# Perception of Emotional Expressions in Different Representations Using Facial Feature Points

Shazia Afzal
Univ. of Cambridge
Computer Laboratory
sa437@cam.ac.uk

Tevfik Metin Sezgin
Koc Univ.
College of Engg.
mtsezgin@ku.edu.tr

Yujian Gao
Univ. of Cambridge
Computer Laboratory
yg259@cam.ac.uk

Peter Robinson
Univ. of Cambridge
Computer Laboratory
pr10@cam.ac.uk

## Abstract

*Facial expression recognition is an enabling technology for affective computing. Many existing facial expression analysis systems rely on automatically tracked facial feature points. Although psychologists have studied emotion perception from manually specified or marker-based point-light displays, no formal study exists on the amount of emotional information conveyed through automatically tracked feature points. We assess the utility of automatically extracted feature points in conveying emotions for posed and naturalistic data and present results from an experiment that compared human raters' judgements of emotional expressions between actual video clips and three automatically generated representations of them. The implications for optimal face representation and creation of realistic animations are discussed.*

## 1. Introduction

The face, as a modality for emotion recognition, has occupied a dominant position in the study of affect in human-machine interaction. This follows from the predominance of facial signs in human perception of emotion [5, 6] as well as the relative advantages it offers over other modalities like speech and physiology. Facial information can be detected and analysed unobtrusively and automatically in real-time, without requiring any specialised equipment except a simple video camera. Even though issues such as occlusion, lighting and pose variation still remain problematic, the field has seen some increasingly good results [13]. Mapping of facial expressions to affective states is however a challenging problem. Facial expressions are not simple read-outs of affective states and their interpretation is largely context-driven. To reduce this complexity for automatic affective inference, measurement and interpretation of facial expressions has traditionally been separated. However, in order to move from expression recognition to expression interpretation it is necessary to discriminate between facial configurations that have a psychological significance from those that have morphological value [4]. The success of this transition depends to a large degree on how much of the information relevant for affect perception is actually captured - or missed, by the techniques employed in facial affect analysis.

This paper investigates properties of one such method, namely facial feature point tracking, to explore the information value of automatically tracked facial landmarks in conveying emotions. We report results from an experiment analysing the emotion recognition accuracy of five emotions - interest, confusion, boredom, happiness and surprise, using samples obtained from posed and naturalistic databases of facial expressions, in different representations of varying information detail. Section 2 gives a background of feature point tracking and how it motivates our experiment. Section 3 provides a description of the data compiled for this study while Section 4 details the experimental design and procedure itself. We discuss our results and its implications in Sections 5 and 6.

## 2. Motivation

The typical sequence of steps in an automatic facial expression recognition system is *face acquisition*, followed by *facial feature extraction* and finally *facial expression classification* [6, 7, 13]. Facial feature extraction can be classified as either deformation-based or motion-based. Deformation extraction includes appearance–based techniques while motion extraction is feature-based and includes methods such as facial feature point tracking and geometric face models. Although appearance-based feature extraction methods yield better recognition results, they require extensive pre-processing (e.g. manual alignment and scaling) and are more sensitive to variation in pose, occlusion and lighting. Facial feature tracking on the other hand, is more robust to pose variation and can deal with partial occlusion. It is therefore considered more suitable for real-time automatic emotion classification, and has been used extensively in emotion recognition systems [13].

The motivation for using facial feature point tracking is based on psychological studies that emphasise the role of facial motion in the perception of emotional expressions [c.f. 3]. These employ an adaptation of Johanssen's [9] point-light display technique to analyse the contribution of facial movement in the discrimination of emotions. Point-light displays are constructed by recording blackened faces with numerous white spots or reflective markers while displaying emotional expressions. The white spots or markers are the only visible source of information and serve as the "carriers of motion",

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

independent of any form or appearance information. Feature-point tracking closely models this technique making it attractive for computational modelling and communication. Feature-point based representation thus reduces the complexity of visual input and provides a simplified depiction of the otherwise rich visual experience [12] while preserving the temporal structure of facial expressions which is crucial for emotion interpretation [3]. Because feature-point based representations are not affected by the idiosyncrasies of individuals' facial appearance, they also enable development of more generalisable computational techniques.

However, the actual utility of a feature-point-based representation in affect recognition depends on the degree to which affective information can be conveyed through the specific set of features used. Therefore it is important to understand how well a set of feature points can convey affective content, especially if the feature points are to be used for automatic facial expression recognition. This paper is an attempt in this direction. In particular, we measure the amount of affective information that can be conveyed by a set of 25 facial feature points used extensively in the automated affect recognition literature. This is done by asking human raters to identify emotions in sequences that are generated from automatically tracked feature points of videos displaying facial affect. In order to account for the effects of different representations on raters' judgements, we use three representation formats ranging from elementary point-light representations to intermediate stick-figure models, to complete and finer 3D ones. We make use of state-of-art automatic feature tracking technology to generate video sequences of different descriptive detail and compare human raters' performance on emotion perception.

# 3. Data Preparation

Our dataset for this experiment included samples taken from four different databases. These were selected to represent a range of posed and naturalistic experimental control conditions. For posed data samples, the Cohn-Kanade DFAT database [10] and the Mind Reading DVD [1] were selected. The DFAT database consists of image sequences of facial displays acted out by different encoders under explicit instructions from an experimenter. The Mind Reading DVD on the other hand consists of emotion samples from actors given example scenarios rather than specific instructions on facial displays. For the naturalistic data we used samples collected from simulated driving scenarios and a computer-based learning setting. The former falls into the category of induced emotion while the latter is completely naturalistic.

Three expert coders labelled the data samples from the different databases to create a final corpus of emotion samples. Five examples for each of interest, boredom, confusion, happiness and surprise were taken from each database based on perfect agreement by all three coders. The DFAT database lacked examples of interest and boredom giving us a dataset of 65 samples. Mean duration of the selected video clips was 3.46 seconds ($\sigma$ =1.99). Table 1 shows the distribution of samples and their average duration for each emotion category per database.

Table 1: Sample distribution across emotion categories

| *Database \ Emotion* | *DFAT* | *Mind Reading* | *Natural* | *No.* | *Duration (sec)* |
|---|---|---|---|---|---|
| Interest | - | 5 | 5 | 10 | $\mu$=4.1<br>($\sigma$=2.1) |
| Bored | - | 5 | 5 | 10 | $\mu$=5.2<br>($\sigma$=0.9) |
| Confusion | 5 | 5 | 5 | 15 | $\mu$=3.1<br>($\sigma$=2.2) |
| Happiness | 5 | 5 | 5 | 15 | $\mu$=3.0<br>($\sigma$=1.9) |
| Surprise | 5 | 5 | 5 | 15 | $\mu$=2.7<br>($\sigma$=1.8) |
| *No.* | 15 | 25 | 25 | **65** | |
| *Duration (sec)* | $\mu$=1.0<br>($\sigma$=0) | $\mu$=4.8<br>($\sigma$=1.3) | $\mu$=3.6<br>($\sigma$=1.8) | | **$\mu$=3.46<br>($\sigma$=1.99)** |

For each of the 65 video samples, three representations at varying levels of information detail were generated – point-based, stick-figures, and 3D animations. These representation formats were chosen because of their perceptual significance as well as their relevance in animation techniques. The renderings for each were generated using automatically tracked facial feature points on the original video clips. This controls for variation across displays and thus enables true comparison of human perceptual performance across displays [12]. In all, our final corpus containing the original 65 emotion samples and their three levels of representation totalled 260 video clips at 25fps.

## 3.1. Point-based Displays

The point-based representation was created from the output of an automatic face-tracker on a black background to resemble the point-light experimental stimuli (see Section 2). The face-tracker used to generate the point-based displays was selected after a careful review of available feature-point trackers, both research and commercial. This FaceTracker[1] is state-of-art in automatic facial feature point tracking and requires no manual pre-processing or calibration. It is resilient to limited out-of-plane motion, can deal with a wide range of physiognomies and can also track faces with glasses or facial hair.

## 3.2. Stick-figure Models

The stick-figure displays formed the next level of representation. A stick figure is an elementary drawing made of lines and dots, and was created by adding minimal detail to the landmarks, i.e., connecting the automatically tracked feature-points using straight lines and sketching eyes using typical shape. Eye height was

[1] http://www.nevenvision.com: Licensed from Google Inc.

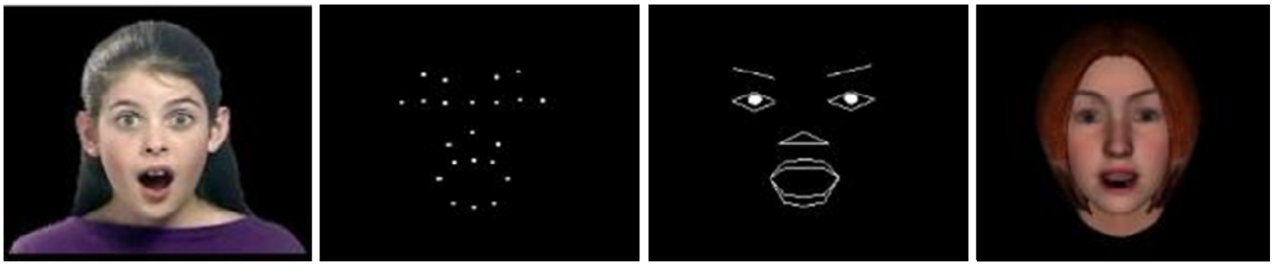


Figure 1: Example of the three representations generated from the original video using automatically tracked feature points

empirically computed as half of its width. Compared to point-based displays, the stick-figure representation presents the rough outline of facial features, and is therefore more face-like and familiar to people. The stick-figure models were also rendered on black background consistent with point-based displays.

### 3.3. 3D XFace Animations

XFace is an open source toolkit used for the creation of 3D animated facial expressions and displays. It implements an MPEG-4-based facial animation mechanism [11] and can generate 3D facial animation by simply inputting the facial animation parameters (FAPs). We chose XFace animation as the third representation because of its simplicity in usage and feature support for rendering animations using FAPs. FAPs are the basis of MPEG-4 Animation of synthetic face models. The automatically tracked feature points were directly converted into a set of FAPs for driving the animations.

Figure 1 shows examples of the different representations generated from an original emotion sequence showing ‘Surprise’.

## 4. Experiment Design

The objective of our study was to ascertain the information value of automatically tracked feature points in conveying emotions. Sample videos of selected emotions were used to generate three different facial representations. The aim was to analyse the perceptual differences in emotion recognition in different forms of representation. We were also interested to know whether elementary representations like point-based and stick-figure models made emotion perception easier and more accurate, or whether a complex 3D representation allowed for finer distinction. More specifically, we wanted to compare –

- How affect recognition accuracy differed across the three generated representations used in displaying facial information
- How affect recognition accuracy differed across databases
- How affect recognition accuracy differed across emotions
- How inter-rater agreement varied across these experimental conditions, and
- How affect recognition accuracy compares to the individual affect decoding ability of participants.

The experiment was designed as a within-subjects repeated measures study where each participant labelled all the sample video clips. To minimise order and practice effects, the presentation of clips was randomised across and within each participant.

### 4.1. Participants

14 participants (8 male, 6 female) in the age-group of 20 to 34 volunteered to take part in this study. All had normal or corrected-to-normal vision and were fluent in English. They were of diverse ethnicities and were reimbursed for their participation.

### 4.2. Stimulus Materials

The dataset compiled from selected original samples and their representations (as discussed in Section 3) formed the stimulus material for the experiment. For better visual fidelity all video sequences were presented at 320 x 240 pixels on a black background. See illustration in Figure 1.

### 4.3. Labelling Interface

The labelling procedure was programmed as a computer-based interface. It allowed participants to watch randomly presented video clips and label them for emotions. Labelling was disabled while a video was playing. In addition, no media controls like pause, rewind or forward were made available, except for a replay control which too was disabled while a video was playing. This was to ensure that all participants watched a video clip in its entirety without selective play and then marked an emotion label. A cross-hair was displayed between consecutive video clips to fixate attention and clear the mind from previous visualisation.

### 4.4. Procedure

Participants completed the experiment individually in our usability lab. Written instructions were provided about the nature of the task. They were informed that they would be shown different facial displays which they were required to judge for emotional content. Participants also read through an emotion word list before starting the experiment. This was to acquaint them with some emotion terms used in everyday language. After signing for consent and providing demographic information the participants underwent a brief training session.

Table 2: Mean percentage recognition accuracy for each emotion under each of the representations used; **O** - Original videos; **P** - Point-light displays; **S** - Stick-figure models and **X** - XFace animations

| **Database \ Emotion** | **DFAT** | | | | **MindReading** | | | | **Natural** | | | | ***Overall per emotion*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ***O*** | ***P*** | ***S*** | ***X*** | ***O*** | ***P*** | ***S*** | ***X*** | ***O*** | ***P*** | ***S*** | ***X*** | |
| **interested** | - | - | - | - | 77.1 | 40.0 | 51.4 | 35.7 | 51.4 | 27.1 | 41.4 | 44.3 | 46.1 |
| **confused** | 42.9 | 17.1 | 28.6 | 21.4 | 72.9 | 27.1 | 35.7 | 50.0 | 58.6 | 24.3 | 34.3 | 30.0 | 36.9 |
| **bored** | - | - | - | - | 91.4 | 35.7 | 31.4 | 15.7 | 81.4 | 42.9 | 50.0 | 8.6 | 44.6 |
| **happy** | 95.7 | 92.9 | 95.7 | 74.3 | 92.9 | 47.1 | 64.3 | 5.7 | 71.4 | 30.0 | 32.9 | 17.1 | 60.0 |
| **surprised** | 97.1 | 72.9 | 87.1 | 62.9 | 91.4 | 77.1 | 75.7 | 45.7 | 45.7 | 14.3 | 22.9 | 18.6 | 59.3 |
| ***Overall per database*** | 65.7 | | | | 53.2 | | | | 37.4 | | | | |

The training session was the same for all participants and consisted of carefully selected eight videos, two from each of the different representation formats. The videos used as stimuli for training were not included in the experimental set and were sampled from both posed and naturalistic databases.

After the training session, participants began with the labelling task, where the stimuli were presented to them in a randomised order. They were given the option to replay a video as many times as they wanted but were instructed to follow their initial reaction as much as they could. For each video a maximum of two labels – primary and secondary, was allowed. The secondary label was optional and participants were asked to make use of this sparingly. The option of labelling more than one way was provided in order to incorporate some level of flexibility in emotion labelling. To avoid fatigue, three short-breaks were scheduled during the labelling session for each participant.

After the labelling session, participants were prompted to fill up an Emotional Quotient (EQ) Test [2]. This is a 40-item self-administered questionnaire used to assess emotional intelligence. The experiment ended with participants filling up a post-experiment questionnaire and providing feedback.

## 4.5. Measures

The following measures were defined and computed:

- The primary label given to the video was considered as the true response emotional label for the presented video.
- The secondary label, when present, was used as an indicator of ambiguity and co-occurrence of emotions.
- The EQ test scores were used as supplementary information to interpret the effect of emotion decoding ability on this task of emotion perception.
- Although, the difficulty level in labelling a video was computed using replay count and decision time, results from this are not presented here.

## 5. Results

Lexical emotion terms often overlap in their meaning therefore, responses for the 'Other' category were post-processed using an emotions taxonomy [1]. For example, emotion terms like puzzled, unsure or baffled were considered to refer to the same emotion as they all belong to the emotion group 'Confused.'

The primary ratings obtained from all the participants after parsing the 'Other' category where present, were compared with the ground-truth labels of the videos. Recognition accuracy of emotions in each of the representations is compared in Table 2 above. Happiness and Surprise show higher overall recognition rates. Note that the chance recognition rate was 20% for all trials.

Furthermore, normality of data was rejected with $p<0.05$ using the Lilliefors goodness-of-fit test for composite normality. We therefore use non-parametric tests to assess statistical significance.

*Recognition across Representations*

The recognition rates were highest for the original videos, followed by stick-figure models, point-light displays, and then XFace animations. A paired comparison of participants' recognition rates for each representation scheme was performed using the non-parametric Wilcoxon Signed Rank Test. Table 3 shows the test results obtained and establishes that the difference in recognition rate across representations is statistically significant.

Table 3: Wilcoxon Signed Rank Test ($\alpha < 0.05$)

| **Wilcoxon Paired Signed Rank Test** | **p** |
|---|---|
| Original, Point-Light | < 0.001 |
| Original, Stick-Figure | < 0.001 |
| Original, XFace | < 0.001 |
| Point-Light, Stick-Figure | < 0.005 |
| Point-Light, XFace | < 0.005 |
| Stick-Figure, XFace | < 0.001 |

*Recognition across Databases*

The effect of database on recognition rates was measured by performing a paired comparison of participants' recognition rates for each database. Table 4 shows the Wilcoxon Signed Rank Test results comparing the recognition rates across the three databases. All values are statistically significant. Participants had the highest recognition rates for the DFAT Cohn-Kanade database followed by the Mind Reading database. Recognition rate was lowest on naturalistic data.

Table 4: Wilcoxon Signed Rank Test ($\alpha < 0.05$)

| **Wilcoxon Paired Signed Rank Test** | **p** |
|---|---|
| Mindreading ,DFAT | < 0.001 |
| Mindreading, Natural | < 0.001 |
| DFAT, Natural | < 0.001 |

*Inter-Rater Reliability*

Since the experiment involves multiple raters rating multiple categories, we compute Fleiss' kappa as a measure of inter-rater reliability or agreement [8]. Table 5 shows the inter-rater agreement kappa values. Except for the original videos, point-light displays and the DFAT Cohn-Kanade database, which show moderate to fair agreements, kappa values indicate very low agreements.

Table 5: Inter-Rater Agreement

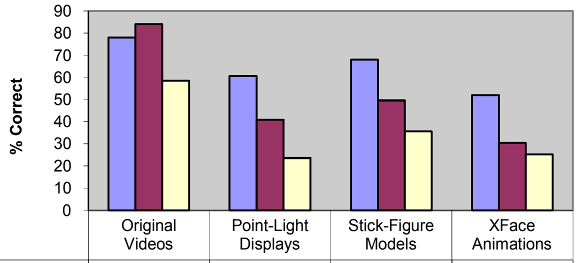
| **Representation** | **Kappa** | **Agreement** |
|---|---|---|
| Original | 0.53 | Moderate |
| Point-Light | 0.16 | Fair |
| Stick-Figure | 0.23 | Fair |
| XFace | 0.15 | Slight |
| **Database** | **Kappa** | **Agreement** |
| DFAT | 0.37 | Moderate |
| MindReading | 0.19 | Fair |
| Natural | 0.07 | Slight |

*Emotional Quotient (EQ) and Recognition Accuracy*

Spearman's rank correlation was computed to test the relationship between EQ scores of participants with their recognition accuracy across the representation schemes or databases sampled. No significant correlation was observed between EQ and recognition accuracy.

## 6. Discussion

The study was designed to investigate how the accuracy of emotion recognition is affected by the nature and representation format of the stimuli. The results provide new insights into perception of emotion from automatically generated facial displays. We find that overall recognition accuracy on videos decreases as we move towards natural data and that this is true for any representation scheme used, so that displays obtained using posed DFAT database are better recognised as compared to those from MindReading or Natural databases. Figure 2 shows the trend in recognition accuracy for each representation across databases used.

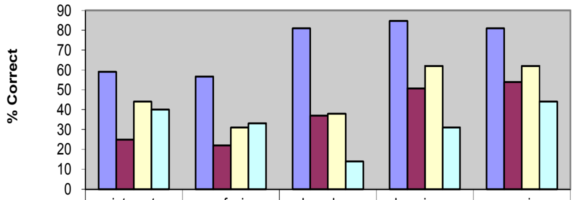


| | Original Videos | Point-Light Displays | Stick-Figure Models | XFace Animations |
|---|---|---|---|---|
| ■DFAT | 78 | 60.7 | 68 | 52 |
| ■MindReading | 84 | 40.8 | 49.6 | 30.4 |
| □Natural | 58.4 | 23.6 | 35.6 | 25.2 |

Figure 2: Recognition accuracy across representations

As expected, original videos show higher recognition rates consistently across representations and databases. Surprisingly however, the stick-figure models show relatively higher level of recognition accuracy compared to both the point-light and 3D XFace animations. This suggests that an intermediate-level of representation, where only outline of facial expressions is provided, affords better perception of emotion. Comparison of inter-rater agreement shows a similar trend (see Table 5) where the kappa value for stick-figure models is marginally higher. A possible explanation is that stick-figure models provide the necessary cues that may then trigger emotion judgements from instinctive mental representations. Any more or less detail in such artificial renderings, as presented though using point-light displays and XFace animations, may be counter-intuitive. If stick-figure models are perceived as better encoders of emotions, then this has implications for synthesis of emotions using computer animations. It is possible that the abstraction level of a stick-figure model allows rendering flaws to be ignored and to focus attention on emotionally salient movements. In contrast, complex models like the 3D XFace animations may enhance flaws in renderings thereby diverting attention to non-significant areas or artefacts.

In fact, we see a similar pattern recurring while comparing recognition of specific emotions across the representations used. See Figure 3 that shows the mean recognition rates for the five emotions in each of the viewed representation scheme. We also find that recognition rates for certain emotions like Happiness and Surprise are consistently higher irrespective of the representation scheme used or the database sampled This implies that the facial feature points commonly employed for emotion recognition using facial expression analysis may not be sufficient to discriminate patterns for all emotions. Emotions like Confusion and Boredom for instance, are accompanied by subtle changes in the face which are not adequately captured by the 24 facial feature points. Except for a few selected works these emotions are in fact rarely addressed in facial expression based emotion recognition [13] and where done, show low classification rates. This does raise an interesting debate on whether there is a limit to emotion recognition using facial feature point tracking even if it is perfected. This suggests that certain emotions are better recognised using feature tracking, while as for others, a hybrid or alternative method (e.g. appearance-based method) may be more effective.



| | interest | confusion | boredom | happiness | surprise |
|---|---|---|---|---|---|
| Original Videos | 59 | 56.7 | 81 | 84.7 | 81 |
| Point-Light Displays | 25 | 22 | 37 | 50.7 | 54 |
| Stick-Figure Models | 44 | 31 | 38 | 62 | 62 |
| Xface Animations | 40 | 33 | 14 | 31 | 44 |

Figure 3: Recognition accuracy across emotion categories

### 6.1. Limitations

The automatic face tracker used in this study tracks only 24 facial landmarks. We acknowledge that this could have affected the recognition accuracy but one of our objectives was indeed to assess how effectively affect-related information was encoded in the facial feature points currently used in facial expression recognition technology.

The number of samples for each category of emotion was small. This was a design constraint as the number of samples to be viewed per participant was already too big (260) and adding more emotion categories would have complicated the experiment. In future, we plan to use the results and observations from this study to repeat the experiment as a between-subjects study with a larger sample size. Also, our ground-truth for videos was based on perfect agreement between three experts which limited our choice of emotion categories to those available in our naturalistic data sources. Consequently, the acted samples (DFAT) lacked boredom and interest.

Finally, although we report significance in recognition rates by performing paired comparisons on representation scheme and database, we would like to observe any interaction effects between the three factors in our design namely, representation scheme, database and emotion. We hope to explore this in future by using multi-way statistical analysis. Future research can also look at how recognition accuracy varies by gender and/or duration of stimuli as well as analysing any systematic confusion that occurs in emotion judgements.

## 7. Summary & Conclusions

The objective of this study was to examine how effectively facial feature points encode emotional expressions. We have presented results from our experiment comparing judgements on five emotions - interest, confusion, boredom, happiness and surprise, in three different representations – point-light displays, stick-figure models and XFace animations, generated from original emotional clips taken from posed and naturalistic databases. Using state-of-art facial feature point tracking, we have attempted to study how affect-related information is compromised when represented in terms of 24 facial feature points. We find that emotion recognition accuracy is higher for original videos, followed by stick-figure models over both point-light displays and XFace animations. We also find that recognition rate and inter-rater agreement decreases as we move from posed data to natural data. Importantly, certain emotions seem to be better discriminable than others irrespective of the representation or nature of stimuli. The results have interesting implications in terms of optimal representation and interplay of facial displays in emotion judgements as well as in analysing the perceptual quality and realism of computer animations.

## 8. Acknowledgements

We acknowledge funding support from the Gates Cambridge Trust and EPSRC, and also thank all our participants.

## 9. References

[1] Baron-Cohen, S., Golan O., Wheelwright S., and Hill J. (2004). Mind Reading: The Interactive Guide to Emotions. Jessica Kingsley Publishers, London

[2] Baron-Cohen, S. & Wheelwright, S. (2004). The Empathy Quotient: An investigation of Adults with Asperger Syndrome or High Functioning Autism, & Normal Sex Differences, Journal of Autism & Developmental Disorders, 34 (2), pp. 163-175

[3] Bassilli, J.N. (1978). Facial Motion in the Perception of Faces & of Emotional Expression, Journal of Experimental Psychology: Human Perception & Performance, 4(3), 373-379

[4] Cohn, J. F. & Schmidt, K. L. (2004). The timing of facial motion in posed & spontaneous smiles. Wavelets, Multiresolution and Information Processing, 2, 1-12.

[5] Darwin, C. (1872). *The Expression of The Emotions in Man and Animals*, London : Murray

[6] Ekman, P. (1982) *Emotion in the Human Face*. Cambridge Univ. Press, Cambridge

[7] Fasel B. & Luettin J. (2003). Automatic Facial Expression Analysis: A Survey. Pattern Recognition, 36(1), 259-275

[8] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psy. Bulletin, 76(5), pp. 378–382

[9] Johansson, G (1973). Visual perception of biological motion & a model for its analysis, Perception & Psychophysics, 14, pp. 201-211

[10] Kanade, T., Cohn, J.F., & Tian, Y. (2000), "Comprehensive Database for Facial Expression Analysis", FG 2000, France.

[11] Pandzic, I.S. & Forchheimer, R. (2002). MPEG-4 Facial Animation: The Standard, Implementation & Applications. Published by John Wiley & Sons.

[12] Thomas, S.M. & Jordan, T.R. (2001). Techniques for the production of point-light & fully illuminated video displays from identical recordings, Behaviour Research Methods, Instruments, & Computers, 33(1), 59-64

[13] Zeng, Z., Pantic, M., Roisman, G.I. & Huang, T.S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, & Spontaneous Expressions, IEEE Trans. PAMI, 31(1), pp. 39-58