Automatic Happiness Strength Analysis of a Group of People using Facial Expressions

Sagiri Prasanthi[#1], Maddali M.V.M. Kumar[*2],

[#1]*PG Student,* [#2]*Assistant Professor*

[#1, #2]*Department of MCA, St. Ann's College of Engineering & Technology, Andhra Pradesh, India*

Abstract *- The latest improvement of social media has given users a stand to socially involve and interact with a higher population. Lakhs of videos, photos and group images are being uploaded daily by users on the web from different events and social gatherings. There is a collective concern in designing systems capability of under-standing human expressions of emotional attributes and affective displays. As images and videos from social events generally hold multiple subjects, it is an important step to study these sets of people. In this paper, we study the problem of happiness strength analysis of a set of people in a group image using facial expression analysis. A user awareness study is showed to understand several attributes, which affect a person's awareness of the happiness strength of a group. We detect the difficulties in developing an automatic mood analysis system and propose model built on the attributes in the study. An in the wild image based database is gathered. To functional the methods, both quantitative and qualitative experiments are done and applied to the problem of shot selection, event summarization and album creation. The experiments illustration that the attributes defined in the paper provide useful information for theme expression analysis, with results close to human awareness results.*

Keywords - *Facial expression recognition, group mood, unconstrained conditions.*

I. INTRODUCTION

Facial expression analysis has seen much research area in the recent times. Though, little care has been given to the estimation of the whole expression theme conveyed by a set 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 like a party, wedding or a graduation ceremony. Generally, these videos and photos were recorded in different conditions and may contain one or more subjects. From a view of automatic emotion analysis, these varied scenarios have received less attention in the affective enumerating community.

Consider an descriptive example of concluding the mood of a set of people posing for a group photograph at a school reunion. To scale the present emotion detection algorithms to work on this type of data *in the wild*, there are several difficulties to overcome such as emotion modelling in various sets of people, labelled data, and face analysis. Expression analysis has been a deep studied problem, focusing on inferring the emotional state of a single subject only. This paper discusses the problem of mood analysis of people in the group automatically. Here, we are fascinated in knowing an individual's strength of happiness and its role to the whole mood of the scene. The context can setup various factors like relative position of the person in the image, their distance from the camera and the level of face obstruction.

This model information based on a group graph, entrench these features in our method and stance the problem in a probabilistic graphical model built on a relatively weighted soft assignment. Analyzing the theme expression took by groups of people in images is an unfamiliar problem that has many real world applications: image search, retrieval, representation and browsing; event summarization and highlight creation; candid photo shot selection; expression apex detection in video; video thumbnail creation etc. A recent Forbes magazine article [1] deliberates the lack of capability of present image search engines to use context. Information like mood of a group can be used to model the context. These problems, where group mood information can be utilized, are a motivation for exploring the various group mood models. One basic methodology is to average the happiness intensities of all people in a group. However, the awareness of the mood of a group is defined by attributes like where people stand, how much of their face is visible etc. These social attributes play an key role in defining the whole happiness an image conveys.

II. LITERATURE REVIEW

A. Bottom-Up Techniques

Tracking sets of people in a crowd has been of specifically interest lately [2]. Based on trajectories built 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 watching 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 make attributes like age, gender and mien. In [3], the scene level happiness averages the individual persons' smiles. Though, in reality, group emotion is not an averaging model [8], [9]. There are attributes, which affect the awareness of a group's emotion and the emotion of the group itself. The literature in social psychology advices that group emotion can be conceptualized in different ways and is best represented by pairing the top-down and bottom-up approaches [8], [9].

Another fascinating bottom-up method is [10] proposed group classification for recognizing urban tribes. Low-level features like colour histograms, and high-level features like gender, age, hat and hair, were used as attributes to learn a Bag-of-Words (BoW) - based classifier. To add the group context, a histogram relating the distance between two faces and the number of over lapping 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 set of people based on their social group type".

B. Top-Down Techniques

In an interesting top-down method, [5] proposed contextual features based on the group building 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 people in a group in an image, which is discussed in this paper. Their experiments on images gotten from the web, show an impressive increase in performance when the group context is used. In other top-down approach, [6] model the social connection between people standing together in a group for aiding recognition. The social connections are inferred in unseen images by learning them from weakly labelled images. A graphical model based on social connections, such as 'father-child' and 'mother-child', and social connection features like relative height, height variance and face ratio. In [12] a face discovery method based on exploring social features like on social event images, is proposed. In object finding 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 connection between people, e.g. between friends on Facebook, using a Conditional Random Field (CRF) [16]. Recently, [17] proposed a framework for picking 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 wished 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 built on various attributes like blinking of eye, clarity of face and face mien. 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 collaborative of audio-video features, like motion, shot switch rate, frame brightness, pitch, and bandwidth, roll off, and spectral flux. However, expression information for individuals or groups in the scenes was not used.

The literature for analyzing a single subject's smile is rich. One prominent method by [19] proposed a new image-based database labelled for smiling and nonsmiling images and assessed various state of the art methods for smile detection. However, in the existing literature, the faces are measured independent of each other. For computing the role of each subject, two types of factors affect group level emotion analysis: (1) Local factors (individual subject level): age, gender, face mien and visibility, eye blink etc. (2) Global factors: where do people stand, with whom people stand etc. In this paper, the concentration is on visibility of face, smile strength, relative face size and relative face distance. Labelled images containing sets of people are required, which we collect from Flickr.

III. MODEL

We gathered a labelled *in the wild* database called HAPpy PEople Images (HAPPEI) – from Flickr holding 4886 images. A program was developed to automatically search and download images, which had keywords associated with sets of people and events. A total of 40 keywords were used (e.g. 'ceremony + graduation', 'people + party', 'photo + group', 'marriage', 'bar', 'reunion', 'function', 'convocation').

Fig. 1 (a): A collage of sample images in HAPPEI.

After taking the images, a Viola-Jones object detector trained on different data was executed on the images. Only images holding more than one subject were kept. False detections were manually

removed. Figure 1(a) shows a collage of images from the database. All images were annotated with a group level mood strength ('neutral' to 'thrilled'). Moreover, in the 4886 images, 8500 faces were manually annotated for face level happiness strength, obstruction strength and mien by four human labelers, who annotated different images. The mood was represented by the happiness strength corresponding to six stages of happiness: Neutral, Small Smile, Large Smile, Small Laugh, Large Laugh and Thrilled (Figure 1(b)).

Fig. 1 (b): Sample face level happiness strength labels in HAPPEI

As a reference during labelling, when the teeth of a member of a label Smile. The LabelMe [21] based Bonn annotation tool [22] 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 awareness 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 built on the awareness of the labelers. One can see this work as a stepping stone to group mood analysis. The aim of the model proposed in this work is to infer the perceived group mood as closely as possible to human observers.

IV. EXPERIMENTS

A. Face Processing Pipeline

Given an image, Viola-Jones (VJ) object detector [23] models trained on frontal and profile faces are applied to the images. For extracting the fiducial points, the part-based point detector of [20] is applied. The resulting nine points describe the location of the right and left corners of both eyes, the center 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 gathered from Flickr, having different scenarios and multifaceted backgrounds, standard face detectors like VJ object detector, result in a fairly high false positive rate (13.6%). To minimize this error, a non-linear binary SVM [24] is trained. The exercise set has samples of faces and non-faces. For face examples, all correct positives from the output of the VJ detector applied to 1300 images from the HAPPEI database are selected. For non-faces, the examples are manually selected from the same VJ output. To create a large number of untruthful positives from real world data, an image group having 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

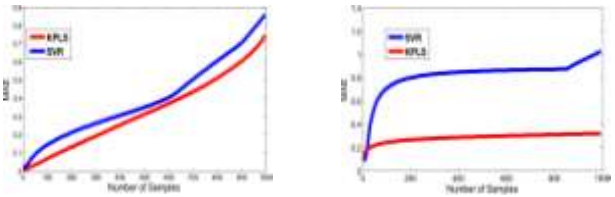
B. Implementation Details

Given a test image I containing group G, the faces in the group are first detected and aligned, then cropped to a size of 70 x 70 pixels. For the happiness strength detection, PHOG features are mined from the face. Here, the pyramid level L = 3, angle range = [0-360] and bin count = 16. The number of latent variables is picked as 18 after experiential 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 strength 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 $\Upsilon = 0{:}1$, which are weights that control the effect of 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, GEMw and GEM_{LDA}), both quantitative and qualitative experiments are performed. 2000 faces are used for training and 1000 for testing of the happiness and obstruction strength regression models.

C. Human Label Comparison

The Mean Average Error (MAE) is used as performance measure. The performance of the KPLS based obstruction strength and happiness strength estimators is compared with Support Vector Regression (SVR) [53] based obstruction strength and happiness strength estimators. Figure 2 displays the comparison based on the MAE scores.



(a) Happiness Strength (b) Obstruction Strength

Fig. 2: Comparison of happiness and obstruction strength methods.

The MAE for obstruction strength is 0.79 for KPLS and 1.03 for SVR. The MAE for happiness strength estimation for KPLS is 0.798 and for SVR 0.965. Table 1 shows the MAE comparison of GEM, GEMw and GEM_{LDA}. As hypothesized, the effect of adding social features is evident in the lower MAE in GEMw and GEM_{LDA}.

Method	GEM	GEM_w	GEM_{LDA}
MAE	0.455	0.434	0.379

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

D. User Study

A total of 15 subjects participated in a two-part user survey and were wished to a) rate happiness intensities in 40 images and b) rate the outcome of the methods for their outcome of the top 5 happiest images from an event. At this time, the users were requested to provide a score in the range of 0 (not good at all) to 5 (very good) for the methods for social events each. The users did not know, which output belonged to which method. For part a), Figure 3 shows the output. Note that the happiness scores computed by the GEMw are close to the mean human score and are well within the range of the standard deviation of the human labelers' scores. In the top row having high happiness strength in the set of people shows in images. The groups in the lower row have a lower happiness strength. From Figure 3, it is obvious that the top and bottom bounds of the happiness strength range assigned by the participants to the row 1 images are generally higher than the strengths assigned to the row 3 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 compared to others. This can be attributed to the difference in awareness of survey participants, as for different people different objects can be more salient.

Fig. 3: The graph defines the assessment of the group mood strength as evaluated by the proposed method with the results from the user study. The row 1 illustrations images with highest strength score and the row 3 shows images which are close to neutral. Please keep that the images are from different events.

For part b), ANOVA tests were implemented with the axiom that adding social context to group mood analysis leads to an estimate nearer to human awareness. For GEM and GEMw, $p < 0.0006$, which is statistically significant in the one-way ANOVA. For GEM and GEM_{LDA}, $p < 0.0002$, which is also statistically compelling

E. Image Ranking from an Event

For comparison, the proposed framework, volunteers were requested to rank a group of images containing a set of people from an event in the next task:

Fig. 4: The row 1 holds images from a ceremony of graduation organized by timestamps. The row 2 holds images ranked by human annotators in order of decreasing happiness strength (from left to right). The row 3 holds images ranked by decreasing happiness strength (from left to right) by GEMw

Given a social event with the same or diverse 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 reducing amount of happiness strength. Figure 4 is a screenshot of an event ranking experiment. In the row 1, the images are set based on their timestamp, i.e. when they were shot. The row 2 shows the ranking by human labelers. The highest happiness strength image is on the left and decreases from left to right. In comparison, the output of GEMw is in row 3, where the proposed method ranked the images in order of their decreasing happiness strength.

F. Candid Group Shot Selection

There are conditions in social congregations when multiple photographs are taken for the same subjects in a similar scene within a short span of time. Due to the vibrant nature of sets of people, it is a difficult task to identify the most favorable expression together in a set of people. Here, the group mood analysis method is enforced to shot selection after a number of pictures have been taken. In Figure 5, the rows show the shots appropriated at short intervals. The GEMw ranks the images having the same subjects and the best image i.e. Highest happiness proportion is displayed in the fourth column.

Fig. 5: Candid Group Shot Selection: Each row signifies a series of photos of the same people. The fourth column is the selected shot built on the maximum score from GEMw.

V. CONCLUSIONS

Social events creates several group shots. In this paper, a framework for estimating the group mood from an image, focusing on positive mood, is proposed. To the best of our knowledge, this is the first work for analyzing group mood built on the structure of a group attributes such as obstruction. An *in the wild* database called HAPPEI is gathered 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 illustration that assigning relative weights to intensities helps to better predict the group mood. The feature augmented topic model based group mood analysis model executives better than the average and weighted group expressions models.

In this work, for inferring the group mood, the global social features are built on the relative location of a person. The aim is to discover of salient or important faces, which can be the leader in the group. An exciting wayis to compute image saliency and weight the confidence of subjects who fall in the highly salient area. A natural extension lead of the proposed work is adding negative emotion group images to the database and framework [25]. Further, human body mien can be merged with the face analysis of a set of people. Based on recent work by [26], body mien can convey affect information. In the images downloaded from the internet, there can be challenges like face blur and obstruction due to neighbours in a group. This can make the inference of the mood of a person non-trivial. Body mien 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 like one mentioned in the Kansei image retrieval systems [27]. In the future, social context factors like age and gender, will be explored.

REFERENCES

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

[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.

[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.

[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.

[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.

[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, 2010, pp. 169–182.

[7] C. Kü blbeck 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.

[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.

[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, and 2001.

[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, 2012, pp. 28–35.

[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.

[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.

[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.

[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, 2008, pp. 1–8.

[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.

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

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

[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.

[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.

[20] 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), 2006, pp. 899–908

[21] 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.

[22] F. Korc and D. Schneider, "Annotation tool," University of Bonn, Department of Photogrammetry, Tech. Rep. TR-IGG-P-2007-01, 2007.

[23] 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.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001

[25] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The More the Merrier: Analysing the Effect of a Group of People In Images," in Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2015.

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

[27] N. Berthouze and L. Berthouze, "Exploring kansei in multimedia information," Kansei Engineering International, vol. 2, no. 2, pp.1–10, 2001.