Role of Group Level Affect to Find The Most Influential Person in Images

Shreya Ghosh and Abhinav Dhall

Learning Affect and Semantic Image analysIs (LASII) Group
Indian Institute of Technology Ropar
http://iitrpr.ac.in/lasii/
{shreya.ghosh,abhinav}@iitrpr.ac.in

Abstract. Group affect analysis is an important cue for predicting various group traits. Generally, the estimation of the group affect, emotional responses, eye gaze and position of people in images are the important cues to identify an important person from a group of people. The main focus of this paper is to explore the importance of group affect in finding the representative of a group. We call that person the "Most Influential Person" (for the first impression) or "leader" of a group. In order to identify the main visual cues for "Most Influential Person", we conducted a user survey. Based on the survey statistics, we annotate the "influential persons" in 1000 images of Group AFfect database (GAF 2.0) via LabelMe toolbox and propose the **"GAF-personage database"**. In order to identify "Most Influential Person", we proposed a DNN based Multiple Instance Learning (Deep MIL) method which takes deep facial features as input. To leverage the deep facial features, we first predict the individual emotion probabilities via CapsNet and rank the detected faces on the basis of it. Then, we extract deep facial features of the top-3 faces via VGG-16 network. Our method performs better than maximum facial area and saliency-based importance methods and achieves the human-level perception of "Most Influential Person" at group-level.

Keywords: Important person, Group of people, Group level affect.

1 Introduction

Nowadays, Social Networking sites have created a huge audience for everyone. A large number of images are uploaded every day on various social portals such as Facebook, Instagram, Google+, LinkedIn and others. These images mainly contain multiple subjects with a nice variety of context, lighting conditions, camera quality and other factors. Due to the above-mentioned reasons, affective computing community gets an opportunity to analyze the pattern of these data in terms of affect, behaviour, cohesiveness, event information, kinship, group norms and culture for a group of people. Moreover, when a group of people pose for a photograph, there exists some reason behind it. It may be some sort of social events (birthday, wedding, cultural festival, meetings after a long time and so on), professional reasons (office meetings, office party, interview etc.) or something else. Thus, it will be interesting to find out who is most "important" personality in the above-mentioned context-photographs.

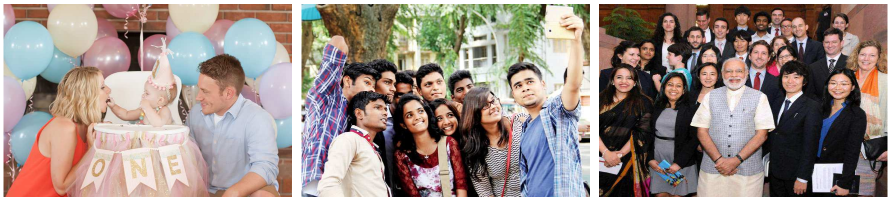

Fig. 1: The left image is a group where the baby in the centre is the most “important”. In the centre image, although it is a friend circle, still from the survey the boy holding the phone is the “important” one. Finally, in the rightmost image it’s about socially prominent people but without this info, our model tries to predict **“Who is the most important person?”**

“Importance” is an ambiguous term in case of real-world images. It has many perspectives such as photographer’s point of view, social norms and viewers’ (third person) perspective. When a photographer takes some photographs, he/she aims to capture some sort of “importance” in that image. Sometimes the main aim of the photographer remains unknown to the third person or viewer. In most of the cases, the camera angle and focus play a vital role in the perception of importance. Generally, human being pays attention to the larger object in an image instead of the background i.e. size and sharpness of an object draws attention. According to social norms, the relative position of people also matters a lot. Especially in the case of social events and office party, mostly the important one will be in the centre. Although in family scenarios, these things vary a lot. In many cases, like the rightmost image of figure 1, there is a presence of socially prominent personality. While annotating those images, people presume them to be most important. Predicting an important person in such images is really a challenging task. Our proposed method does not have such bias as it tried to predict “important” person on the basis of visual cues and group emotion intensities without any prior information.

Despite the above-mentioned challenge, there are several other challenges, such as diverse changes in human pose, action, appearance and occlusion involved in this task. Moreover, there is a lot of variation in context, background, illumination and lighting conditions. The automated system has to take care of facial and image level information to deal with these challenges. In some recent works, [25] attempted to detect “key actors” via attention model which takes human action and appearance as input. Solomon et al. [30] trained a regression model on spatial and saliency information to infer relative importance between two people. [22] used semantic information such as interactions between persons, eye gaze information which is essentially used to infer about the importance of persons. Some psychological studies [26] reveal that group emotion also plays a vital role in the identification of most influential[1] person (in other words leader).

The main objective of our study is to answer the following questions-

- How useful are face-level and group affect features for predicting an important person in an image?

[1] Please note that we use important and influential terms interchangeably throughout the paper.

- What are the factors, which affect the perception of the important person in a group image?

Automatic identification of the most influential person at first impression (first look) has several real-world applications. It can be used for im2text applications [30] (generating sentences that describe an image), event summarization, image retrieval, web crawling, "smart-cropping" of images [30] and ranking of personal photos etc.

Our contributions in this paper are as follows:

1) We propose an automatic "Most Influential Person" detection method via group level emotion. It performs better than our three baselines as mentioned in Section 3.

2) We labelled GAF 2.0 [5] dataset with "influential person" annotation and proposed "GAF-personage" database.

The rest of the paper structure is as follows: Section 2 is all about the prior work in this field. Section 3 describes the dataset, data annotation and survey statistics. Section 4 is about our approach towards this problem. Section 5 contains the details of the experiments we conducted on behalf of our method. Finally, Section 6 states the conclusion and future scope of this project.

2 Prior Work

One of the first group related analysis was proposed by Ge et al. [12] using a bottom-up hierarchical clustering algorithm. The motivation of the paper is to spread situation awareness and evacuation planning in real-time especially in case of huge conjugation. Further, several studies are conducted in order to understand several group traits.

2.1 Finding Important Persons

Recently, Li et al. [22] propose a Hybrid Interaction Graph (HIG) to rank people present in an image. This HIG includes spatial score, action score, appearance score and attention score. Spatial score and appearance score correspond to the location and attributes of the persons respectively. Action score indicates pose of the person and attention score includes eye gaze as an attribute.

In another interesting work, Solomon et al. [30] propose a measure of importance in terms of person level features such as position, scale, sharpness, facial pose and occlusion. Results show that there is a small correlation between importance and visual saliency. A text corresponding to each image is also generated which describes the image.

2.2 Importance in Images

Several works [31,18,35] study the importance of objects in an image. Yamaguchi et al. [35] define "importance" via several human perceived factors which are related to compositions (i.e. size and location of objects), semantics (i.e. object type, scene type along with its description strength) and context of the given image. The results also state that in any image "person" can be classified as the most important.

There is a huge difference between the "image level importance" and "important person" [30,22] as it requires a more coarse level understanding of the image.

2.3 Image Saliency

Several studies [8,14] try to figure out the part of the image which draws the viewer's attention. Mostly, human mind judges on the basis of image saliency. Jiang et al. [20] study image level saliency in crowd images. The main objective of the paper [20] is to find salient regions in images and use these as a feature to predict the crowded context as well as the crowd levels. Here, multiple kernel learning (MKL) is used for feature integration and extraction of important information.

However, there is a significant difference between image saliency and importance. Saliency [16] tries to predict the most eye-catching regions in image whereas importance takes context and other factors into account.

2.4 Group Affect

The first group affect analysis was conducted by Dhall et al. [6] where both facial and contextual information are taken into consideration. [4] divides group affect analysis approaches into two broad categories: bottom-up and top-down approach. The bottom-up approaches first analyze the group-members individually and then evaluate the contribution of these members towards the overall group's mood. The main motivation behind the top-down approach is to determine global factors and it's impacts on the perception of group level emotion. Dhall et al. [4] propose the use of low-level features for inferring an individual's happiness intensity and then pooled it at a global level.

In another interesting study, Hernandez et al. [15] conduct an interesting experiment at MIT, where the facial expression of the people passing through the corridor was analyzed for the presence of smile. The number of smiles are averaged at a given point to decide the overall group-level mood. Barsade et al. [2] propose that the social norms and its constraints (i.e. interpersonal cohesion and individual emotional responses) are the important cues for group emotion. In another paper, Gallagher et al. [10] argue that social context plays an important role in group-level scenarios. They modelled the group as a min-span tree. The task in the paper is to infer the gender and age of group members using the group-level contextual information. Dhall et al. [5] compute a scene level descriptor to encode the background information along with the facial and body cues. Huang et al. [17] model the group using a conditional random field and represent faces with a local binary pattern variant.
Mou et al. [23] perform an interesting study of human-affect on individual and group scenarios. They create three models as mentioned below:

1) An individual model which is trained with an individual level dataset.
2) Group model which is trained with a group dataset and
3) Combined model is the hybrid fused model of above two.

Smith et al. [29] argue that the group-level emotion is different from individual emotion. In order to predict an individual's role in the overall group emotion, one should study two factors.

First, a person's involvement in a group.

Second, his/her behaviour with the group members.

2.5 Multiple instance Learning

Multiple Instance Learning was introduced by Dietterich et al. [7] for drug activity prediction. Andrews et al. [1] propose two SVM based MIL methods for classification. The methods are named mi-SVM (for instance-level classification) and MI-SVM (bag-level classification). There are several papers [11,38] which use neural networks to explore this problem. Most of the computer vision tasks such as face detection [36], segmentation [33] and so on can fit into multiple instance learning framework.

In a recent paper, Xu et al. [34] propose a weakly supervised deep learning based MIL method in medical image processing. Further, Zhu et al. [39] propose a multiple instance learning methods with salient windows. The main aim of this method is unsupervised object detection. Wu et al. [32] use both CNN and DNN based multiple instance learning methods for image classification as well as image auto-annotation task. Zhu et al. [40] propose a deep multi-instance framework (sparse label assignment) for the breast cancer classification task. In a recent archive paper, Ilse et al. [19] propose attention based multiple instance learning framework for learning Bernoulli's distribution (at bag label).

3 Dataset Collection

Group AFfect dataset (GAF 2.0) is proposed by Dhall et al. [5] which contains group images in real-world scenarios. The images are collected via web crawling. Event and group related keywords such as party, family, protest, club, graduation ceremony and so on are used to find group images. These images are labelled into three group level emotion categories (positive, negative and neutral). We choose 1000 images from GAF 2.0 dataset[2] which uniformly belongs to three classes i.e. positive, negative and neutral respectively. We use MTCNN face detection library [37] to select those 1000 images which contains three and more than three faces. Further, we conduct a survey to observe how people decide the "importance" in a given group image.

3.1 Survey and Data Annotation

We conduct a survey of 10 images via Google form over 50 people having different occupations (for example student, corporate employee, govt. employee, professor and manager). The snapshots of the form is shown in figure 2. One has to choose an option on the support of "who seems to be most influential person?" Besides, one has to give reasons regarding his/her choice. The order of the faces in survey images are selected at random and the number assigned to a face in an image remains same throughout the survey.

The survey statistics and results are shown in figure 3. The first row in figure 3 describes the age distribution and gender distribution of the participants respectively. From the top left image of figure 3, it is observed that the age of the participants are varied from 17-57 years. There are 59.6% male participants and 40.4% females participants (in the top right image of figure 3). From the participants' responses regarding

[2] The datasets mentioned in [22,30] are not publicly available in the respective websites.

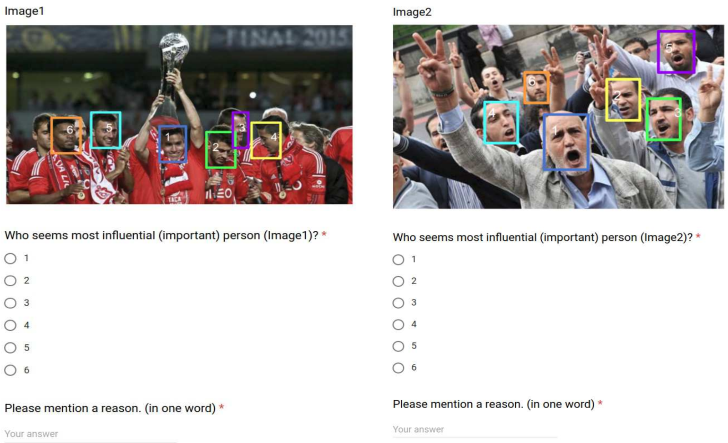

Image1

Who seems most influential (important) person (Image1)? *

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6

Please mention a reason. (in one word) *

Your answer

Image2

Who seems most influential (important) person (Image2)? *

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6

Please mention a reason. (in one word) *

Your answer

Fig. 2: These are the snapshots of the survey form.

their respective choices, we form a word cloud (in the middle images of 2^{nd}, 3^{rd} and 4^{th} rows of figure 3). From this statistics, we observed that people labelled on the basis of the image level, face level and position features. For example, the image present in 2^{nd} row, the main focus is on the context feature i.e. trophy. The frequent occurrence of *'happy', 'smiling', 'angry', 'front'* and *'centre'* keywords throughout the responses indicate group level emotion and position information. We use group affect for choosing the same number of faces across images for further analysis because in the survey result people mention emotion attributes for choosing a particular face (for example 'happy', 'smiling' and 'angry').

Keeping all of this factors in mind, 3 annotators annotate the proposed dataset **"GAF-personage"** via LabelMe online toolbox [27]. Before starting annotation, we explained the survey statistics and the trends of the choices (for example in family scenarios mainly children are given preference, socially prominent people present in an image get preference and so on). For the baseline, we choose the central face of the image, image saliency and maximum facial area which will be discussed in the next subsection.

3.2 Baselines for the Importance Model

From the survey statistics, we observe that people mainly focus on the center of an image. To consider the central face of an image as baseline, we first determine the central pixel of the image. Then, we find the nearby face via the distance between the center of the image and the tip of the nose of the detected face. This is a very weak baseline because in the real world scenarios it is not necessary that the photographer is in front of the main subject.

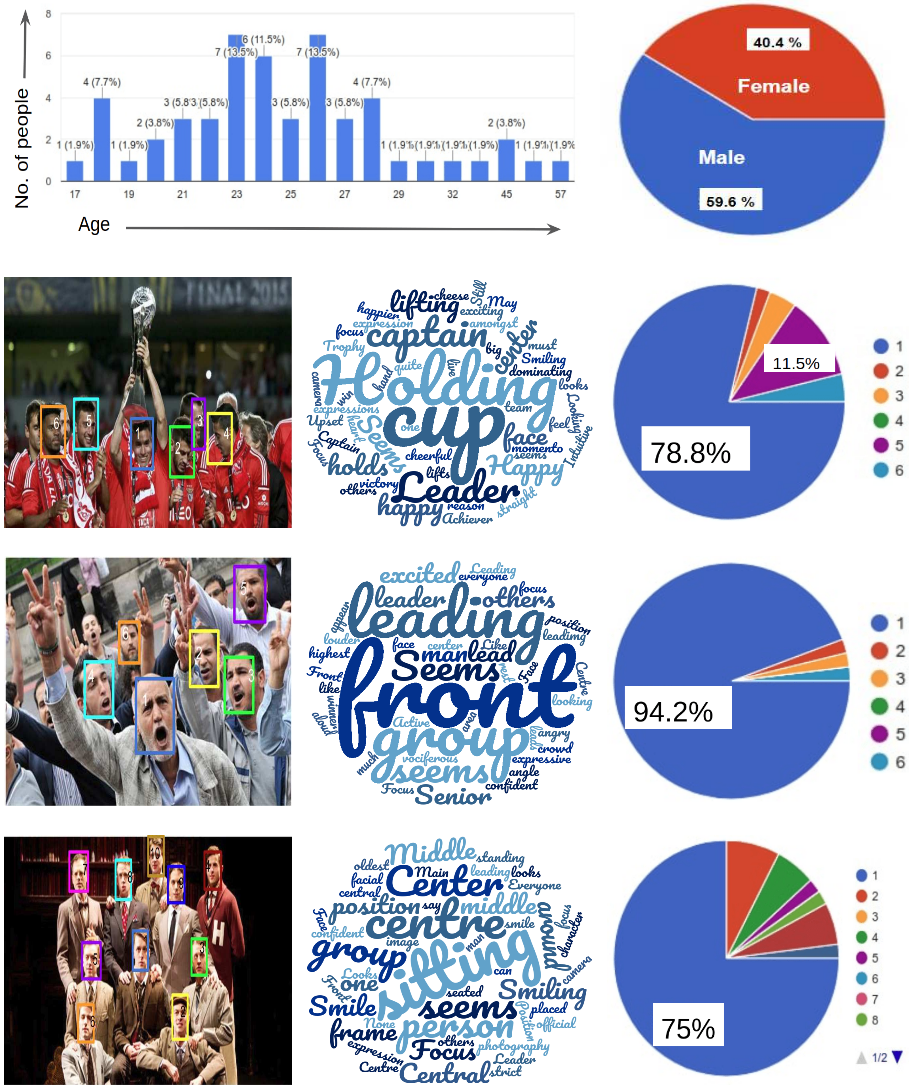

Fig. 3: This figure describes the survey results. The top left image describes the age distribution and top right image describes the gender distribution of the participants. In case of the top left figure, the x-axis indicates the age distribution and y-axis indicates percentage. For the 2^{nd}, 3^{rd} and 4^{th} row, the first column is the given survey image, the second column is the reason specified in the survey to choose “influential” person and the third column is about the voting results.

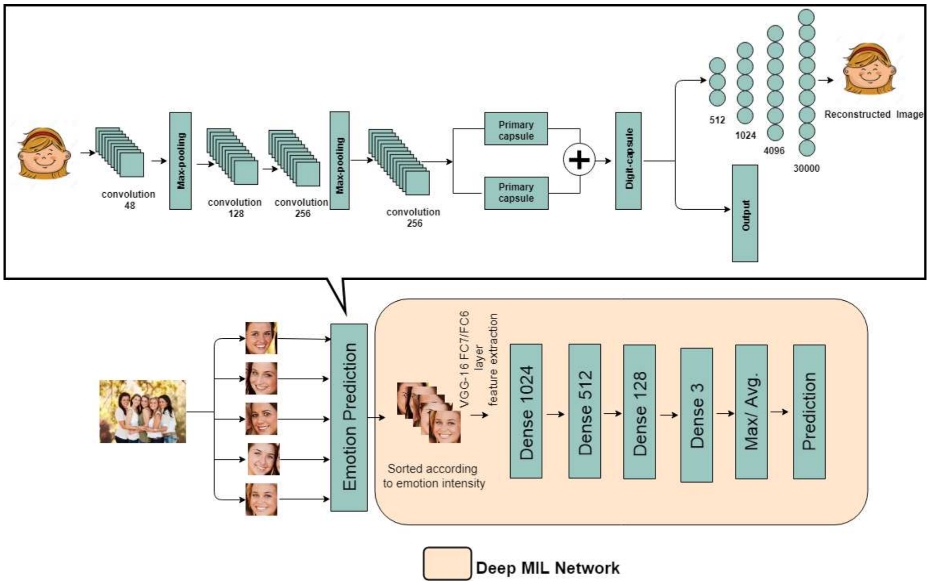

Fig. 4: This figure describes the overall pipeline of the proposed method. The top box describes the CapsNet structure which is used for emotion prediction. The bottom yellow coloured DNN is our proposed Deep MIL architecture.

We choose image saliency as another baseline because generally people judges on the basis of salient regions. The five-fold cross-validation accuracy of the saliency based prediction and ground truth is shown in table 2.

Similarly, we choose the maximum facial area as another baseline because it is also an important factor to identify most important person [35]. The five-fold cross-validation accuracy of the maximum facial area based prediction and ground truth is shown in table 2.

Thus, we choose three baselines for the “importance-model”:
1) Center of the image,
2) Image Saliency and
3) Maximum facial area.

4 Proposed Network

In this section, we describe our proposed method. Our proposed pipeline is shown in figure 4 which consists of two structures. The top box predicts individual-level emotion and the bottom coloured network is deep MIL based DNN which is used for final prediction.

Given an image containing three or more than three people, we first detect faces using MTCNN library [37] which uses three-stage cascaded CNN to detect faces. According to [26], group affect and leadership (influential person) are correlated. Due to this, we first sort the faces according to the emotional intensity of each face and choose the top three faces because our labelled images contain three and more than three faces.

4.1 Emotion Intensity Estimation

Recently, Sabour et al. [28] proposed capsule network which is able to capture spatial information. Due to this reason, it is used for capturing facial expressions [13] and AU detection [9] in recent studies. There are several challenges in leveraging facial information as it can be occluded, blurred or rotated. We trained the model structure as mentioned in [13] (refer figure 4 upper box). From this model, we get class wise group affect probabilities of each face which gives us the information about which face contributes more to the group affect. After sorting, we take only top three faces because the minimum three faces are present in our dataset images.

4.2 Deep MIL (DNN) for Importance Estimation

In **Multiple Instance Learning** [32] terms, the image is considered as a bag $X = \{X_1, .., X_N\}$, where each X_i is the elements inside the bag (also called instance/ feature vector) where X_i's are the ith element in the bag and N is the total number of elements in the bag. In our case, the group image stands for the bag which contains N elements (faces).

In order to keep the number of elements in each bag uniform, we sort the faces according to their respective emotional intensity provided by CapsNet. Further, we extracted VGG-16 FC7 and FC6 layer features of the top 3 faces sorted by the method mentioned above. This VGG-16 network is pre-trained on the VGG-face dataset which contains identity corresponding to faces. Thus, the FC layers of the VGG network contains high-level facial features which leverage facial structure and pose related information.

From the FC7 and FC6 layers of VGG face network, we got 4096-dimensional facial information which is further passed through several FC layers as shown in table 1. At last, we took the maximum and average of the features to predict the overall concept of the most influential person.

For our experiment, we use Swish [24] as an activation function instead of ReLU. The Swish activation function is defined as $f(x) = x.sigmoid(x)$. It has few properties (like unbounded above and bounded below) similar to ReLU and few different properties (like smooth and non-monotonic). The bounded below property of the above function helps in regularization of the network. Similarly, it does not reach near zero gradient due to its "unbounded above" property. Thus, the network can train at a faster rate. Besides, it's self-gating properties allows scaler values only instead of multiple gating inputs.

5 Experiments & Results

In this section, we will describe experimental details. For implementation, we use Keras [3] deep learning library with Tensorflow backend. The data, labels and code will be made publicly available (link).

Layers	**Input**	**Output**	**Layer Details**
Dense	b,3,4096	b,3,1024	1024
Activation	b,3,1024	b,3,1024	Relu/Swish
Dense	b,3,1024	b,3,512	512
Activation	b,3,512	b,3,512	Relu/Swish
Dropout	b,3,512	b,3,512	0.5
Dense	b,3,512	b,3,128	128
Activation	b,3,128	b,3,128	Relu/Swish
Dropout	b,3,128	b,3,128	0.3
Dense	b,3,128	b,3,3	3
Activation	b,3,3	b,3,3	Relu/Swish
Max-Pooling & Flatten	b,3,3	b,3	3(1-D)

Table 1: This is the detail architecture of the proposed network. Here, b refer to the batch size.

5.1 Emotion Prediction Network

We train a capsule network to predict emotions. The structure of the network is mentioned in the figure 4. We train the network on RAF-DB [21] dataset which contains approximately 30k facial images with 7-dimensional expression distribution (happy, sad, surprise, neutral, fear, angry and disgust). The input image passes through several convolution layers followed by max pooling layer before entering into two parallel primary capsule layers. A capsule is a set of nested neural network layers where a neural layer resides inside another. We use ‘adam’ optimizer with its default settings in keras library to train this network. The loss is the same as the original paper [28], that is margin loss for classification and mean square error for image reconstruction. It reaches the accuracy within 20-25 epochs without any data augmentation.

From group images, we detect faces via MTCNN [37] face detection library. We resize the detected faces to 100×100 dimension and predict corresponding emotion probabilities. Then, we sort the faces according to this probabilities and take the top three faces for further analysis. We observe that among 1000 images 713 images (approx. 71.3%) have their respective ground truth in this top-3 category.

5.2 Deep MIL Network (DNN)

First, we extract the VGG-16 FC7 and FC6 layer features of each top 3 faces which passes through several dense layers as mentioned in the table 1 before prediction. We first use parallel three networks for each top 3 faces. Then, we take maximum and average of the outputs to predict final “influential person”.

In order to train this Deep MIL network, we use mean square error as loss function and SGD optimizer with learning rate 0.01 and momentum 0.9 without any learning

MIL fold	Accuracy (%) (for VGG16 FC7 feature with max-pooling)	Accuracy (%) (for VGG16 FC7 feature with avg-pooling)	Accuracy (%) (for VGG16 FC6 feature with max-pooling)	Accuracy (%) (for VGG16 FC6 feature with avg-pooling)	Accuracy (%) (max facial area)	Accuracy (%) (Saliency [16])
1^{st} fold	57.49	62.50	76.00	73.50	57.20	55.20
2^{nd} fold	65.00	60.00	73.00	73.50	57.20	60.40
3^{rd} fold	66.50	60.00	71.50	73.50	55.82	56.30
4^{th} fold	56.49	65.50	77.00	76.00	49.30	50.50
5^{th} fold	62.50	75.00	74.50	76.00	53.40	56.30
Avg	**61.60**	**64.60**	**74.40**	**74.50**	**53.32**	**55.74**

Table 2: Experimental results for finding the most "influential person" in an image.

rate decay. Instead of avg-pooling, we also tried with max-pooling before prediction but the results are better in terms of accuracy in case of avg-pooling.

5.3 Result Analysis

From the table 2, we can conclude that our method is performing better than the maximum facial area concept as well as the saliency concept. Although saliency plays an important part in case of labelling as human mind make the first perception mostly by judging saliency of an image. Similarly, the maximum facial area also plays an important role in the perception of "most influential person" because in the survey results for second image, people choose both the frontal face (having maximum area as well) to be important. We also observe that the average pooling before prediction performs better than max pooling because average pooling infers overall statistics from an image where max pooling deals with specific statistics from images.

From the table 3, we can say that our model can detect most important person more precisely than 2^{nd} most important person. In a more fine-grained analysis, we can observe that in case of negative group-affect scenario it is almost similar in both cases because of expression intensity and camera angle. The situation changes in the case of happy images where the precision is relatively lower. The main reason behind this

Class	Precision (most important)	Precision (2nd most important)
Negative	0.6667	0.6667
Positive	0.4530	0.3125
Neutral	0.8547	0.6837
Overall	**0.6520**	**0.5479**

Table 3: This table contains the precision of emotion wise person prediction because GAF 2.0 dataset consists of group emotion images which consists of three classes (positive, negative and neutral).

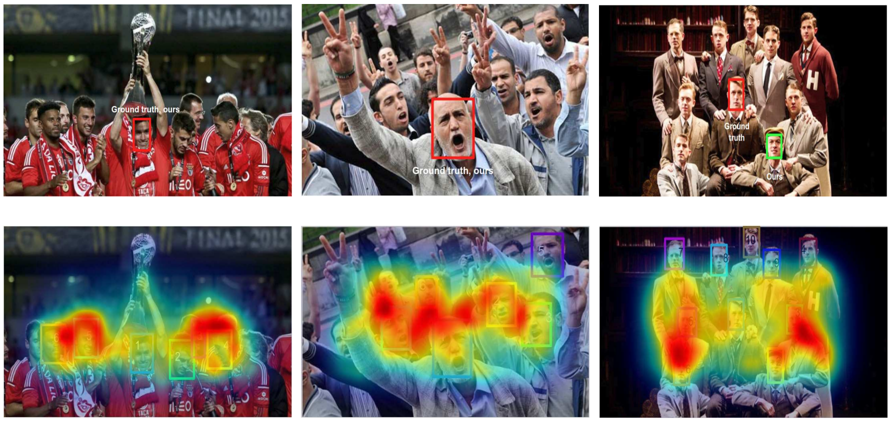

Fig. 5: This figure shows the output of our model. The left and middle image of the first row predicts the most important person correctly but for the first row rightmost image it selects the person with green boundary box as the most important person. We computed the saliency map via Matlab toolbox proposed by Hou et al. [16]. From saliency map, it is clear that there is a difference between saliency and importance.

is that the smile intensity changes a lot over images and the classifier get confused to choose the “influential person”.

6 Conclusion & Future Work

We study the importance of group affect to predict the most influential person. We first sort the faces according to the emotional intensity and then treat the problem as multiple instance learning problem. The results are better than the maximum facial area in the image and image saliency. Thus, we conclude that -

- Both face-level and group level features are important for predicting an important person in an image. When we sort the top three faces, we observed that 73.3% important people are included. Thus, it is clear that group level feature (here group affect) is important. Similarly, we perform our MIL experiments on the basis of deep facial features and results show that it is an important feature.
- From the survey, it is observed that the position of the person is an important motivation behind the “important” perception. Along with that facial cues, overall group affect is also a relevant indication.

In our pipeline, we have not included the position, body-pose, personal attributes, personality and eye gaze information. Besides, we can also analyze the fashion quotient of the group image especially in some social context such as a wedding, prize distribution ceremony, birthday party and so on. It will be interesting to combine all these factors to predict the probable “Leader” of a group.

7 Acknowledgement

We acknowledge the support of NVIDIA for providing us TITAN Xp G5X 12GB GPU for the research purpose. We are thankful to the anonymous reviewers for their insightful comments and helpful suggestions to improve the quality of this paper. We would also like to thank the members of LASII lab for their support.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in neural information processing systems. pp. 577–584 (2003)
2. Barsade, S.G., Gibson, D.E.: Group emotion: A view from top and bottom. Composition. (1998)
3. Chollet, F., et al.: Keras (2015)
4. Dhall, A., Goecke, R., Gedeon, T.: Automatic group happiness intensity analysis. IEEE Transactions on Affective Computing,2015
5. Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: From individual to group-level emotion recognition: Emotiw 5.0. In: ACM ICMI (2017)
6. Dhall, A., Joshi, J., Radwan, I., Goecke, R.: Finding happiest moments in a social context. In: Asian Conference on Computer Vision. pp. 613–626. Springer (2012)
7. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)
8. Elazary, L., Itti, L.: Interesting objects are visually salient. Journal of vision **8**(3), 3–3 (2008)
9. Ertugrul, I.O., Jeni, L.A., Cohn, J.F.: Facscaps: Pose-independent facial action coding with capsules
10. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: IEEE CVPR (2009)
11. Garcez, A.d., Zaverucha, G.: Multi-instance learning using recurrent neural networks. In: Neural Networks (IJCNN), The 2012 International Joint Conference on. pp. 1–6. IEEE (2012)
12. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012)
13. Ghosh, S., Dhall, A., Sebe, N.: Automatic group affect analysis in images via visual attribute and feature networks. In: IEEE International Conference on Image Processing (ICIP). IEEE (2018)
14. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in neural information processing systems. pp. 545–552 (2007)
15. Hernandez, J., Hoque, M.E., Drevo, W., Picard, R.W.: Mood meter: counting smiles in the wild. In: ACM UbiComp (2012)
16. Hou, X., Harel, J., Koch, C.: Image signature: Highlighting sparse salient regions. IEEE transactions on pattern analysis and machine intelligence **34**(1), 194–201 (2012)
17. Huang, X., Dhall, A., Zhao, G., Goecke, R., Pietikäinen, M.: Riesz-based volume local binary pattern and A novel group expression model for group happiness intensity analysis. In: BMVC (2015)
18. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. International journal of computer vision **100**(2), 134–153 (2012)
19. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712 (2018)

20. Jiang, M., Xu, J., Zhao, Q.: Saliency in crowd. In: European Conference on Computer Vision. pp. 17–32. Springer (2014)
21. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593. IEEE (2017)
22. Li, W.H., Li, B., Zheng, W.S.: Personrank: Detecting important people in images. In: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. pp. 234–241. IEEE (2018)
23. Mou, W., Gunes, H., Patras, I.: Alone versus in-a-group: A comparative analysis of facial affect recognition. In: ACM Multimedia (2016)
24. Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941 (2017)
25. Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., Fei-Fei, L.: Detecting events and key actors in multi-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3043–3053 (2016)
26. Redl, F.: Group emotion and leadership. Psychiatry **5**(4), 573–596 (1942)
27. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. International journal of computer vision **77**(1-3), 157–173 (2008)
28. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in Neural Information Processing Systems. pp. 3856–3866 (2017)
29. Smith, E.R., Seger, C.R., Mackie, D.M.: Can emotions be truly group level? evidence regarding four conceptual criteria. Journal of personality and social psychology (2007)
30. Solomon Mathialagan, C., Gallagher, A.C., Batra, D.: Vip: Finding important people in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4858–4866 (2015)
31. Spain, M., Perona, P.: Measuring and predicting object importance. International Journal of Computer Vision **91**(1), 59–76 (2011)
32. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3460–3469 (2015)
33. Wu, J., Zhao, Y., Zhu, J.Y., Luo, S., Tu, Z.: Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 256–263 (2014)
34. Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Eric, I., Chang, C.: Deep learning of feature representation with multiple instance learning for medical image analysis. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. pp. 1626–1630. IEEE (2014)
35. Yamaguchi, K., Stratos, K., Berg, A.C., Sood, A., Mitchell, M., Mensch, A., Goyal, A., Han, X., Dodge, J., Daume, H., et al.: Understanding and predicting importance in images. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3562–3569. IEEE (2012)
36. Zhang, C., Platt, J.C., Viola, P.A.: Multiple instance boosting for object detection. In: Advances in neural information processing systems. pp. 1417–1424 (2006)
37. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
38. Zhou, Z.H., Zhang, M.L.: Neural networks for multi-instance learning. In: Proceedings of the International Conference on Intelligent Information Technology, Beijing, China. pp. 455–459 (2002)

39. Zhu, J.Y., Wu, J., Xu, Y., Chang, E., Tu, Z.: Unsupervised object class discovery via saliency-guided multiple class learning. IEEE transactions on pattern analysis and machine intelligence **37**(4), 862–875 (2015)
40. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 603–611. Springer (2017)