings when multiple photographs are taken for the same subjects in a similar scene within a short span of time. Due to the dynamic nature of groups of people, it is a challenging task to get the most favourable expression together in a group of people. Here, we experiment with our group happiness method for shot selection after a number of pictures have been taken. In Figure 6, the rows are the shots taken at short intervals. The $\mathbf{GEM}_w$ ranks the images containing the same subjects and the best image (highest happiness quotient) is displayed in the fourth column. More experiments and visual outputs can be found in the supplementary material.

6 Conclusion

Social gathering events generate many group shots. We propose a framework for estimating the theme expression of an image, focussing on happiness, containing a group of people based on social context. To the best of our knowledge, this is the first work for analysing mood of a group based on the structure of the group and local attributes such as occlusion. We collected an 'in the wild' database called HAPpy PEople Images (HAPPEI) from Flickr based on keyword search. It is

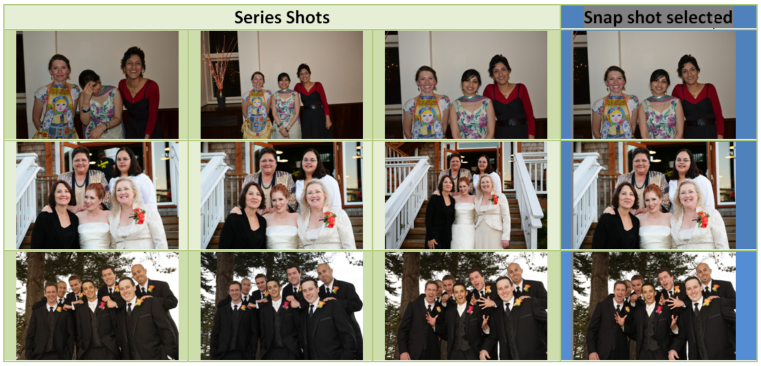


Fig. 6. Candid Group Shot Selection: Here, each row represents a series of photographs of the same people. The fourth column is the selected shot based on the highest score from $\mathbf{GEM}_w$ (Eq. 9). Please refer to the supplementary material for more experiments.

labelled at both image and face level. From the perspective of social context, we deal with the global structure of the group. We assign relative weights to the happiness intensities of individual faces in a group, so as to estimate their contribution to the group mood. We show that assigning relative weights to intensities helps in better predication of the group mood. Further, topic model based group expressions model performs better than the average and weighted group expressions model. We hope that the proposed framework and database will be able to contribute and instigate research in group expressions analysis.

In the future, we will explore other social context factors such as age, gender and their effect on mood analysis of a group. As this work focuses on defining the framework and social constraint, we have used standard detector algorithms and descriptors. Further extensions of the work can benefit immensely from robust face detection and alignment and the use of new, faster and more discriminative descriptors. In its current form, the proposed framework has a limitation w.r.t. extreme poses. In future, we will add pose handling and develop a system, which selects group photographs that can benefit immensely from pose information. Further, computer vision problems such as early event detection and abnormal event detection for multiple subjects can be dealt with this framework.

References

1. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence (2009) 39–58
2. Whitehill, J., Littlewort, G., Fasel, I.R., Bartlett, M.S., Movellan, J.R.: Toward Practical Smile Detection. IEEE TPAMI (2009) 2106–2111

3. Gallagher, A., Chen, T.: Understanding Images of Groups Of People. In: IEEE CVPR. (2009) 256–263
4. Wang, G., Gallagher, A.C., Luo, J., Forsyth, D.A.: Seeing people in social context: Recognizing people and social relationships. In: ECCV (5). (2010) 169–182
5. Stone, Z., Zickler, T., Darell, T.: Autotagging facebook: Social network context improves photo annotation. In: IEEE CVPR. (2008)
6. Hernandez, J., Hoque, E.: MIT Mood Meter (2011) http://www.moodmeter.media.mit.edu.
7. Fiss, J., Agarwala, A., Curless, B.: Candid portrait selection from video. ACM Trans. Graph. (2011) 128
8. Eichner, M., Ferrari, V.: We Are Family: Joint Pose Estimation of Multiple Persons. In: ECCV. (2010) 228–242
9. Lin, D., Tang, X.: Quality-Driven Face Occlusion Detection and Recovery. In: IEEE CVPR. (2007)
10. Rosipal, R.: Nonlinear Partial Least Squares: An Overview. In: Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques. ACCM, IGI Global (2011) 169–189
11. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: IEEE CVPR. (2011) 657–664
12. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: IEEE ICCV. (2009) 24–31
13. Schwartz, W.R., Guo, H., Davis, L.S.: A robust and scalable approach to face identification. In: ECCV. (2010) 476–489
14. Bosch, A., Zisserman, A., Munoz, X.: Representing Shape with a Spatial Pyramid Kernel. In: CIVR. (2007) 401–408
15. Gehrig, T., Ekenel, H.K.: Facial action unit detection using kernel partial least squares. In: ICCV Workshops. (2011) 2092–2099
16. Parikh, D., Grauman, K.: Relative attributes. In: ICCV. (2011) 503–510
17. Tsai, G., Xu, C., Liu, J., Kuipers, B.: Real-time indoor scene understanding using bayesian filtering with motion cues. In: ICCV. (2011) 121–128
18. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: NIPS. (2001) 601–608
19. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: CIVR. (2007) 494–501
20. Blei, D.M., McAuliffe, J.D.: Supervised Topic Models. In: NIPS. (2007)
21. Zhu, J., Ahmed, A., Xing, E.P.: Medlda: maximum margin supervised topic models for regression and classification. In: ICML. (2009) 158
22. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE MultiMedia **19** (2012) 34–41
23. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. IJCV (2008) 157–173
24. Korc, F., Schneider, D.: Annotation tool. Technical Report TR-IGG-P-2007-01, University of Bonn, Department of Photogrammetry (2007)
25. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: IEEE CVPR. (2001) 511–518
26. Everingham, M., Sivic, J., Zisserman, A.: Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In: BMVC. (2006) 899–908
27. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001) http://www.csie.ntu.edu.tw/ cjlin/libsvm.
28. Dhall, A., Asthana, A., Goecke, R., Gedeon, T.: Emotion recognition using PHOG and LPQ features. In: FG, FERA workshop. (2011) 878–883