Automatic Group Happiness Intensity Analysis

Article *in* IEEE Transactions on Affective Computing · January 2015

DOI: 10.1109/TAFFC.2015.2397456

CITATIONS
84

READS
768

3 authors:

Abhinav Dhall
Indian Institute of Technology Ropar
97 PUBLICATIONS **2,711** CITATIONS
SEE PROFILE

Roland Goecke
University of Canberra
177 PUBLICATIONS **4,539** CITATIONS
SEE PROFILE

Tom Gedeon
Australian National University
433 PUBLICATIONS **6,055** CITATIONS
SEE PROFILE

Automatic Group Happiness Intensity Analysis

Abhinav Dhall, *Member, IEEE,* Roland Goecke, *Member, IEEE,* and Tom Gedeon, *Senior Member, IEEE*

Abstract—The recent advancement of social media has given users a platform to socially engage and interact with a larger population. Millions of images and videos are being uploaded everyday by users on the web from different events and social gatherings. There is an increasing interest in designing systems capable of understanding human manifestations of emotional attributes and affective displays. As images and videos from social events generally contain multiple subjects, it is an essential step to study these groups of people. In this paper, we study the problem of happiness intensity analysis of a group of people in an image using facial expression analysis. A user perception study is conducted to understand various attributes, which affect a person's perception of the happiness intensity of a group. We identify the challenges in developing an automatic mood analysis system and propose three models based on the attributes in the study. An 'in the wild' image-based database is collected. To validate the methods, both quantitative and qualitative experiments are performed and applied to the problem of shot selection, event summarisation and album creation. The experiments show that the global and local attributes defined in the paper provide useful information for theme expression analysis, with results close to human perception results.

Index Terms—Facial expression recognition, group mood, unconstrained conditions.

1 Introduction

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 a group of people in an image. With the growing popularity of data sharing and broadcasting websites such as YouTube and Flickr, every day users are uploading millions of images and videos of social events such as a party, wedding or a graduation ceremony. Generally, these videos and images were recorded in different conditions and may contain one or more subjects. From a view of automatic emotion analysis, these diverse scenarios have received less attention in the affective computing community.

Consider an illustrative example of inferring the mood of a group of people posing for a group photograph at a school reunion. To scale the current emotion detection algorithms to work on this type of data in the wild, there are several challenges to overcome such as emotion modelling of groups of people, labelled data, and face analysis. Expression analysis has been a long studied problem, focussing on inferring the emotional state of a single subject only. This paper discusses the problem of automatic mood analysis of a group of people. 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, 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 groups of people in images is an unexplored problem that has many real-world applications: image search, retrieval, representation and browsing; event summarisation and highlight creation; candid photo shot selection; expression apex detection in video; video thumbnail creation etc. A recent Forbes magazine article [1] discusses the lack of ability of current image search engines to use context. Information such as the mood of a group can be used to model the context. These problems, where group mood information can be utilised, are a motivation for exploring the various group mood models. One basic approach is to average the happiness intensities of all people in a group. However, the perception of the mood of a group is defined by attributes such as where people stand, how much of their face is visible etc. These social attributes play an important role in defining the overall happiness[1] an image conveys.

2 Key Contributions

1) An automatic framework for happiness intensity analysis of a group of people in images based on the social context.
2) A weighted model is presented, taking into consideration the global and local attributes that affect the perceived happiness intensity of a group.
3) A labelled 'in the wild' database containing images of groups of people is collected using a semi-automatic process and compared with existing databases.

The remainder of the paper is organised as follows: Section 3 discusses prior work on various aspects of the visual analysis of a group of people. Section 4 describes

• *A. Dhall and R. Goecke are with the University of Canberra. T. Gedeon is with the Australian National University E-mail: {abhinav.dhall, tom.gedeon}@anu.edu.au, roland.goecke@ieee.org*

1. This paper uses the terms 'mood', 'expression' and 'happiness' interchangeably for describing the mood of a group of people in an image.

the problems and challenges involved in automatic group expression analysis. The details of a 149-user survey investigating attributes affecting the perception of mood are discussed in Section 5. An 'in the wild' database collection method is detailed in Section 6. Section 7 discusses a basic model based on averaging. Global context based social features are presented in Section 8. Local context based on occlusion intensity is described in Section 9. The global and local contexts are combined and applied to the averaging approach in Section 10. The manual attributes are combined with data-driven attributes in a supervised hierarchal Bayesian framework in Section 11. Section 12 discusses the results of the proposed frameworks, including both quantitative and qualitative experiments.

3 Literature Review

The analysis of a group of people in an image or a video has recently received much attention in computer vision. 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 [2], [3], [4]; b) **Top-down** methods: The group/sub-group information is used as a prior for inference of subject level attributes [5], [6].

3.1 Bottom-up Techniques

Tracking groups of people in a crowd has been of particular interest lately [2]. Based on trajectories constructed from the movement of people, [2] propose a hierarchical clustering algorithm which detects sub-groups in crowd video clips. In an interesting experiment, [3] installed cameras at four locations on the MIT campus and tried to estimate the mood of people looking into the camera and compute a mood map for the campus using the Shore framework [7] for face analysis, which detects multiple faces in a scene in real-time. The framework also generates attributes such as age, gender and pose. In [3], the scene level happiness averages the individual persons' smiles. However, in reality, group emotion is not an averaging model [8], [9]. There are attributes, which affect the perception of a group's emotion and the emotion of the group itself. The literature in social psychology suggests that group emotion can be conceptualised in different ways and is best represented by pairing the top-down and bottom-up approaches [8], [9].

In another interesting bottom-up method, [10] proposed group classification for recognising urban tribes (a group of people part of a common activity). Low-level features, such as colour histograms, and high-level features, such as age, gender, hair and hat, were used as attributes (using the Face.com API) to learn a Bag-of-Words (BoW)-based classifier. To add the group context, a histogram describing the distance between two faces and the number of overlapping bounding boxes was computed. Fourteen classes depicting various groups, such as 'informal club', 'beach party' and 'hipsters', were used. The experiments showed that a combination of attributes can be used to describe a type of group. In 'Hippster wars' [11], a framework based on clothes related features was proposed for classifying a group of people based on their social group type.

3.2 Top-down Techniques

In an interesting top-down approach, [5] proposed contextual features based on the group structure for computing the age and gender of individuals. The global attributes described here are similar to [5]'s contextual features of social context. However, the problem in [5] is inverse to the problem of inferring the mood of a group of people in an image, which is discussed in this paper. Their experiments on images obtained from the web, show an impressive increase in performance when the group context is used. In another top-down approach, [6] 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. A graphical model based on social relationships, such as '*father-child*' and '*mother-child*', and social relationship features, such as relative height, height difference and face ratio. In [12] a face discovery method based on exploring social features, such as on social event images, is proposed.

In object detection and recognition work by [13], scene context information and its relationship with the objects is described. Moreover, [14] acknowledges the benefit of using global spatial constraints for scene analysis. In face recognition [15], social context is employed to model the relationship between people, e.g. between friends on Facebook, using a Conditional Random Field (CRF) [16].

Recently, [17] proposed a framework for selecting candid shots from a video of a single person. A physiological study was conducted, where 150 subjects were shown images of a person. They were asked to rate the attractiveness of the images and mention attributes, which influenced their decision. Professional photographers were also asked to label the images. Further, a regression model was learnt based on various attributes, such as eye blink, clarity of face and face pose. A limitation of this approach is that the samples contain a single subject only.

[18] proposed affect based video clip browsing by learning two regression models, predicting valence and arousal values, to describe the affect. The regression models learnt on an ensemble of audio-video features, such as motion, shot switch rate, frame brightness, pitch, bandwidth, roll off, and spectral flux. However, expression information for individuals or groups in the scenes was not used.

The literature for analysing a single subject's happiness / smile is rich. One prominent approach by [19] proposed a new image-based database labelled for smiling and non-smiling images and evaluated several state of the art methods for smile detection. However, in the existing literature, the faces are considered independent of each other. For computing the contribution of each subject, two types of factors affect group level emotion analysis: (1) Local factors (individual subject level): age, gender, face visibility, face pose, eye blink etc. (2) Global factors: where do people stand, with whom people stand etc. In this paper, the focus is on face visibility, smile intensity, relative face size and relative face distance. Labelled images containing groups of people are required, which we collect from Flickr.

4 CHALLENGES

The following subsections discuss the challenges in creating an automatic system for happiness intensity inference.

4.1 Attributes

Human perception of the mood of a group of people is very subjective. [9] argue that the mood of a group is composed by two broad categories of components: top-down and bottom-up. Top-down is the affective context, i.e. attributes such as group history, background, social event etc., which have an effect on the group members. For example, a group of people laughing at a party displays happiness in a different way than a group of people in an office meeting room. From an image perspective, this means that the scene/background information can be used as affective context. The bottom-up component deals with the subjects in the group in terms of attributes of individuals that affect the perception of the group's mood. It defines the contribution of individuals to the overall group mood.

From now on, the top-down component is referred to as 'global context' and the bottom-up component as 'local context'. There can be various attributes, which define these two components. For example, global context contains but is not limited to scene information, social event information, who is standing with whom, where are people standing in an image and w.r.t. the camera. Local context, i.e. subject specific attributes, cover an individual's mood/emotion, face visibility, face size w.r.t. neighbours, age, gender, head pose and eye blink. To further understand these attributes, a perception user study is performed. The study and its results are detailed in Section 5.

4.2 Data and Labelling

Data simulating 'in the wild' conditions is a major challenge for making emotion recognition methods work in real-world conditions. Generally, emotion analysis databases are lab-recorded and contain a single subject in an image or video. It is easy to ask people to pose in a laboratory, but acted expressions are very different from spontaneous ones. Anyone working in emotion analysis will attest to the difficulty of collecting spontaneous data in real-world conditions. For learning and testing an emotion analysis system, labelled data containing groups of people in different social scenarios is required. Once the data is available, the next task is labelling. According to [20], moods are low-intensity, diffuse affective states that usually do not have a clear antecedent. Mood can be positive/pleasant and negative/unpleasant. The type of labelling (discrete or continuous) required is problem dependent. The database proposed in this paper – HAPPEI – is labelled for neutral to pleasant mood with discrete levels.

4.3 Visual Analysis

Inferring the group mood involves classic computer vision tasks. As a pre-processing step, the first challenge is face and facial landmark detection. Ideally, for a facial dynamics analysis system, one will want a subject independent facial landmark detector [21]. Further, [21] argue that a subject dependent facial parts method, such as Active Appearance Models (AAM) [22], performs better than subject independent Constrained Local Models (CLM) [23]. However, if a proper descriptor is used on top of previously aligned faces from a subject independent detector, the aligment error can be compensated for. Moreover, [24] show the effectiveness of the Mixture of Pictorial Structures [24] over CLM and AAM for facial landmark localisation when there is much head movement. Motivated by these arguments, the parts based model of [25][2] is used here (Section 12). The images in the new database HAPPEI were downloaded from Flickr, creating the challenge of real-world varied illumination scenarios and differences in face resolution. To overcome these, LPQ and PHOG are used, as LPQ is robust to varied illumination and PHOG is robust to scale [26].

5 SURVEY

To understand the attributes (Section 4.1) affecting the perception of the group mood, a user study was conducted. Two sets of surveys were developed. In the first part (Figure 1), subjects were asked to compare two images for their apparent mood and rate the one with a higher positive mood. Further, they were asked various questions about the attibutes/reasons, which made them choose a specific image/group out of the two images/groups. A total of 149 subjects participated in this survey (94 males / 55 females). There are a total of three cases in the first survey. Figure 1 shows the questions in one of the cases. Cases 1, 2 and 3 in Figure 2 describe the analysis of the responses of the participants for the three cases in the survey. On the left of the figures, the two images to be compared are displayed.

The images in the survey were chosen on the basis of two criteria: (1) to validate the hypothesis that adding an occlusion attribute to the model decreased the error (user perception vs. model output), which was noticed in the earlier experiments in [27]. Therefore, in one case, two images shot in succession were chosen, in which one of the subjects covered his face in the first shot (Figure 2 Case 1). It is interesting to note that a larger number of survey participants (69.0%) chose image B in Figure 2 Case 1 as having a more positive mood score on the scale of neutral towards thrilled. Out of these 69.0%, 51.1% chose 'faces being less occluded' as one of the reasons. The other dominating attribute for their decision was the larger number of people smiling (54.6%). Both attributes are correlated; it is easier to infer the expression of a person when the face is clearly visible.

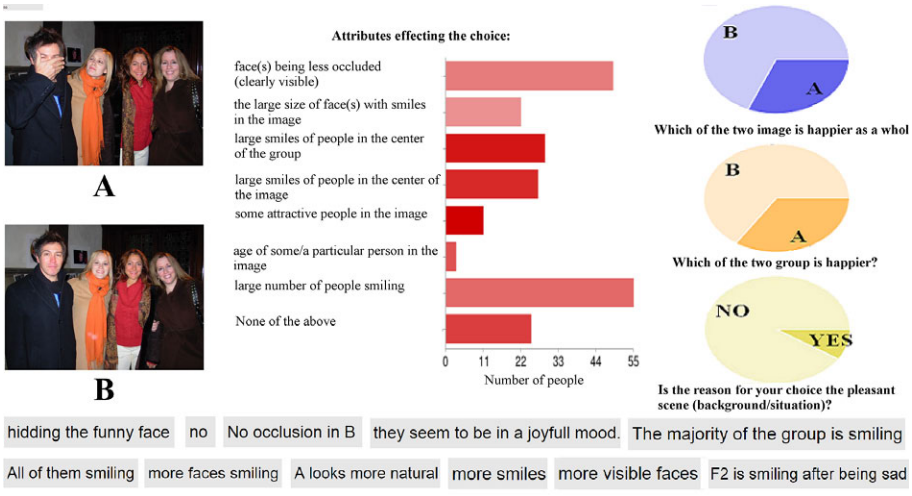
(2) In the other case (Figure 2 Case 2), two images were randomly chosen. 76.0% of participants ranked Image A higher than Image B. 46.0% of participants chose 'larger number of people smiling' as the most dominant attribute (49.1% for participants who selected Image A and 37.1% for participants who selected Image B). 46.4% of

2. [24] report that their method works better than [25]; however, due to its lower computational complexity [25] is used here.

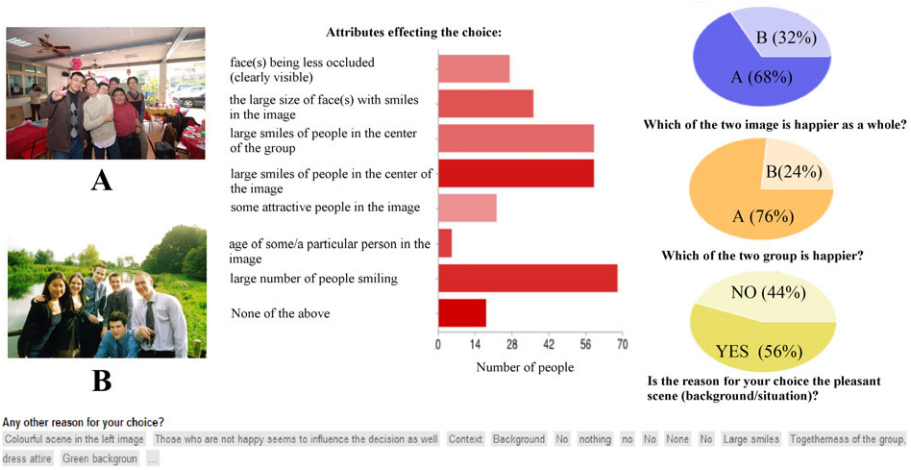
Fig. 1. Screenshot of a case in the user survey (Section 5).

participants who chose Image A selected 'large smiles of people in the center of the group'. Also, 56.0% of the participants chose the presence of a pleasant background as the reason for their selection. In the question: 'Any other reason for your choice?', participants pointed to the context, background, togetherness of the group, and party like scenario in Image A (Figure 2 Case 2), which made them consider the group in Image A being happier. Other considerable responses were body pose, people closer to each other in the group and spontaneous facial expressions (i.e. when subjects are not explicitly posing in front of the camera). For the question: 'Please define the scene in one word in Image A' and 'Please define the scene in one word in Image B', the majority of the participants defined the scene in the context of mood such as 'relaxed', 'pleasant', 'interesting', 'happening', or 'enjoyable'.

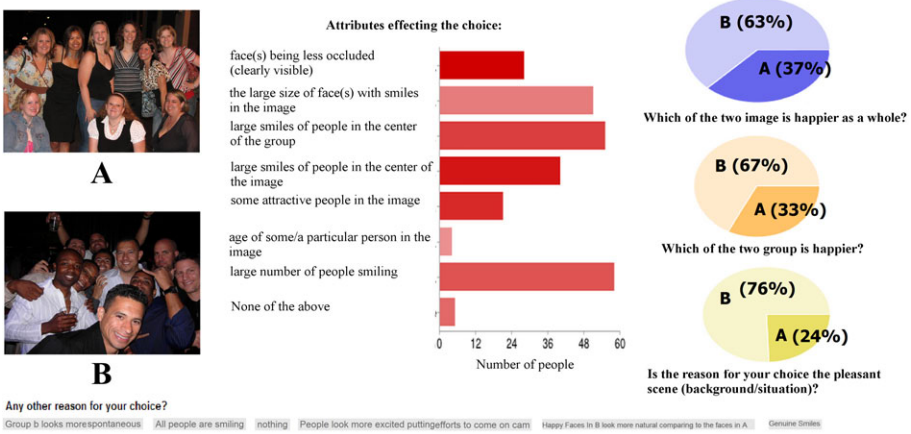
In Figure 2 Case 3, 41.0% of the survey participants chose 'the large size of face(s) with smiles in the image'. For the question 'What are the dominating attributes/characteristics of the leader(s) in the group that affect the group's mood?', the participants mentioned the large size of face(s) with smiles in the image, the centre location of the subject, large smiles, and people standing closer. Based on the user survey, in the following sections, the attributes which are discussed further are relative location of members of a group, relative face size to estimate if a person is in the front or back, face visibility/occlusion and mood of a member. In a recent study by Jiang et al. [28], the authors conducted an eye tracking based study to locate the objects and attributes, which are salient in group/crowd images. Their saliency related attributes are similar to our findings. Based on observations from the fixations, [28] proposed high-level features related to attributes such as face size, face occlusion and pose.



(a) Case 1



(b) Case 2



(c) Case 3

Fig. 2. Results of the analysis of some cases in the survey (Section 5).

6 Database Creation and Labelling

Popular facial expression databases, such as CK+ [29], FEEDTUM [30] and MultiPIE [31], are all 'individual' centric databases, i.e. contain a single subject only in any particular image or video. For the problem in this work, there are various databases, which are partially relevant [19], [4]. In an interesting work, [19] proposed the GENKI database. It was collected from Google images shot in unconstrained conditions containing a single subject per image smiling or non-smiling. However, it does not fulfill our requirements as the intensity level of happiness is not labelled at both image and face level. [4] propose a dynamic facial expressions database – *Acted Facial Expressions In The Wild* (AFEW) – collected from movies. It contains both single and multiple subject videos. However, there are no intensity labels present and multiple subject video clips are few in number. Therefore, a new labelled database for image-based group mood analysis is required.

Databases such as GENKI and AFEW have been compiled using semi-automatic approaches. [19] used Google images based on a keyword search for finding relevant images. [4] used a recommender system based approach where a system suggested video clips to the labellers based on emotion related keywords in closed caption subtitles. This makes the process of database creation and labelling easier and less time consuming. Inspired by [4], [19], a semi-automatic approach is followed. 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 come with rich associated labels, which contain very useful information describing the scene in the images.

We collected a labelled 'in the wild' database – called HAPpy PEople Images (HAPPEI) – from Flickr containing 4886 images. A Matlab based program was developed to automatically search and download images, which had keywords associated with groups of people and events. A total of 40 keywords were used (e.g. 'party + people', 'group + photo', 'graduation + ceremony', 'marriage', 'bar', 'reunion', 'function', 'convocation'). After downloading the images, a Viola-Jones object detector trained on different data (frontal and pose models in OpenCV) was executed on the images. Only images containing more than one subject were kept. False detections were manually removed. Figure 3(a) shows a collage of images from the database.

All images were annotated with a group level mood intensity ('neutral' to 'thrilled'). Moreover, in the 4886 images, 8500 faces were manually annotated for face level happiness intensity, occlusion intensity and pose by four human labelers, who annotated different images. The mood was represented by the happiness intensity corresponding to six stages of happiness: *Neutral*, *Small Smile*, *Large Smile*, *Small Laugh*, *Large Laugh* and *Thrilled* (Figure 3(b)). As a reference during labelling, when the teeth of a member of a group were visible, the happiness intensity was labelled as a laugh. When the mouth was open wide, a *Thrilled* label was assigned. A face with a closed mouth was assigned the label *Smile*. The LabelMe [32] based Bonn annotation tool [33] was used for labelling. It is interesting to note that ideally one would like to infer the mood of a group by the means of self-rating along with the perception of mood of the group. In this work, no self-rating was conducted as the data was collected from the internet. In this database, the labels are based on the perception of the labellers. One can see this work as a stepping stone to group mood analysis. The aim of the models (Sections 7, 10 and 11.1) proposed in this work is to infer the perceived group mood as closely as possible to human observers.

(a) A collage of sample images in HAPPEI.

(b) Sample face level happiness intensity labels in HAPPEI.

Fig. 3. The HAPPEI database.

7 Group Expression Model

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

$$\text{GEM} = \frac{\sum_i \mathcal{I}_{\mathcal{H}i}}{s} \quad . \tag{1}$$

In this simple formulation, both global information, e.g. the relative position of people in the image, and local information, e.g. the level of occlusion of a face, are ignored. In order to add the bottom-up and top-down components (Section 4.1), it is proposed here to add these social context features as weights to the process of determining the happiness intensity of a group image. The experiments (Section 12) on HAPPEI confirm the positive effect of adding social feature weights to GEM. In the next section, methods for computing the global context are discussed.

8 Global Context

Barsade and Gibson [8] as well as Kelley and Barsade [9] emphasise the contribution of the top-down component

Fig. 4. *Top Left:* Image with mood intensity score = 70. *Top Right:* Min-span tree depicting connection between faces. *Bottom Left:* Happiness intensity heat map generated using the GEM model (mood intensity score = 81). *Bottom Right:* Happiness intensity heat map with social context. The contribution of the faces with respect to their neighbours (F2 and F4) towards the overall intensity of the group is weighted (mood intensity score = 71.4). Adding the global context feature reduces the error.

to the perception of the happiness intensity of a group. Here, the top-down contribution represents the effect of the group on a subject. Furthermore, in the survey (Section 5), participants mentioned attributes such as location and face size of subjects in a group. In this section, we formulate the weights for describing the top-down component. The tip of the nose $\mathbf{p_i}$ is considered as the position 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 $\mathbf{p_m}$. Prim's minimal spanning tree algorithm [34] is computed on $\mathcal{G}$, which provides information about the relative position of people in the group with respect to their neighbours. In Figure 4, the min-span tree of the group graph is shown.

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 (intraocular distance), $d_i = ||\mathbf{l} - \mathbf{r}||$. The relative face size $\theta_\mathbf{i}$ of f_i 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, it is assumed that the expression intensity of their faces contributes more to the perceived group mood intensity than the faces of people standing in the back. Eichner and Ferrari [35] 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

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

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

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

where parameters α and β control the effect of these weight factors on the global weight. Figure 5 demonstrates the effect of the global context on the overall output of GEM.

9 Local Context

In the previous section, global context features, which compute weights on the basis of two factors: (1) where do people stand in a group and (2) how far are they away from the camera, are defined. The bottom-up component of the framework is now defined in this section. The local

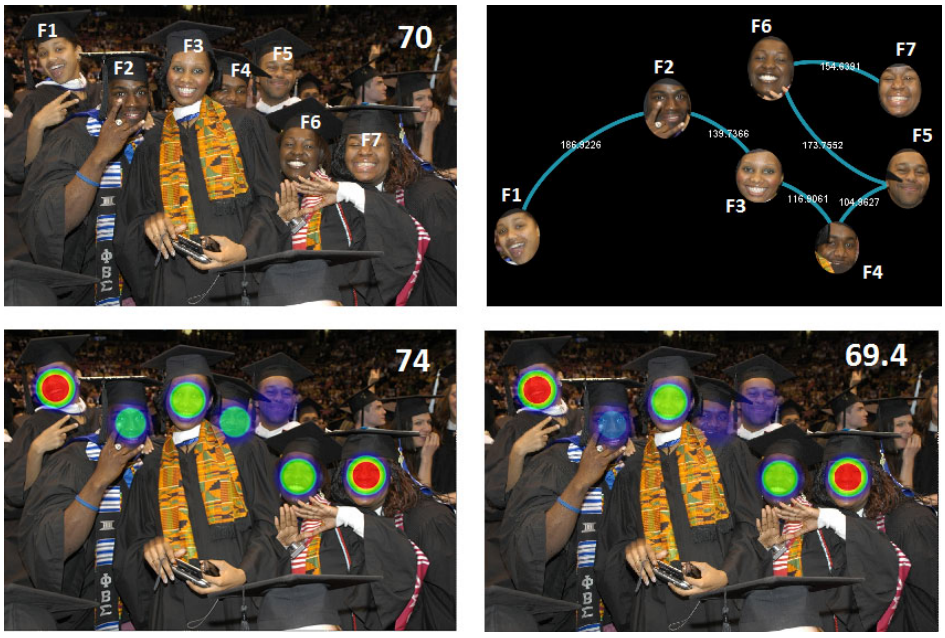


Fig. 5. *Top Left:* Image with happiness intensity score=70. *Top Right:* Min-span tree showing connection between faces. *Bottom Left:* Happiness intensity heat map. *Bottom Right:* Happiness intensity heat map with social context, the contribution of the occluded faces (F2 and F4) towards the overall intensity of the group is penalised.

context is described in terms of an individual person's level of face visibility and happiness intensity.

Occlusion Intensity Estimate: Occlusion in faces, 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 [36] introduced an automatic occlusion detection and rectification method for faces via GraphCut-based detection and confidence 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. Because of this, 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 [36], it is proposed 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) [37] regression framework. The PLS set of methods has recently become very popular in computer vision [38], [39], [40]. In [39] and [40], PLS is used for dimensionality reduction as a prior step before classification. [38] use KPLS based regression for simultaneous dimensionality reduction and age estimation and show that KPLS works well for face analysis when the feature vector is high dimension. For the occlusion intensity estimation problem, 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, the PLS framework defines a decomposition of matrices $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \quad (5)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{E} \quad (6)$$

where $\mathbf{T}$ and $\mathbf{U}$ are the $n \times p$ score matrices of the p extracted latent projections. The $N \times p$ matrix $\mathbf{P}$ and $M \times p$ matrix $\mathbf{Q}$ denote the corresponding loading matrices and the $n \times N$ matrix $\mathbf{E}$ and $n \times M$ matrix $\mathbf{F}$ denote the residual matrices that account for the error made by the projection. The classic NIPALS method [41] is used to solve the optimisation criteria

$$\begin{aligned}[cov(\mathbf{t}, \mathbf{u})]^2 &= [cov(\mathbf{Xw}, \mathbf{Yc})]^2 \\ &= max_{|r|=|s|=1}[cov(\mathbf{Xr}, \mathbf{Ys})]^2 \end{aligned} \quad (7)$$

where $cov(\mathbf{t}; \mathbf{u}) = \mathbf{t}^{\mathbf{T}}\mathbf{u}/n$ is the sample covariance between the score vectors $\mathbf{t}$ and $\mathbf{u}$. The score vectors $\{\mathbf{t}_i\}_{i=1}^{p}$ are good predictors of $\mathbf{Y}$ and the inner relation between the score vectors $\mathbf{t}$ and $\mathbf{u}$ is given by $\mathbf{U} = \mathbf{TA} + \mathbf{H}$, where $\mathbf{A}$ is a $p \times p$ diagonal matrix and $\mathbf{H}$ is the residual matrix.

To perform classification, the regression matrix $\mathbf{B}$ is calculated as

$$\mathbf{B} = \mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} \quad . \quad (8)$$

For a given test sample matrix X_{test}, the estimated labels matrix $\hat{Y}$ is given by

$$\hat{\mathbf{Y}} = \mathbf{X}_{test}\mathbf{B} \quad (9)$$

For a detailed description, readers may refer to [37]. Now, for a non-linear mapping, the kernel trick can be

applied to the PLS method. $\mathbf{X}$ is substituted with $\mathbf{\Phi} = [\Phi(x_1), ..., \Phi(x_n)]^T$, which maps input data to a higher-dimensional space. The kernel matrix is then defined by the Gram matrix $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$, in which the kernel function defines each element $\mathbf{K}_{i,j} = k(x_i, x_j)$. Therefore, Equations 8 and 9 can be rewritten as

$$\hat{\mathbf{Y}} = \mathbf{K}_{test}\mathbf{R} \quad (10)$$

$$\mathbf{R} = \mathbf{U}(\mathbf{T}^T\mathbf{K}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} \quad (11)$$

where $\mathbf{K}_{test} = \mathbf{\Phi_{test}}\mathbf{\Phi}^T$ is the kernel matrix for test samples.

The input sample vector x_i is a normalised combination of Hue, Saturation and the PHOG [42] 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 6). 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

$$\lambda_\mathbf{i} = ||1 - \gamma y_i|| \quad (12)$$

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 [43].

10 Weighted Group Expression Model

The global and local contexts defined in Eq. 4 and 12 are used to formulate the relative weight for each face f_i as

$$\pi_i = \lambda_i \psi_i \quad (13)$$

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 13, the new weighted GEM is defined as

$$\text{GEM}_w = \frac{\sum_i \mathcal{I}_{\mathcal{H}i}\pi_i}{s} \quad . \quad (14)$$

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

11 Social Context as Attributes

The social features described above can also be viewed as manually defined attributes. From the survey (Section 5), it is evident that along with the social context features, there are many others such as age, attractiveness, and gender. The

Fig. 6. Manually defined attributes.

assumptions in GEM_w do not hold true for some scenarios, for example: when a baby is in the lap of the mother. In this case, the system will give a higher weight to the mother and a lower weight to the baby, assuming that the mother is in the front and the baby is in the background. In order to implicitly add the effect of other attributes, a feature augmentation approach is presented next.

Lately, attributes have been very popular in the computer vision community (e.g. [44]). Attributes are defined as high-level semantically meaningful representations. They have been, for example, successfully applied to object recognition [44], scene analysis [45] and face analysis [46].

Based on the regressed happiness intensities, the attributes are defined as *Neutral*, *Small Smile*, *Large Smile*, *Small 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 6 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 [47], low-level feature based attributes are computed. 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 a combination of manual and data-driven attributes. A weighted bag of visual words based on extracting low-level features is computed.

Furthermore, a topic model is learnt using Latent Dirichlet allocation (LDA) [48]. The manually defined and weighted data-driven attributes are combined to form a single feature.

11.1 Augmented Group Expression Model

Topic models, though originally developed for document analysis, have been successfully applied to computer vision problems. One very popular topic modelling technique is LDA [48], 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.

[46] introduced people-LDA, where topics were modelled around faces in images along with titles from news. The work of [46] has some similarities to the method proposed here but the single biggest difference is that [46] create topics around single people rather than around a group of people. The proposed group model creates topics around happiness intensitie for a group of people. For learning the topic model, a dictionary is learnt first.

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, [49] defined a weighted soft 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

$$\mathbf{t_k} = \sum_{i}^{N} \sum_{j}^{M_i} \frac{1}{2^{i-1}} sim(j,k), \tag{15}$$

where M_i represents the number of face f_j whose i^{th} 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, it is interesting to add the global attributes as weights here. It is intuitive to note that the weights affect the frequency component of words, which here represent faces in a group. Therefore, it is similar to applying weights in GEM_w and looking at the contribution based on a neighbourhood analysis of a particular subject under consideration. As the final goal is to understand the contribution of each face f_i towards the happiness intensity of its group $\mathcal{G}$, the relative weight formulated in Eq. 13 is used to define a 'relatively weighted' soft-assignment. Eq. 15 can then be modified as

$$\mathbf{t_k} = \sum_{i}^{N} \sum_{j}^{M_i} \frac{\psi_j}{2^{i-1}} sim(j,k) \quad . \tag{16}$$

Now, along with weights for each nearest visual word for a patch, another weight term is being introduced, 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, [50] proposed the Supervised LDA (SLDA) by adding a response variable for each document. It was shown to perform better for regression and classification tasks. Using a supervised topic model is a natural choice for HAPPEI, as the human annotated labels for the happiness intensities at the image level are present. The document corpus is the set of groups $\mathcal{G}$. The word here represents each face in $\mathcal{G}$. The Max Entropy Discriminant LDA (MedLDA) [51] is computed for topic model creation and test label inference. The LDA formulation for groups is referred to as GEM_{LDA}. In the results section, the average model GEM is compared with the weighted average model GEM_w and the feature augmented topic model GEM_{LDA}.

12 Experiments

12.1 Face Processing Pipeline

Given an image, Viola-Jones (VJ) object detector [52] models trained on frontal and profile faces are applied to the images. For extracting the fiducial points, the part-based point detector of [25] is applied. The resulting nine points describe the location of the left and 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, an affine transform is applied.

As the images were collected from Flickr, containing different scenarios and complex backgrounds, classic face detectors, such as the VJ object detector, result in a fairly high false positive rate (13.6%). To minimise this error, a non-linear binary SVM [53] is trained. The training set contains samples of faces and non-faces. For face samples, all true positives from the output of the VJ detector applied to 1300 images from the HAPPEI database are selected. For non-faces, the samples are manually selected from the same VJ output. To create a large number of false positives from real world data, an image set containing monuments, mountains and water scenes (but no persons facing the camera) is constructed. To learn the parameters for SVM, five-fold cross validation is performed.

12.2 Implementation Details

Given a test image I containing group $\mathcal{G}$, the faces in the group are first detected and aligned, then cropped to a size of 70×70 pixels. For the happiness intensity detection, PHOG features are extracted from the face. Here, the pyramid level $L = 3$, angle range = $[0-360]$ and bin count $= 16$. The number of latent variables is chosen as 18 after empirical validation. PHOG is scale invariant. The use of PHOG is motivated by [54], where PHOG performed well for facial expression analysis.

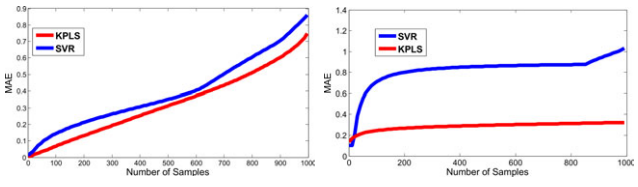
The parameters for MedLDA are $\alpha = 0.1$, $k = 25$, for SVM $fold = 5$. 1500 documents are used for training and 500 for testing. The range of labels is the group mood intensity range [0-100] with a step size of 10. For learning the dictionary, the number of words k is empirically set to 60. In Eq. 4 and 12, the parameters are set to $\alpha = 0.3$, $\beta = 1.1$ and $\gamma = 0.1$, which are weights that control the effect of

Fig. 7. The graph describes the comparison of the group mood intensity as calculated by the proposed method with the results from the user study. The top row shows images with high intensity score and the bottom row shows images which are close to neutral. Please note that the images are from different events.

manual attributes. Adding the power of 2 (to the Equations 4 and 16) results in a smooth curve based on the weight values. For a fair comparison between the three proposed models (GEM, GEM_w and GEM_{LDA}), both quantitative and qualitative experiments are performed. 2000 faces are used for training and 1000 for testing of the happiness and occlusion intensity regression models.

12.3 Human Label Comparison

The Mean Average Error (MAE) is used as performance measure. The performance of the KPLS based occlusion intensity and happiness intensity estimators is compared with Support Vector Regression (SVR) [53] based occlusion intensity and happiness intensity estimators. Figure 8 displays the comparison based on the MAE scores. The MAE for occlusion intensity is 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 in the lower MAE in GEM_w and GEM_{LDA}.



(a) Happiness Intensity (b) Occlusion Intensity

Fig. 8. Comparison of happiness and occlusion intensity methods.

Method	GEM	GEM_w	GEM_{LDA}
MAE	0.455	0.434	0.379

TABLE 1
Comparison of GEM, GEM_w and GEM_{LDA}.

12.4 User Study

A total of 15 subjects participated in a two-part user survey and were asked to a) rate happiness intensities in 40 images and b) rate the output of the three methods for their output of the top 5 happiest images from an event. Here, the users were asked to provide a score in the range of 0 (not good at all) to 5 (very good) for the three methods for three social events each. The users did not know, which output belonged to which method. For part a), Figure 7 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. The group of people in images in the top row have high happiness intensity. The groups in the lower row have a lower happiness intensity. From Figure 7, it is evident that the upper and lower bounds of the happiness intensity range assigned by the participants to the top row images are generally higher than the intensities assigned to the bottom row images. The average standard deviation of the happiness intensities is 1.67. It is interesting to note that for some images, there was a high variation as

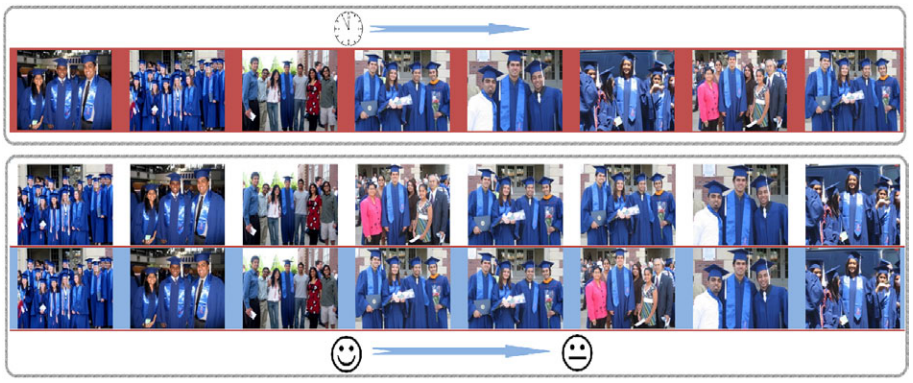


Fig. 9. The top row contains images from a graduation ceremony organised by timestamps. The second row contains images ranked by human annotators in order of decreasing happiness intensity (from left to right). The third row contains images ranked by decreasing happiness intensity (from left to right) by $\mathbf{GEM}_w$.

compared to others. This can be attributed to the difference in perception of survey participants, as for different people different objects can be more salient.

For part b), ANOVA tests were performed with the hypothesis that adding social context to group mood 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.

12.5 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 in the following task: Given a social event with the same or different people present in one or more photographs, the happiest moment of the event is to be found. Therefore, all the images are ranked on the basis of their decreasing amount of (perceived) happiness intensity. Figure 9 is a screenshot of 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 by human labellers. The highest happiness intensity image is on the left and decreases from left to right. In comparison, the output of $\mathbf{GEM}_w$ is in row 3, where the proposed method ranked the images in order of their decreasing happiness intensity.

12.6 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 find the most favourable expression together in a group of people. Here, the group mood analysis method is applied to shot selection after a number of pictures have been taken. In Figure 10, the rows show 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.

13 Conclusions

Social events generate many group shots. In this paper, a framework for estimating the group mood from an image, focussing on positive mood, is proposed. To the best of our knowledge, this is the first work for analysing group mood based on the structure of a group and local attributes such as occlusion. An 'in the wild' database called HAPPEI is collected from Flickr based on keyword search. It is labelled at both image and face level. From the perspective of social context, the global structure of the group is explored. Relative weights are assigned to the happiness intensities of individual faces in a group, so as to estimate their contribution to the perceived group mood. The experiments show that assigning relative weights to intensities helps to better predict the group mood. The feature augmented topic model based group mood analysis model performs better than the average and weighted group expressions models.

In this work, for inferring the group mood, the global social features are based on the relative location of a person. The aim is to find salient or important faces, which can be the leader in the group. An interesting direction is to compute image saliency and weight the confidence of subjects who fall in the highly salient area. A natural extension of the proposed work is adding negative emotion group images to the database and framework [55]. Further, human body pose can be merged with the face analysis of a group of people. Based on recent work by [56], body pose can convey affect information. In the images downloaded from the internet, there can be challenges such as face blur and occlusion due to neighbours in a group. This can make the inference of the mood of a person non-trivial. Body pose information can be fused with face information for robust inference. Attributes such as clothes colours and background scene details can also give important information about the social event and, hence, aid in inferring the mood of a group. The mood value of the group can be fused with other attributes such as the one mentioned in the Kansei image retrieval systems [57]. In the future, social context factors, such as age and gender, will be explored.

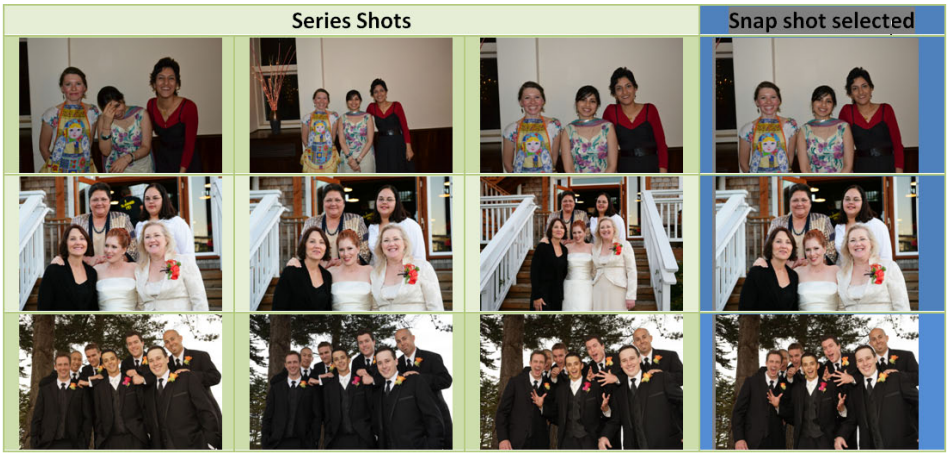


Fig. 10. Candid Group Shot Selection: Each row represents a series of photographs of the same people. The fourth column is the selected shot based on the highest score from $\mathbf{GEM}_w$ (Eq. 14).

REFERENCES

[1] M. Caroll, "How tumblr and pinterest are fueling the image intelligence problem," *Forbes*, Jaunary 2012. 1

[2] 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. 2

[3] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard, "Mood meter: counting smiles in the wild," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 301–310. 2

[4] 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. 2, 6

[5] 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)*, 2009, pp. 256–263. 2

[6] 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)*, 2010, pp. 169–182. 2

[7] C. Küblbeck and A. Ernst, "Face detection and tracking in video sequences using the modifiedcensus transformation," *Image Vision Computing*, vol. 24, no. 6, pp. 564–572, 2006. 2

[8] 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, 6

[9] 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. 2, 3, 6

[10] 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)*, 2012, pp. 28–35. 2

[11] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 472–488. 2

[12] Y. J. Lee and K. Grauman, "Face discovery with social context," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2011, pp. 1–11. 2

[13] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 763–770. 2

[14] D. Parikh, C. L. Zitnick, and T. Chen, "From appearance to context-based recognition: Dense labeling n small images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8. 2

[15] 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)*, 2008, pp. 1–8. 2

[16] O. K. Manyam, N. Kumar, P. N. Belhumeur, and D. J. Kriegman, "Two faces are better than one: Face recognition in group photographs," in *Proceedings of the International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–8. 2

[17] J. Fiss, A. Agarwala, and B. Curless, "Candid portrait selection from video," *ACM Transaction on Graphics*, p. 128, 2011. 2

[18] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 1853–1856. 2

[19] 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. 2, 6

[20] J. P. Forgas, "Affect in social judgments and decisions: A multiprocess model," *Advances in experimental social psychology*, vol. 25, pp. 227–275, 1992. 3

[21] 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. 3

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 1998, pp. 681–685. 3

[23] J. M. Saragih, S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *International Conference of Computer Vision (ICCV)*, September 2009. 3

[24] 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)*, 2012, pp. 2879–2886. 3

[25] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video," in *Proceedings of the British Machine and Vision Conference (BMVC)*, 2006, pp. 899–908. 3, 10

[26] A. Dhall, A. Asthana, and R. Goecke, "A ssim-based approach for finding similar facial expressions," in *Proceedings of the IEEE International Conference on Automatic Faces and Gesture Recognition and Workshop FERA*, 2011, pp. 815–820. 3

[27] 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)*, 2012, pp. 613–626. 3

[28] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 17–32. 4

[29] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition and Workshops (CVPRW)*, 2010, pp. 94–101. 6

[30] F. Wallhoff, "Facial expressions and emotion database," 2006, http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html. 6
[31] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2008, pp. 1–8. 6
[32] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008. 6
[33] F. Korc and D. Schneider, "Annotation tool," University of Bonn, Department of Photogrammetry, Tech. Rep. TR-IGG-P-2007-01, 2007. 6
[34] R. C. Prim, "Shortest connection networks and some generalizations," *Bell system technical journal*, vol. 36, no. 6, pp. 1389–1401, 1957. 7
[35] M. Eichner and V. Ferrari, "We Are Family: Joint Pose Estimation of Multiple Persons," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010, pp. 228–242. 7
[36] D. Lin and X. Tang, "Quality-Driven Face Occlusion Detection and Recovery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–7. 8
[37] R. Rosipal, *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*. ACCM, IGI Global, 2011, ch. Nonlinear Partial Least Squares: An Overview. 8
[38] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 657–664. 8
[39] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 24–31. 8
[40] W. R. Schwartz, H. Guo, and L. S. Davis, "A Robust and Scalable Approach to Face Identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010, pp. 476–489. 8
[41] H. M. Blalock, *Quantitative Sociology: International perspectives on mathematical and statistical model building*. Academic Press, 1975, ch. Path models with latent variables: The NIPALS approach. 8
[42] 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)*, 2007, pp. 401–408. 9
[43] 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)*, 2011, pp. 2092–2099. 9
[44] D. Parikh and K. Grauman, "Relative attributes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 503–510. 9
[45] F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 524–531. 9
[46] V. Jain, E. G. Learned-Miller, and A. McCallum, "People-lda: Anchoring topics to people using face recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8. 9, 10
[47] 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)*, 2011, pp. 121–128. 9
[48] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2001, pp. 601–608. 9
[49] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the ACM international conference on Image and Video Retrieval (CIVR)*, 2007, pp. 494–501. 10
[50] D. M. Blei and J. D. McAuliffe, "Supervised Topic Models," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2007. 10
[51] J. Zhu, A. Ahmed, and E. P. Xing, "Medlda: maximum margin supervised topic models for regression and classification," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009, p. 158. 10
[52] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. I–511. 10
[53] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001, http://www.csie.ntu.edu.tw/ cjlin/libsvm. 10, 11
[54] 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 workshop FERA*, 2011, pp. 878–883. 10
[55] 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 *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015. 12
[56] 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. 12
[57] N. Berthouze and L. Berthouze, "Exploring kansei in multimedia information," *Kansei Engineering International*, vol. 2, no. 2, pp. 1–10, 2001. 12

Abhinav Dhall is a postdoctoral research fellow at the Vision and Sensing Group, Human-Centred Technology Research Centre, University of Canberra and an adjunct research fellow at the Australian National University. He received his PhD in Computer Science from the Australian National University in 2014. He was awarded the Best Doctoral Paper Award at ACM ICMI 2013, Best Student Paper Honourable mention at IEEE FG 2013 and Best Paper Nomination at IEEE ICME 2012. His research interests are in computer vision for affective computing and social signal processing.

Roland Goecke is Professor of Affective Computing at the University of Canberra and an adjunct senior research fellow at the Australian National University. He is the Director of the Human-Centred Technology Research Centre and leads the Vision and Sensing Group, University of Canberra. He received his Masters degree in Computer Science from the University of Rostock, Germany, in 1998 and his PhD in Computer Science from the Australian National University, Canberra, Australia, in 2004. His research interests are in affective computing, pattern recognition, computer vision, human-computer interaction and multimodal signal processing.

Tom Gedeon is Chair Professor of Computer Science at the Australian National University and leads the Information and Human-Centred Computing Group at the Research School of Computer Science. His BSc and PhD are from the University of Western Australia. He is a former president of the Asia-Pacific Neural Network Assembly and a former President of the Computing Research and Education Association of Australasia. He serves on journal advisory boards as member or editor.