Context based Facial Expression Analysis in the Wild

Abhinav Dhall
School of Computer Science, CECS, Australian National University, Australia
abhinav.dhall@anu.edu.au
http://users.cecs.anu.edu.au/∼adhall

***Abstract*—With the advances in the computer vision in the past few years, analysis of human facial expressions has gained attention. Facial expression analysis is now an active field of research for over two decades now. However, still there are a lot of questions unanswered. This project will explore and devise algorithms and techniques for facial expression analysis in practical environments. Methods will also be developed for inferring the emotion of a group of people. The central hypothesis of the project is that close to real-world data can be extracted from movies and facial expression analysis on movies is a stepping stone for moving to analysis in the real-world. The data extracted from movies carries with it rich meta-data information which will be useful for exploring the role of context in emotion recognition in the wild. For the analysis of groups of people various attributes effect the perception of mood. Study will be conducted and thorough literature survey will be performed on what are these contextual attributes. A system which can classify the mood of a group of people in videos will be developed and will be used to solve the problem of efficient image browsing and retrieval based on emotion.**

I. Introduction

In the recent years, social media has become very popular. On social portals such as YouTube and Flickr, users are uploading millions of images and videos related to social events everyday. This has created new challenges for efficient multimedia retrieval, browsing and management. Researchers have been using low-level audio and visual features combined with meta-data such as tags for retrieval and browsing. We hypothesise that as these images and videos generally contain people, facial expression analysis can be used as a meta-data information as well for affective browsing and retrieval. Generally, the video clips and images uploaded by users have been recorded in different conditions and may contain one or more subjects. From the perspective of automatic expression analysis, these diverse scenarios are unaddressed. We call these cases: 'facial expression analysis in the wild', where 'wild' corresponds to the diverse environment and conditions, in which the data has been captured. This PhD project aims at developing algorithms, techniques & applications for Facial Expression Recognition (FER) in practical environments. Further, models are proposed for modelling the expression of both individual and groups of people.

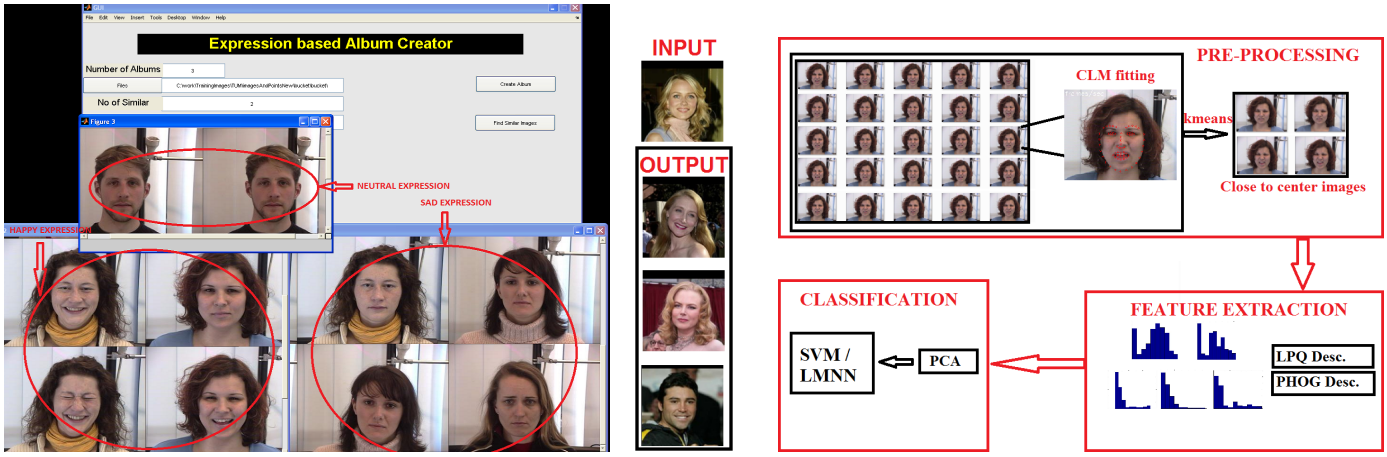
II. Motivation & Background

Affective computing was coined by Prof. Rosalind Picard in the 1990s, since then it has come a long way. Affective computing encompasses where sub-problems such as facial expression analysis [29], body pose analysis [15] and medical problems such as depression [13], pain [23] and stress analysis [22]. In particular interest to this project, facial expression analysis has been a long studied problem [28], [3]. FER methods can be divided into static and temporal FER methods. Static methods generally deal with a frame based FER [3], [24]. Human facial expressions are dynamic in nature. Psychologist [1] have argued that the temporal information is necessary and more informative for FER. Temporal FER methods deal with videos. There has been interesting research in the temporal FER [20], [8], [12]. In this PhD project, both static and temporal facial expression methods are explored. Though temporal FER is preferred but in scenarios where only images are available (such as Flickr.com), FER methods need to infer the expression using static information only. FER methods can also be classified on the basis of number of subjects in a sample: individual or group. Given that the social event data on the internet may contain multiple subjects, recently Dhall *et al.* [10] have proposed models based on social context for mood analysis groups of people.

Broadly, FER systems can also be broadly divided into three categories based on the type of feature representation used: a) shape features based FER methods: where a geometric representation of the face is used [5]; b) appearance feature based FER method: where texture information is used [3] [6] and c) hybrid FER methods which use both shape and appearance descriptors [17]. From designing a FER method which can work in real-world conditions, choosing the right descriptor is essential such that the facial dynamics are captured and the representation does not reach Bayes risk. Also, from an information retrieval perspective, affect is used as an attribute. For inferring affect, the systems generally need to be fast enough for efficient retrieval. This poses the problem of selecting robust features which can be computed efficiently.

Generally FER methods have been limited to lab-controlled data. This poses a complex problem of extending and porting the methods created for lab controlled data to real-world conditions. There has been little work on affect analysis in real-world scenarios. One of early work is by Zhang et al. [30] who analysed the affect of movies for efficient browsing of videos. However, this method does not take into consideration the facial expressions which are a strong cue. Neither this nor earlier affect analysis methods take into consideration the presence of multiple subjects. However, for moving to real-world scenarios, one needs to take these situations (multiple people in a scene) into consideration. This can be used for browsing and retrieving images and videos which contain subject(s) in tough conditions.

Due to the nature (lab-controlled, posed) of the currently available databases [16], [26], [21], current FER systems seldom use context information. Context recently has caught attention in a lot of vision based face analysis problems [11], [27], [18], [19]. In an interesting work, Gallahger *et al.* use the neighborhood information in group photographs as a prior for inference of age and gender. In their experiments, they found that age and gender inference can gain using the context computed based on the neighbors of a subject. [18], use the neighbor subject's information for inference of identity of a subject.



(a) (b)

Fig. 1. a) Expression based album creation and album by similar expression [5]. b) Key-frame based emotion detection [8].

III. Challenges

To transfer the current facial expression algorithms to work on data in the wild, there are several challenges. Consider an illustrative example of categorization i.e. assigning an emotion label to a video clip of a subject(s) protesting at the Tahrir square in Egypt during the 2011 protests. In order to learn an automatic system which can infer the label representing the expression, we require labelled data containing video clips representing different expressions in diverse settings, along with a label which defines the emotional state. Traditionally, emotion recognition has been focussing on data collected in very controlled environments, such as research laboratories. Ideally, one would like to collect spontaneous data in real-world conditions. However, as anyone working in the emotion research community will testify, collecting spontaneous data in real-world conditions is a tedious task. As the subject in the example video clip generally moves his/her head, it poses another challenge of out-of-plane head movements. Head pose normalisation methods are required to capture the temporal dynamics of facial activity. With analyzing spontaneous expressions comes the problem of occlusion as subjects move their arms and hands as part of the non-verbal communication, this inter-occlusion needs to be handled for correct label inference.

The complexity of such video clips (like the one in the example above) increases with the presence of multiple subjects. Research in this field has been focussing on recognition of a single subject's emotion i.e. given a video clip or image, only a single subject is present in it. However, data being uploaded on the web, specially revolving around social events such as the illustrated example contains groups of people. Group mood analysis finds it's application in opinion mining, image and video album creation, image visualisation and early violence prediction among others. There has been work in psychology on analysis of emotion of group of people, cues from this can be taken on creating models for handling group emotions. The **major challenges** for group mood analysis are: 1) labelled data representing various social scenarios; 2) robust face and fiducial points detector (this is relevant for a single subject scenario as well, though not for lab-controlled scenario databases)and 3) models which can take into consideration the affective compositional effects and the affective context. A simple solution to group mood analysis is emotion averaging. However, in real-world conditions averaging is not ideal. This motivates us to research for models which accommodate various attributes that effect the perception of group mood and their interaction.

IV. Proposed Methods

A. Facial Expression Based Album Creation

With the advancement in digital sensors, users captures a lot of images in scenarios like social events and trips. This leads to a complex task of efficiently retrieving and browsing through these huge collection of images. Generally people are the main focus of interest in social events. A structural similarity based method is proposed in [5]. Given an image with a face, fiducial points are computed using constrained local models. These fiducial points are used to compute a new geometric feature called Expression Image (EI). EI captures the shape of a face which is representation of an expression. Now given images in an album, EI is computed for each image and a structural similarity based clustering algorithm is applied. This creates clusters representing different emotions and such cluster representatives can be used as emotive thumbnails for browsing an album.

Further a 'browse by expression' extension is proposed, where given an input image with a subject showing a particular expression, the system retrieves images with similar expressions. Figure 1(a) defines the method output, the three sub groups in red circles are the cluster centres generated by similar expressions and the second illustration with a celebrity input image and the result images are similar expression images. Facial performance transfer [2] and similar expression [6] based classification have been explored as extension to this method.

B. Facial Expression Recognition Challenge

As part of the PhD project, an emotion detection method is proposed based on selecting key-frames [6]. Fiducial points are extracted using CLM and clustering is performed on the normalised shape points of all the frames of a video clip. The cluster centres are then chosen as the key-frames on which texture descriptors are computed. On analysing visually, the cluster centres corresponded to various stages of an expression i.e. onset-apex-offset. The method preformed well on the both task (subject independent and dependent) in the first FERA 2011 challenge [25]. Figure 1(b) describes the steps involved in the system.

C. Facial Expressions In The Wild

As discussed in the Section III, data simulating 'in the wild' conditions is the first challenge for making FER methods work in real-world conditions. To over come this, it is proposed to extract data from movies [9]. Even though movies are made in controlled conditions, they still resemble real-world conditions and clearly actors

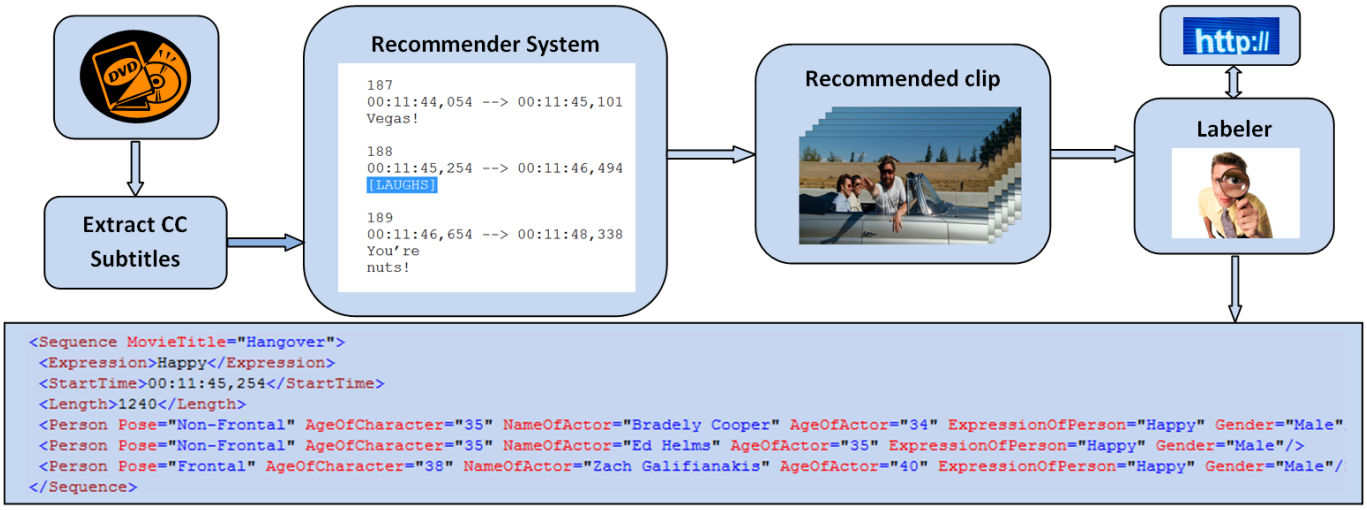


Fig. 2. **AFEW database creation pipeline:** Given a dvd movie, closed captions are extracted. The system scans the closed caption for emotion related keywords. Once a relevant keyword is found, the corresponding video clip is played to the human labeller. The human labeller then uses the internet movie database to label the meta-data. Please note that the meta-data contains context specific information - age of the actor and character, gender and subject identity.

in good movies try to emulate natural expressions. It is very difficult to collect spontaneous expressions in challenging environments. A semi-automatic recommender system based method is proposed for creating an expression dataset. The recommender system scans movie DVD's closed caption subtitles for emotion related keywords. Video clips containing these keywords are presented to an annotator, who then decides if the clip is useful. Meta-data information such as identity of actor, age, gender are stored for each video clip. This temporal dataset is called Acted Facial Expressions In The Wild database [9]. It contains 1426 short video clips and has been downloaded 50+ times in the past 14 months (http://cs.anu.edu.au/few).

Figure 2 defines the database construction process. Along with the expression information (*Angry*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad* and *Surprise*) of the subject(s) in the video, meta-data such as age of the actor, age of the character, gender and pose are also labelled. This meta-data can be used to introduce context. For example, it is interesting to see the effect of age and gender on FER systems. Further, information such as subject identity can be used to study the evolution of expression in actors.

A frame based database: Static Facial Expressions In The Wild (SFEW) has been extracted from AFEW. Currently there are 700 images. Strict experimentation protocols have been defined for both AFEW and SFEW. The experiments on the databases show the short coming of current state-of-art FER methods which perform very well on lab-controlled data.

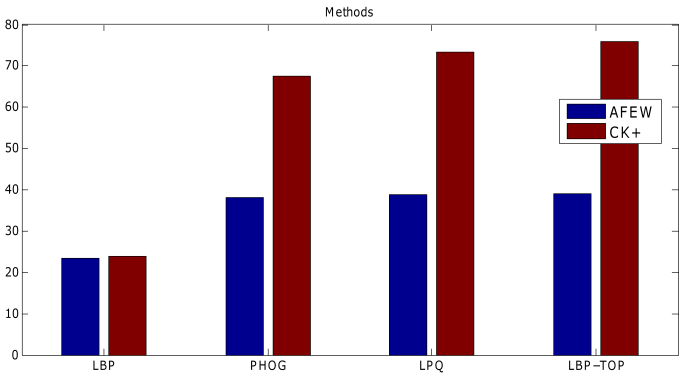


Fig. 3. Performance comparison of four texture descriptors: LBP, PHOG, LPQ and LBP-TOP on CK+ and AFEW. The results clearly show the gap in performance induced when the data is close to real-world scenarios.

1) Experiments: Figure 3, describes the performance of state-of-art descriptors on CK+ and AFEW. It is evident from the graph that the current methods do not scale well when tested on real-world conditions. On investigating further, it was found that there was a sharp drop in performance due to poor face detection. Further, in databases such as the CK+, the apex frame is known and is used for LBP-TOP. However, it is non-trivial to localise the apex frame in the AFEW database as there is no constraint on when the onset, apex and offset of the expression should be present.

For progress in the field of FER, a grand challenge and workshop: Emotion Recognition In The Wild (EmotiW)[1] is being organised as part of the ACM International Conference on MultiModal Interaction (ICMI 2013). Researchers are invited to test and extend their state-of-art methods on real-world data.

D. Group Mood Analysis

Images and videos uploaded on internet and in movies generally have multiple subjects. Emotion analysis of group of people is dependent on two main parts: the member's contribution and the scene context. As a pilot study, data is downloaded from Flickr based on keywords related to social events such as marriage, convocation, party etc. Face detector is applied on downloaded images and fast rejection is performed, if an image has less than three people. Then images are labelled for each person's happiness intensity, face clarity and pose. Also the mood of the group in an image is labelled. The database contains 8500 labelled faces and 4886 images.

1) Average model: A simple group analysis model is averaging of individual person's expression intensities (mood). The mood is inferred by learning regression model on face data from the HAPPEI database [10]. Kelly et al. [14] argue that the mood of a group is composed by two broad category of components: top-down and bottom-up. Top-down ('global context') is the affective context i.e. the effect induced by attributes such as group history, background, social event etc. Top-down has an effect on the group members. For example a group of people laughing in a party display 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

[1] http://cs.anu.edu.au/few/emotiw

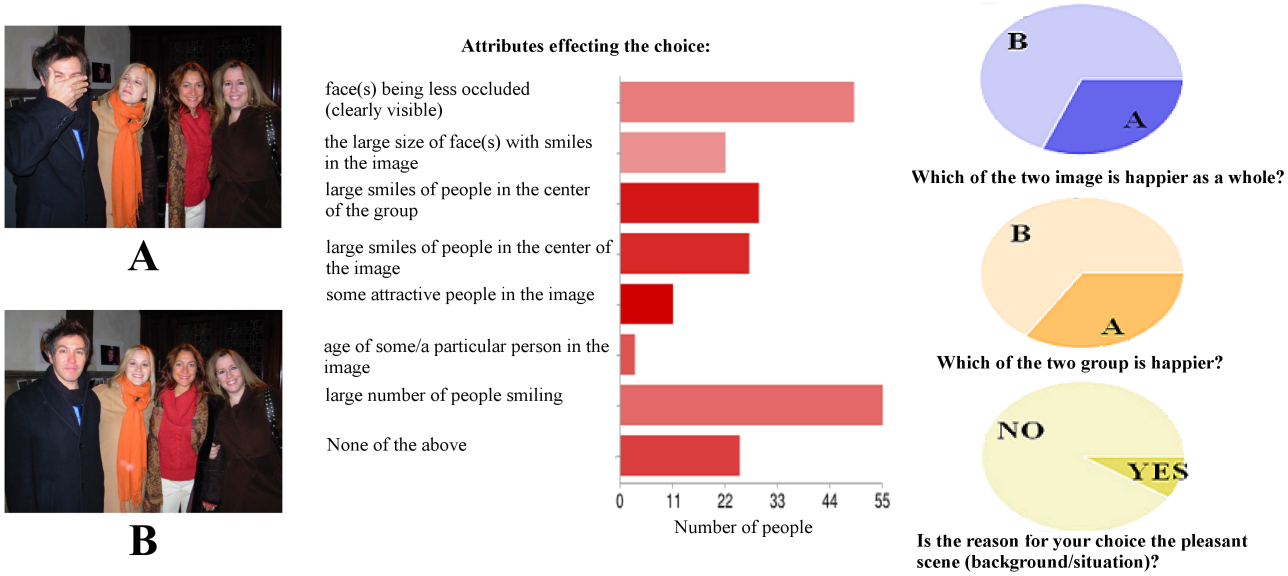


Fig. 4. The figure describes a case from the survey. On the left are the two images which were shown to the participants. In the middle the bar graphs shows the votes which each attribute got from the participants. The pie charts on the right show the voting for which group/image has a better mood.

be used as affective context. Bottom-up ('local context') component deals with the subject in the group. Attributes of individuals that effect the perception of group mood. The bottom-up component defines the contribution of individuals to the overall group mood.

Global context refers (but not limited to) to scene information, social event information, who is standing with whom, where are people standing in an image and with respect to the camera etc. Local context i.e. individual specific attributes cover individual's mood/emotion, face visibility, face size with respect to neighbors, age, gender, head pose and eye blink etc. To further understand these attributes a perception study was performed. In the survey (Figure 4), total of 150 individuals participated. The survey asked individuals to rank the happiness in set of two different images which contain group of people. To understand their perception behind making a decision various questions were asked such as: 'is your choice based on: large smiling faces; large number of people smiling; context and background; attractive subjects; age of a particular subject' and so on. After analysing this data various attributes are defined.

2) Weighted model: To add the context in the averaging model, a group of people in an image are represented as a fully connected graph, where faces are the vertices and the distance between two faces is the weight of the edges. Min-span tree is computed and it tell's the neighbor of each subject in the group. The neighborhood information is used to calculate the relative face size and relative face location of the subject. The hypothesis behind computing the relative face size, is that human visual system focuses mainly on salient faces. These salient faces are large and clear faces. This can also be argued as subjects standing in the front of the group have larger face sizes as compared to the ones standing in the back. Therefore, subjects in the front are more salient as their faces are clearer and larger. Figure 6(a), describes the effect of relative face size attribute. Further, the contribution of mood of a subject who is standing away from the group needs to be penalised. Relative distance is computed to penalise the subjects which are not members of the group i.e. standing away from the group [10].

Fig. 5. Group shot selection example. The first three columns show successively shot images and the fourth column is the recommended image based on highest mood score [10].

The local context is computed on the bases of a subject's mood and face clarity. Face clarity can also be seen as intra or inter-occlusion which leads to poor visibility of a subject's face. Regression models are learnt on HAPPEI database to infer the face occlusion and mood intensity [10]. The global and local context are applied as weights to the averaging model.

3) Feature augmented model: Group mood analysis is a weakly labelled problem. Even though a number of attributes (Figure 6(b)) are labelled in the training data, the survey conducted showed that there are factors such as age and gender too. To incorporate this information to the weighted group expression model, topic model is learnt. The attributes are augmented with a bag of words model which is learnt on low-level features extracted from faces. The augmented feature is then used to learn a graphical model. Experiments show that the performance of augmented feature based topic model is superior to that of weighted and average group expression models. Along with the quantitative analysis performed for comparing the proposed group expression models, various qualitative experiments are also conducted.

An interesting application of group expression analysis is group shot selection. Images are shot in succession/burst mode and mood is used as the deciding factor. Figure 5 describes the experiment, the fourth column displays the selected frame for each row of successive shots. The mood value of the group can be fused with other attributes such as the one mentioned in the Kansei image retrieval systems [4].

V. Research Progress and Future Work

The work discussed above has been published at various venues ([8], [9], [10], [5], [7]. Recently, in affective computing interesting work has come up describing body expression analysis [15]. In the mood analysis survey, participants mentioned body pose, background and clothes also being strong attributes which effected their perception about the mood of a group. In some scenarios where a face may not be

(a) (b)

Fig. 6. a) The figure describes the effect of social attributes on the expression intensity estimation of the image. *Top Left:* Original image, which a human labeler gave a happiness intensity score of 70. *Top Right:* Min-span tree depicting connection between faces where the edge weight is the distance between the faces. *Bottom Left:* Happiness intensity heat map using averaging; note how some faces relatively smaller in size are given a high weighting. *Bottom Right:* Happiness intensity heat map with social context, the contribution of face size and relative position of faces (F1 and F3) towards the overall intensity of the group is penalised. b) Attributes based approach for mood inference.

of sufficient resolution or blurred or occluded, body pose can provide crucial information about the mood of the person. Therefore, body expression, background and clothes describing attributes are being analysed and fused into the current model.

VI. Acknowledgement

My PhD research is sponsored by AusAid's Australian Leadership Award scholarship. I am thankful to my supervisors: Dr. Roland Goecke, Prof. Tom Gedeon, Dr. Simon Lucey and other collaborators at the Australian National University, University of Canberra, Commonwealth Scientific and Industrial Research Organisation and the University of California San Diego.

References

[1] Z. Ambadar, J. Schooler, and J. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, pages 403–410, 2005.

[2] A. Asthana, M. de la Hunty, A. Dhall, and R. Goecke. Facial performance transfer via deformable models and parametric correspondence. *IEEE TVCG*, pages 1511–1519, 2012.

[3] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *IEEE SMC*, 2004.

[4] N. Berthouze and L. Berthouze. Exploring kansei in multimedia information. *Kansei Engineering International*, pages 1–10, 2001.

[5] A. Dhall, A. Asthana, and R. Goecke. Facial expression based automatic album creation. In *ICONIP*, 2010.

[6] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using PHOG and LPQ features. In *IEEE AFGR2011 workshop FERA*, 2011.

[7] A. Dhall and R. Goecke. Group expression intensity estimation in videos via gaussian processes. In *ICPR*, 2012.

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static Facial Expression Analysis In Tough Conditions: Data, Evaluation Protocol And Benchmark. In *ICCVW*, BEFIT'11, 2011.

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. A semi-automatic method for collecting richly labelled large facial expression databases from movies. *IEEE Multimedia*, 2012.

[10] A. Dhall, J. Joshi, I. Radwan, and R. Goecke. Finding happiest moments in a social context. In *ACCV*, 2012.

[11] A. Gallagher and T. Chen. Understanding Images of Groups of People. In *CVPR*, 2009.

[12] T. Gehrig and H. K. Ekenel. Facial action unit detection using kernel partial least squares. In *ICCV Workshops*, pages 2092–2099, 2011.

[13] J. Joshi, A. Dhall, R. Goecke, M. Breakspear, and G. Parker. Neural-net classification for spatio-temporal descriptor based depression analysis. In *Proceedings of the International Conference on Pattern Recognition*, ICPR'12, pages 2634–2638, 2012.

[14] J. R. Kelly and S. G. Barsade. Mood and emotions in small groups and work teams. *Organizational behavior and human decision processes*, 86(1):99–130, 2001.

[15] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, PP(99):1, 2012.

[16] 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 *CVPR4HB10*, 2010.

[17] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn. AAM Derived Face Representations for Robust Facial Action Recognition. In *IEEE AFGR*, pages 155–162, 2006.

[18] O. K. Manyam, N. Kumar, P. N. Belhumeur, and D. J. Kriegman. Two faces are better than one: Face recognition in group photographs. In *IJCB*, pages 1–8, 2011.

[19] A. C. Murillo, I. S. Kwak, L. Bourdev, D. J. Kriegman, and S. Belongie. Urban tribes: Analyzing group photos from a social perspective. In *CVPR Workshops*, pages 28–35, 2012.

[20] M. Pantic, I. Patras, and M. Valstar. Learning spatiotemporal models of facial expressions. In *Int'l Conf. Measuring Behaviour 2005*, pages 7–10, August 2005.

[21] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, ICME'05, 2005.

[22] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke. Modeling stress using thermal facial patterns: A spatio-temporal approach. In *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction*, ACII'13, 2013.

[23] K. Sikka, A. Dhall, and M. Bartlett. Weakly supervised pain localization using multiple instance learning. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, FG'13, 2013.

[24] U. Tariq, J. Yang, and T. S. Huang. Multi-view facial expression recognition analysis with generic sparse coding feature. In *ECCV Workshops (3)*, pages 578–588, 2012.

[25] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. R. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE*

Transactions on Systems, Man, and Cybernetics, Part B, pages 966–979, 2012.

[26] F. Wallhoff. Facial expressions and emotion database, 2006. http://www.mmk.ei.tum.de/ waf/fgnet/feedtum.html.

[27] G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV (5)*, pages 169–182, 2010.

[28] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In *In CVPR*, pages 70–75. IEEE Computer Society, 1994.

[29] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE TPAMI*, pages 39–58, 2009.

[30] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li. Utilizing affective analysis for efficient movie browsing. In *ICIP*, pages 1853–1856, 2009.