Expression Analysis In The Wild: From Individual To Groups

Abhinav Dhall
Research School of Computer Science
Australian National University, ACT 2601, Australia
abhinav.dhall@anu.edu.au

1. 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. For the analysis of groups of people various attributes effect the perception of mood. 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.

Categories and Subject Descriptors

I.5 [**PATTERN RECOGNITION**]; I.4.8 [**IMAGE PROCESSING AND COMPUTER VISION**]: Scene Analysis; I.4.9 [**IMAGE PROCESSING AND COMPUTER VISION**]: Applications; H.5.1 [**INFORMATION INTERFACES AND PRESENTATION**]: Multimedia Information Systems

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Group mood analysis, Expression analysis in the wild

2. OBJECTIVE

Automatic facial expression analysis deals with a computer based inference of a person's internal emotional state. It is an active field of research with applications in the field of human computer interaction, information retrieval and management, affective computing, medical diagnosis such as depression, stress, pain, lie detection and many others. With the growing popularity of data sharing and broadcasting web site such as YouTube, Flickr; everyday users are uploading millions of videos and images of social events. Generally, these video clips and images have been recorded in different conditions and may contain one or more subjects. From the perspective of automatic emotion analysis, these diverse scenarios are unaddressed. This PhD project aims at developing algorithms, techniques & applications for Facial Expression Recognition (FER) in practical environments.

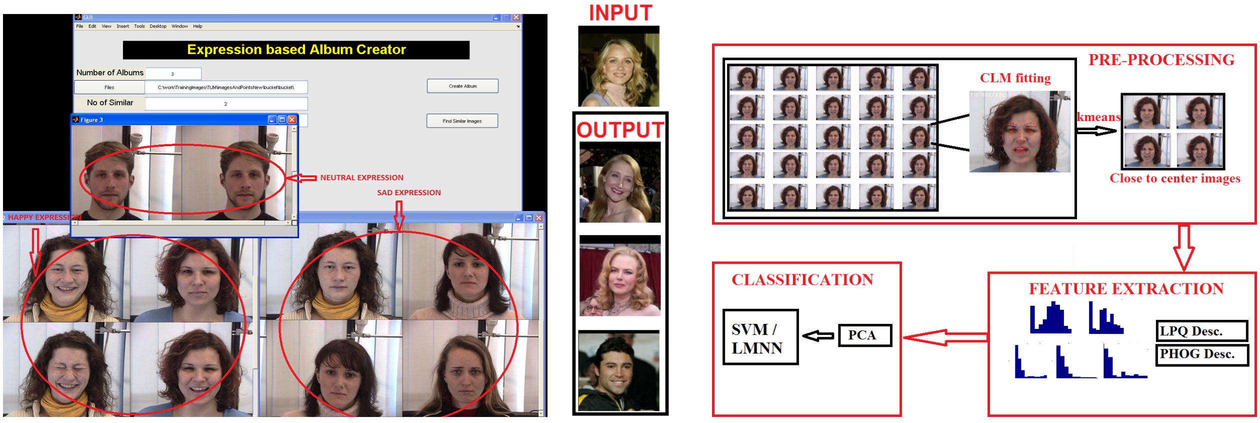
3. MOTIVATION & BACKGROUND

Facial expression analysis has been a long studied problem [13], [3]. FER methods can be divided into static and temporal FER methods. Static methods generally deal with a frame only based FER [3]. Temporal FER methods deal with videos as human facial expressions are dynamic in nature and are a better representation for expression representation [1]. 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, 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.

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 [4] ; 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 [12]. 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. [14] who analysed the affect of movies for efficient browsing

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
Copyright 2013 ACM 978-1-4503-2033-7/13/04 ...$15.00.



(a) (b)

Figure 1: a) Expression based album creation and album by similar expression [4]. b) Key-frame based emotion detection [8].

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.

4. 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 Tahir 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.

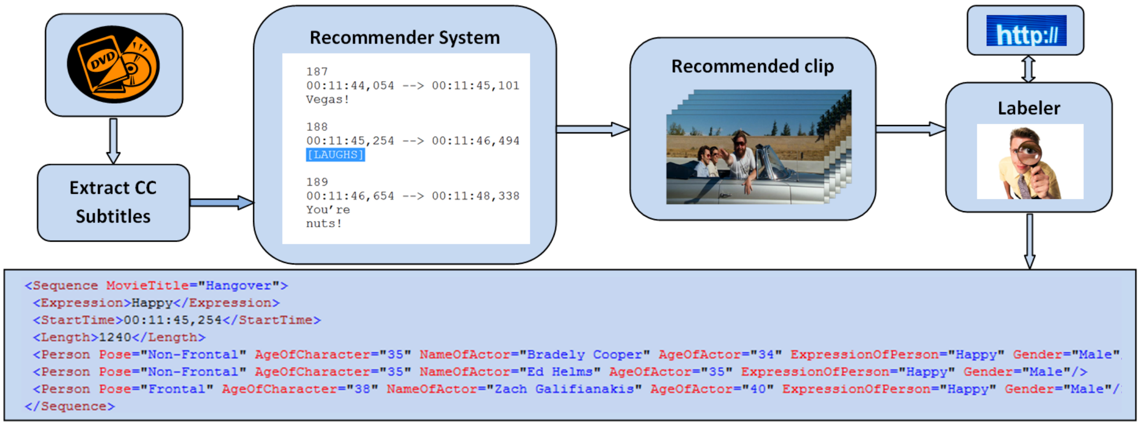
The complexity of such video clips (like the one in the example above) is increased with the presence of multiple subjects. Research in this field has been focussing on recognition of a single subject's emotion i.e. given a video clip or image, only a single subject is present in it. However, data being 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 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.

5. PROPOSED METHODS

5.1 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 [4]. 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



```
187
00:11:44,054 --> 00:11:45,101
Vegas!

188
00:11:45,254 --> 00:11:46,494
[LAUGHS]

189
00:11:46,654 --> 00:11:48,338
You're
nuts!
```

```
<Sequence MovieTitle="Hangover">
 <Expression>Happy</Expression>
 <StartTime>00:11:45,254</StartTime>
 <Length>1240</Length>
 <Person Pose="Non-Frontal" AgeOfCharacter="35" NameOfActor="Bradely Cooper" AgeOfActor="34" ExpressionOfPerson="Happy" Gender="Male"
 <Person Pose="Non-Frontal" AgeOfCharacter="35" NameOfActor="Ed Helms" AgeOfActor="35" ExpressionOfPerson="Happy" Gender="Male"/>
 <Person Pose="Frontal" AgeOfCharacter="38" NameOfActor="Zach Galifianakis" AgeOfActor="40" ExpressionOfPerson="Happy" Gender="Male"/
</Sequence>
```

Figure 2: AFEW database creation pipeline

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.

5.2 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 FERA 2011 challenge. Figure 1(b) describes the steps involved in the system.

5.3 Facial Expressions In The Wild

As discussed in the Section 4, 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 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).

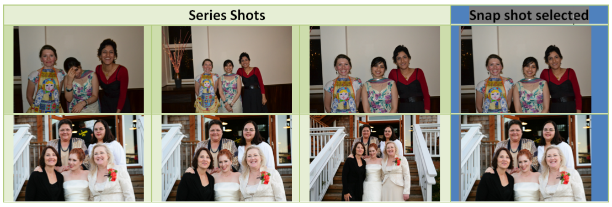


Figure 3: Group shot selection example. The first three columns show successively shot images and the fourth column is the recommended image based on highest mood score [10].

A image only dataset, 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. Figure 2(a) defines the database construction process.

For progress in the field of FER, a grand challenge and workshop: Emotion Recognition In The Wild (EmotiW) is being organised as part of the ACM International Conference on MultiModal Interaction (ICMI 2013).

5.4 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 4000 images.

A simple group analysis model is averaging of individual person's expression intensities. However, humans while perceiving the mood of group of people take into consideration various attributes. To understand these attributes, an online survey was conducted. In the survey, 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

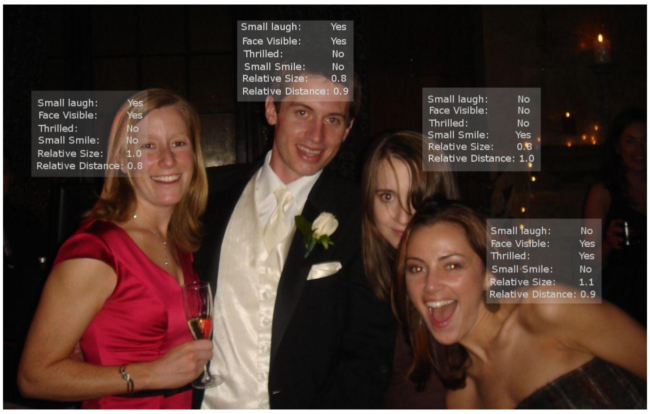


Figure 4: Attribute based group mood inference

decision various questions are 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.

Attributes which effect group mood perception can be categorized into global and local attributes. Global attributes include but not limited to the location of a person in a group, a person's neighbor and the scene. Local attribute comprise of an individual person's mood, face clarity and so on. Figure 2(b) defines the attributes computed for members of a group [10]. These attributes are used as weights to each person's contribution towards the overall group's mood.

Group mood analysis is a weakly labelled problem. Even though a number of attributes 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 3 describes the experiment, the fourth column displays the selected frame for each row of successive shots.

6. RESEARCH PROGRESS

First six months are thorough literature survey. Then similar expressions [4, 5] and FERA challenge [6] are explored. In the second year AFEW [9] and SFEW [8] databases are developed. Group expressions [10] [7] are explored in the third year. Research collaboration with graduate students at the HCC lab at the University of Canberra for depression detection [11].Thesis writing will be undertaken in the first half of the fourth year.

7. 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 and other collaborators at the Australian National University, University of Canberra, Commonwealth Scientific and Industrial Research Organisation and the University of California San Diego.

8. 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] A. Dhall, A. Asthana, and R. Goecke. Facial expression based automatic album creation. In *ICONIP*, pages 485–492, 2010.

[5] A. Dhall, A. Asthana, and R. Goecke. A ssim-based approach for finding similar facial expressions. In *IEEE AFGR2011 workshop FERA*, pages 815–820, 2011.

[6] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using PHOG and LPQ features. In *IEEE AFGR2011 workshop FERA*, pages 878–883, 2011.

[7] A. Dhall and R. Goecke. Group expression intensity estimation in videos via gaussian processes. In *ICPR*, pages 3525–3528, 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, pages 2106–2112, 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] J. Joshi, A. Dhall, R. Goecke, M. Breakspear, and G. Parker. Neural-net classification for spatio-temporal descriptor based depression analysis. In *ICPR*, pages 2634–2638, 2012.

[12] 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.

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

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