Towards a comprehensive 3D dynamic facial expression database

Charlie D. Frowd[1], Bogdan J. Matuszewski[2], Lik-Kwan Shark[2] and Wei Quan[2]
[1]School of Psychology, [2]School of Computing, Engineering and Physical Sciences
University of Central Lancashire
Faculty of Science and Technology, Preston PR1 2HE
UK
[1]cfrowd@uclan.ac.uk www.uclan.ac.uk/scitech/research/rae2008/psychology/staff_profiles/cfrowd.php
[2]bmatuszewski@uclan.ac.uk www.ecson.org/resources/adsip_database.html

Abstract: Human faces play an important role in everyday life, including the expression of person identity, emotion and intentionality, along with a range of biological functions. The human face has also become the subject of considerable research effort, and there has been a shift towards understanding it using stimuli of increasingly more realistic formats. In the current work, we outline progress made in the production of a database of facial expressions in arguably the most realistic format, 3D dynamic. A suitable architecture for capturing such 3D dynamic image sequences is described and then used to record seven expressions (fear, disgust, anger, happiness, surprise, sadness and pain) by 10 actors at 3 levels of intensity (mild, normal and extreme). We also present details of a psychological experiment that was used to formally evaluate the accuracy of the expressions in a 2D dynamic format. The result is an initial, validated database for researchers and practitioners. The goal is to scale up the work with more actors and expression types.

Key-Words: Emotion, Capture, 3D dynamic, Animation, Human face, Actor, Experiment, Statistical models.

1 Introduction

Human faces have evolved into an impressive range of functions. These include the capture of sensory information, the projection of person identity, the expression of emotion and intention, and for biological functions involved with speech and respiration. The human face itself has increasingly become the subject of psychological experimentation and theorizing, arguably triggered by Charles Darwin in his thesis on emotion in animals and man [1]. During the past 30 years or so, an intense research exercise has been taking place to both confirm and extend Darwin's observations.

The interest in this area is such that psychological research programmes now exist in a broad range of overlapping disciplines. Examples include how we understand, recognize and learn identities from faces, interpret emotion, use gaze cues to interpret intentionality, and to describe, construct and identify images for forensic applications. These projects themselves have included a wide range of subject age groups, and individuals from different backgrounds and cultures.

It is evident that there has been a shift towards the use of more realistic face stimuli in research projects. Early work tended to focus on outline drawings [2], static photographs [3][4][5][6] and even moving clips [7][8]. For example, in the seminal work by Ekman [3], static photographs of faces were taken of people producing one of six facial expressions: happiness, sadness, anger, fear, disgust and surprise. The production of each expression was noted to involve a distinct combination of movements in three facial 'components', and also that these expressions were universally recognised across human cultures. A method of describing the expressions was provided: the Facial Action Coding System, FACS. (A good overview of FACS may be found here: [9]).

We now know that static stimuli may not be the most appropriate format for experimentation. For example, there are robust effects of motion in facilitating the perception of subtle facial expressions depicting the above six emotions [10]. Using sophisticated video alignment and editing techniques, recent work has also started to unpack the importance of the various facial areas involved in emotional expression [8][11][12]. It is clear that while some emotions can be recognised very well statically from the face alone, for example happiness, others require extra facial information. For instance, both the expressions of agreement and disagreement require so-called 'rigid head motion' (RHM) – an up-down or left-right head shake, respectively. Others, such as fear, 'don't know' (clueless) and 'don't understand' (confusion) normally involve RHM to some extent, and the perception of the expression is improved when head movements are observed – e.g. [11][12][13]. In fact, dynamic information can be particularly effective in situations where a face is observed under less than optimal conditions such as poor lighting or from a distance [14]. There is even evidence for separate neural cognitive mechanisms for certain types of face processing tasks carried out statically or dynamically [15].

The current project aims to extend the above work by providing a database of dynamic facial expressions in 3D. One benefit of this will be to allow research to be conducted with stimuli that can be viewed from different angles, and in 3D, through the use of 3D glasses, 3D displays, etc. It is clear that improvements to the perception of expressions occur for 2D dynamic relative to 2D static stimuli, and so a progression to 3D dynamic should provide further benefit. Specifically, the ability to extract cues that are ambiguous or difficult to perceive are likely to be particularly improved in 3D, especially in those expressions that involve a depth component – e.g. fear, surprise, confusion.

There are other benefits to building a 3D dynamic database of expressions. Firstly, in the animation industry, one approach to produce animated characters is the construction of statistical models of facial identity and/or emotion – e.g. [16]. Such models require accurate and believable expressions, of the type intended to be captured here. Secondly, in the case of plastic surgery, or intervention therapies by speech therapists, accurate models of facial expression could be both insightful and valuable [17]. Thirdly, composite faces of criminals constructed by eyewitnesses traditionally lack emotion (see [18] for a review), and so the construction of emotional cues may help to trigger cues to identity when published in the media (this notion is based on a potential overlap of information processing between a person's facial identity and their expression [7][19]; it is also clear that at least one type of animated sequence is valuable [21]). Fourthly, there are potentially valuable applications in the human-computer interaction (HCI) area for the accurate perception of facial emotions [13]. Fifthly, there has been some research into the automated recognition of pain expressions, which if ultimately successful, would be very useful in medical environments [17][20]. There are other applications in the security industry and biometrics.

The main objective of the project then is to build a database of 3D facial sequences showing different articulations of expressions. Our intention is to do this initially for a relatively small number of human participants. It will contain a range of expressions, including Ekman's 'universal' six, and will be made available to researchers and practitioners.

In the follow sections, we describe an image capture system capable of good quality dynamic expression capture in 3D. This is followed by the procedures used to capture and then validate the expressions. Finally, as the project is ongoing, details of future plans are discussed.

2 Image capture and 3D viewing

We used the 3dMD dynamic scanner to capture our 3D facial sequences and the JVC GZ-MG142 Everio camcorder to capture 2D texture videos. The 3D scanner contains six digital cameras and two infrared speckle projectors that are positioned on two sides – i.e., one projector and three cameras on each side, as shown in Fig. 1. During the data acquisition process, the scanner projects a random speckle pattern onto a human face and captures a set of 2D images from different angles. Four of these images are used to reconstruct 3D geometry whereas the other two provide textural information for accurate 3D face rendering. Subsequently, it automatically merges images from all cameras and produces a 3D polygon surface mesh of the face with mapped texture. The system is able to capture 600 seconds of dynamic 3D at an acquisition speed of up to 48 frames per second. In the current application, the frame rate was set to 24.

Fig. 1. A photograph of the 3dMD dynamic 3D scanner.

A head-mounted display (HMD) is proposed to be used for viewing the 3D facial sequences. It is a device that is worn on the head and contains a small display in front of each eye. It is able to provide a stereoscopic view of 3D facial sequences to allow participants to observe the dynamic 3D facial expressions from different angles by changing head position. An example HMD is presented in Fig. 2

The room used for photography had the approximate size of 10×3 square meters. The windows of the room were covered in order to block out daylight, which would otherwise have caused lighting artifacts as well as affect the accuracy of the 3D face reconstruction. The lighting system itself used for the data acquisition consisted of four cool-light lamps (two on each side). In the acquisition stage, participants sat in a chair set to at height of about 0.5 meters. The distance from the image capture system to the participant was about 1 meter. The background to the participant was a plain, white-painted wall.

Fig. 2. An example of a head-mounted display

3 Expression capture and processing

Our objective was to capture and evaluate a range of facial expressions using a 3D image capture system. We noted four main issues: the production of accurate expressions, the general process involved in image capture, expression evaluation, and the processing of 3D animated sequences. The first two of these issues are described in the next sections; the validation (analysis) of the expressions is presented in section 4; and the processing of the 3D sequences is described in section 3.4.

3.1 Design

Some expressions can be easily produced on demand, such as happiness, sadness and surprise, while others are harder, fear and pain. For this initial work, we recruited the services of trainee actors, people who have taught themselves in the production of emotional responses. This is an established technique – e.g. [11] – that should enable fairly accurate expressions to be captured.

There are still likely to be individual differences in the accuracy of expression production, even from actors. We therefore recruited 10 actors, as many as was possible within the available timeframe.

The initial data set aimed to capture the most useful expressions. These included the six 'universal' ones, as noted by Ekman and Friesen [3] – happiness, sadness, anger, fear, disgust and surprise – plus one more, pain. To provide useful variations, these seven expressions were captured at three levels of intensity from each actor, 'mild', 'normal' and 'extreme'.

3.2 Participants

The actors who provided the expressions were final year students studying drama at the University of Central Lancashire, Preston, UK. There were 8 females and 2 males and all were in their early twenties.

3.3 Procedure

Actors were informed in advance of the seven expressions that were required (to allow them to practice, if necessary). During the session, actors were tested individually. They were given an initial brief as to the aim of the project and signed a consent form to allow their footage to be processed, stored, analyzed for accuracy by further participants, and distributed for research and non-commercial purposes.

Each person was asked to produce seven expressions at three levels of intensity. The Experimenter mentioned what the first expression was (anger) and the desired intensity level ('mild'), and the actor indicated when s/he was ready. The Experimenter counted down from 'three' to 'one' and triggered the 3D sequence capture. This was repeated for the other two levels ('normal' and 'extreme') and for the seven expressions. The order of expression production was carried out in the following fixed order for each person: anger, disgust, fear, happiness, sadness, surprise and pain. The 'mild' expression was requested first in each case, followed by 'normal' and then 'extreme'. A total of 21 sequences (7 expressions x 3 levels) were collected from each actor.

We note that there was a three to four minute delay after each sequence, the time required for the data to be downloaded from the camera to the PC. In addition, a couple of the actors kindly repeated expression production a couple of days later; these data were not analyzed, but parked for later work. Sessions took about two hours to complete.

Example expressions can be seen in Fig. 3.

3.4 Data processing

After data acquisition, the 2D texture videos were edited using Cyberlink's PowerDirector Express commercial software [22], and made available for expression validation (refer to the next section). The 3D facial sequences are in the process of being cleaned, to fill in 'holes' and to remove unwanted parts in the scans. The cleaning of the 3D facial scan can be done manually using the 3dMD patient software; however, as there are a large number of scans, this is a time-consuming process and so a bespoke program is under development to assist.

4 Validating expression sequences

We were interested in the general processes of expression capture and analysis, and so while both 2D (camcorder) and 3D (3dMD) expression information were captured, only information from the 2D camera was formally evaluated, see below. Later work will consider replicating and extending this analysis in 3D: refer to section 5, below.

4.1 Design

A bespoke computer program was written to present the 2D video clips to participants and to collect their responses. The program presented the clips in a random sequence for each person, to avoid *order* or *fatigue* effects that might bias the data. As there were a fairly large number of sequences in total (210), these were split into two manageable sets and participants looked at one of these.

The program played the clips sequentially alongside a windows dialog with entry boxes for each of the seven expressions. These boxes allowed participants to rate their level of confidence for each expression for each clip. While more than one rating box could be used each time, to allow a person to indicate ambiguity, values were required to sum to 100%. To allow ease of entry, as many clips were assumed to have clearly defined expressions, initial values were set to 0%. This method of data collection provided a powerful within-subjects experimental design for ratings of expression type (anger / disgust / fear / happiness / pain / sadness / surprise) and intensity (mild / normal / extreme).

4.2 Participants

The participants who provided ratings of confidence were 10 staff and students at the University of Central Lancashire, Preston, UK. They comprised 6 females and 4 males (M = 34.8 years, SD = 11.2 years). Each person was given an honorarium of £2.

Fig. 3. Expression production from one of our actors at three intensity levels (01 = 'mild' / 02 = 'normal' / 03 = 'extreme'). Shown are single frames taken from the 2D camcorder at roughly the peak of the expression. Note that, although greyscale images are shown here, recordings were made and are available in colour.

4.3 Materials
Video clips from the 2D camera were trimmed, to remove excess material, and cataloged for presentation to participants. The clips were converted to AVI format at a resolution of 640x480 pixels. A good quality computer monitor was used for presentation; the displayed head size of the actors was approximately 4cm x 4cm. Each sequence lasted about 3 seconds.

4.4 Procedure
Participants were tested individually. The Experimenter told them that they would be shown a set of video clips of actors producing facial expressions and their task was to identify the expressions depicted, by assigning confidence ratings.

A bespoke computer program was used to present the video clips and collect the confidence ratings. Participants were shown an example form that would be used to enter the ratings, and that this contained entry boxes for the seven expressions to be presented – anger, disgust, fear, happiness, pain, sadness and surprise. It was mentioned that each box would accept values in the range of 0 to 100% but, as some expressions would be unambiguous, it was only necessary to enter a single value in the appropriate box for these items. It was also mentioned that, for a given video clip, ratings could be distributed over the various entry boxes so long as scores summed to 100%. Participants were told that the form would provide feedback about this and would only move onto the next clip when this constrained had been satisfied. Note that the form was populated (initialized) with zeros when each face was presented, to facilitate data entry.

Ratings were collected from a total of 210 clips. These comprised of the 10 different actors each person producing the above seven expressions at three levels of intensity (mild, normal, extreme); 10 * 7 * 3 = 210. To make the task manageable for participants here, the clips were split into two equal image sets (of 105 sequences) with each containing all levels from a given actor. Participants were randomly assigned to image set 1 or 2, with equal sampling (i.e. 5 people inspected set 1, and the other 5 people set 2). Thus, when participants were ready, the 105 video clips were presented sequentially in a different random order for each person and participants rated each in turn as instructed. The task was self paced and a break was offered after about 15 minutes. The entire procedure took roughly 30 minutes to complete. At the end, each person was debriefed as to the nature of the study.

4.5 Results
The dataset contains five repeated confidence ratings for each of the 210 video clips, a total of 1050 (210 clips * 5 repeats) individual responses (values between 0 and 100%). Confidence scores had a grand mean of 65.0%; by actor, the mean ranged from 54.7% to 79.7%.

A range of analyses is possible on the data set: overall accuracy, accuracy by actor, accuracy by expression, by level, confusion of expressions, and so forth. For brevity, we present an overall analysis here, by expression and intensity level; these relate to confidence scores given for the 'correct' or intended expression. Individual and group confidence data may be obtained online – see 4.5.2 for details.

4.5.1 Confidence scores
Confidence scores were extracted for the 'correct' expression in each video clip. These data are presented in Table 1. It can be seen that happiness expressions were given near perfect confidence scores, and anger, pain and fear were the worst rated at around 50%. Also, the 'normal' intensity level was somewhat better rated than 'mild', and 'extreme' was also somewhat better than 'normal'.

Table 1. Mean confidence scores for the seven expressions (values in percent).

Intensity	anger	disgust	fear	happiness	pain	sadness	surpise	Mean
mild	47.5	51.5	43.3	90.3	42.6	72.9	57.4	57.9
normal	56.6	78.3	41.5	94.3	51.4	75.6	62.0	65.7
extreme	61.4	80.7	48.4	96.0	56.2	74.0	75.7	70.3
Mean	55.2	70.2	44.4	93.5	50.0	74.2	65.0	64.6

The participant rating data were analyzed by a repeated-measures Analysis of Variance (ANOVA). This was significant for expression type, $F(6,54) = 15.9$, $p < 0.001$, $\eta^2 = 0.64$, and intensity level, $F(1,18) = 12.6$, $p < .001$, $\eta^2 = 0.58$; the interaction between these factors was not significant, $F(12,108) = 1.5$, $p = 0.154$. Analysis of expression type using t-tests with Bonferroni correction (which is necessary due to the large number of contrasts being carried out) is displayed in Table 2. Simple contrasts of the ANOVA for intensity level (one-tailed) indicated that 'normal' expressions were given significantly higher confidence scores that 'mild' ones, $p < 0.001$, with a very large effect size, Cohen's $d = 1.99$, and that there was an approaching significant benefit of 'extreme' over 'normal', $p = 0.083$, $d = 0.79$.

To summarize. Firstly, it can be seen that there are fairly large differences by actor, which is normal – e.g. [11]. Secondly, there are differences in confidence scores between the various expressions, with happiness being produced the best across all actors, and fear being the worst. These are typical results found elsewhere – e.g. [3][8][11]. Thirdly, based on the assumption that higher confidence scores indicate a more accurate identification of expression, mild expressions appear to be the hardest to classify, and extreme the easiest, a sensible outcome. Note that confidence scores increased reliably from mild to normal, but less so from normal to extreme, where the benefit was weaker (a marginally significant improvement).

Table 2. Post hoc analysis using t-tests for the seven expressions (uncorrected probability values)

	anger	disgust	fear	happiness	pain	sadness	surpise
anger	-	0.038	0.042	0.000 *	0.519	0.001 *	0.161
disgust		-	0.010	0.000 *	0.001 *	0.378	0.271
fear			-	0.000 *	0.508	0.001 *	0.044
happiness				-	0.000 *	0.001 *	0.000 *
pain					-	0.001 *	0.065
sadness						-	0.166
surpise							-

*Statistically significant contrast (note that with a Bonferroni correction applied, the significance level of 0.05 should be considered as 0.05/20, or 0.005).

4.5.2 Availability of confidence and image data

A CD containing 2D animated clips and confidence scores for dataset validation may be obtained at www.ecson.org/resources/adsip_database.html. The 3D raw data will be available online this summer, and processed images as soon as possible thereafter.

5 Discussion and future work

The current project is the first step to produce a comprehensive 3D database of facial expressions. We recruited the services 10 actors and actresses to produce seven expressions at three levels of 'intensity'; these were captured using a 3D dynamic system and a 2D camcorder. We also recruited a group of observers to formally analyze the accuracy of the 2D expressions. They did this by providing confidence ratings to each of the video sequences. Analysis of these data suggested that the general approach used was sensible. Both participant ratings and video sequences are now available (see 4.5.2).

We are currently processing the 3D image files, to ensure image consistency (e.g. making sure that there are no 'gaps' remaining in the sequences). One of the next steps is to render these on a 3D viewer, as mentioned in section 2, and then to repeat the database validation – although this can clearly be done by other researchers, which we encourage.

We think the most useful evaluation would be a comparison using four image modes. The first would be a static frame taken from the 2D camera at the 'peak' of the expression (similar to those shown in Fig. 3). This is perhaps the easiest and earliest method [3]. The second would be the same as that carried out in section 4, 2D sequences (camcorder). The third would be dynamic 3D. We predict that there will be a fairly large increase in the accuracy from 2D static to 2D dynamic, as found before, e.g. [10], and a further, albeit less dramatic improvement from 2D dynamic to 3D dynamic. It is perhaps prudent though to include a fourth condition, 3D static, as a comparison to 2D static; one would sensibly predict equivalence between these two conditions. (3D static information is unlikely to be more helpful than that of 2D static for expression perception, but this needs confirming.)

A sensible next step would also be to expand the number of individuals in the database. It is clear that the use of actors, or trainee actors, is a valid approach. While it would appear advantageous to recruit members of the public, and in more natural settings, there are problems – e.g. [11]. Firstly, it is difficult to provide a range of stimuli that would naturally and reliably evoke a range of expressions. Even if this were possible, it would be unethical to try to elicit some emotions: fear, disgust and pain. Also, there is the general problem of knowing exactly which emotion a person was experiencing. Therefore, the use of actors is better. Nevertheless, we do acknowledge that there are certain applications where at least some naturally-produced expressions would be valuable: HCI emotion recognition, as mentioned above (e.g. [12]). However, these expressions might be only used in 2D rather than 3D dynamic, and so the former image mode may be sufficient.

As well as continuing our work on a facial expression recognition system [23], we would also consider revising the expressions captured. It would be advantageous to be more specific about some. For example, surprise clearly comes in two forms – pleasant and unpleasant – and it would be useful to differentiate these (and potentially capture both types). We would, though, consider increasing the number of different expressions. Potential candidates include agreement and disagreement, potentially for the film and animation industry, and confusion and clueless, for psychological research. Suggestions from other researchers are welcome.

ISSN: 1790-5109

ISBN: 978-960-474-114-4

6 Conclusion

This paper outlines the initial work done to produce a validated database of dynamic facial expressions in 3D. We have discussed the general methodology used to capture and process seven expressions, and provided and tested a methodology for validating them from the 2D dynamic footage. Confidence ratings and 2D dynamic data are now available for use by other researchers and practitioners; 3D expression data will be available this summer. We plan to validate the 3D dynamic information from the sequences collected so far and expand the number of actors and expressions.

References:

[1] Darwin, C. *The expression of emotion in man and animals*. New York: D. Appleton and company. 1872.

[2] Benson, P.J., & Perrett, D.I. Visual processing of facial distinctiveness. *Perception*, 23, 1994, pp. 75-93.

[3] Ekman P., & Friesen W.V., Measuring facial movement. *Environmental Psychology and Non-verbal Behavior*, 1, 1976, pp. 56-75.

[4] Bruce, V. Changing faces: Visual and non-visual coding processes in face recognition, *British Journal of Psychology*, 73, 1982, pp. 105-116.

[5] Langton, S.R.H., & Bruce, V. You Must See the Point: Automatic Processing of Cues to the Direction of Social Attention. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 2000, pp. 747-757.

[6] Frowd, C.D., Hancock, P.J.B., & Carson, D. EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Psychology*, 1, 2004, pp. 1-21.

[7] Lander, K., Christie, F., & Bruce, V. The role of movement in the recognition of famous faces. *Memory and Cognition*, 27, 1999, pp. 974-985.

[8] Cunningham, D., Kleiner, M., Wallraven, C., & Bulthoff, H. Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception*, 2, 2005, pp. 251–269.

[9] Sayette, M.A. Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression. *Journal of Nonverbal Behavior*, 25, 2001, pp. 167-85.

[10] Ambadar, Z., & Cohn, J.F. Deciphering the Enigmatic Face: The Importance of Facial Dynamics in Interpreting Subtle Facial Expressions. *Psychological Science*, 16, 2005, pp. 403-410.

[11] Nusseck, M., Cunningham, D. W., Wallraven, C., & Bülthoff, H.H. The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision*, 8, 2008, pp. 1-23.

[12] Busso, C., Grimm, M., & Narayanan, S. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15, 2007, pp. 1075-1086.

[13] Wang, T.H., & Lien, J.J.J. Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation. *Pattern Recognition*, 42, 2009, pp. 962-977.

[14] O'Toole, A.J., Roark, D.A., & Abdi, H. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, 6, 2002, pp. 261-266.

[15] Steede, L.L., Tree, J.J., & Hole, G.J. I can't recognize your face but I can recognize its movement. *Cognitive Neuropsychology*, 24, 2007, pp. 451-466.

[16] Sloan, R., Robinson, B., and Cook, M. Animating believable facial expressions: Is it possible to choreograph perceptually valid emotional expressions? 2009, a3.vox.com/6a00c2251df2578e1d011016165813860b-pi.

[17] Brahnam, S., Chuang, C., Shih, F. Y., Slack, M. R., Machine recognition and representation of neonatal facial display of acute pain. *Artificial Intelligence in Medicine*, 36, 2006, pp. 211-222.

[18] Frowd, C.D., Bruce, V., & Hancock, P.J.B. Changing the face of criminal identification. *The Psychologist*, 21, 2008, pp. 670-672.

[19] Calder, A.J. & Young, A. Understanding the recognition of facial identity and facial expression. *Nature Reviews: Neuroscience*, 6, 2005, pp. 641- 651.

[20] Hancock, P.J.B., Frowd, C.D., & Brodie, E. Niven, C.A. Recognition of Pain Expressions. In A. Cangelosi, G. Bugmann & R. Borisyuk (Eds) *Modelling Language and Action* (pp. 339-348) Singapore: World Scientific Publishing. 2005.

[21] Frowd, C.D., Bruce, V., Ross, D., McIntyre, A., & Hancock, P.J.B. An application of caricature: how to improve the recognition of facial composites. *Visual Cognition*, 15, 2007, pp. 1-31.

[22] www2.cli.co.jp/products/ne/pdirex_eng.htm

[23] Quan, W., Matuszewski, B.J., Shark, L.K., & Ait-Boudaoud, D. Low Dimensional Surface Parameterisation with Applications in Biometrics. *International Conference on Medical Information Visualisation - BioMedical Visualisation (MediVis 2007)*, 2007, pp.15-22.