Automatic Extraction of Facial Action Codes

M. Frank, P. Perona, Y. Yacoob*

Executive Summary

Two systems for capturing and analyzing automatically facial expressions are evaluated and compared. The systems, developed by interdisciplinary research teams at CMU/Pitt and UCSD/Salk, are the first attempt at automatically analyzing spontaneous facial expressions with unconstrained head orientation. The training and test sequences where acquired by Ekman and Frank while the subject was engaged in deception in a realistic scenario. Facial analysis is based on the extraction of action units as defined by Ekman in his classic Facial Action Coding System. While full automation has not yet been achieved the systems correctly classified around 90% of instances of 3 facial movements, or Action Units. The objective of a fully automated system for measuring human expression from facial and body motion and, possibly, speech cues, is a realistic goal for the next 10 years. Applications in intelligence and security, human-machine interfaces, and psychology will be enabled by this technology. The next step to be taken in order to achieve this goal is the acquisition of a 10x larger and better (more sensors and higher resolution) dataset. This should be followed by the creation of a community of interdisciplinary teams working towards this goal.

Introduction

Designing and building machines that can interact with people with the same facility with which people interact with each other is one of the greatest challenges of modern engineering. In order to achieve this goal both computational and psychological research must progress in modeling and analyzing human communication modalities by considering sensory and cognitive systems such as vision, audition and language.

Modeling and analyzing human facial expressions is a vital component of this effort since the most informative window into human emotion is the human face. The pioneering research of Paul Ekman, who codified the action units (AUs) of facial expressions in his Facial Action Coding System (FACS), offers a convenient starting point in this research. This report analyzes and compares two approaches to quantitatively measure facial action units on a video database of spontaneous face expressions.

Facial actions and emotion

Research going back to Darwin (1998/1872) recognized that human emotion is expressed, as well as most easily recognized, through facial expressions (e.g., Ekman, 1994; Izard, 1994). Moreover, there appears to be a specific set of facial expressions generated by the emotions of anger, disgust, fear, happy, sad, surprise, and to a lesser degree contempt, embarrassment, interest, pain, and shame. These emotions are universally generated and recognized across all cultures (Ekman, 1972). They are unbidden, with a particular pattern of morphology and dynamic actions (Ekman & Friesen, 1982; Frank & Ekman, 1993). A number of studies have since documented the relationship between these facial expressions of emotion and

* Authorship is alphabetical. Authors' contact information: Mark Frank, SCILS, Rutgers University, 4 Huntington St., New Brunswick, NJ 08901, mgfrank@scils.rutgers.edu; Pietro Perona, Department of Electrical Engineering, Caltech 136-93, Pasadena, CA 91125, perona@caltech.edu; Yaser Yacoob, UMIACS, University of Maryland, College Park, MD, 20742, yaser@umiacs.umd.edu.

the physiology of the emotional response. For example, researchers have found a significant relationship between facial expressions of emotion and a) the self-report of various emotional experiences (Ekman, Friesen, & Ancoli, 1980); b) unique autonomic nervous system (ANS) physiological profiles in American actors, the elderly, and the Mnangkabau - a matrilineal, Indonesian society (Ekman, Levenson, & Friesen, 1983; Levenson, Ekman, & Friesen, 1990; and Levenson, Ekman, Heider, & Friesen, 1992); and c) specific Central Nervous System (CNS) patterns of hemispheric brain activation, as measured by the electroencephalogram (EEG; Davidson, 1984; 1992). Thus, facial expressions seem to be the most visible and reliable clues to the presence or absence of human emotional response.

Applications

These universal expressions of emotion have many real world applications. First, they have shown utility as markers of various states of social functioning. For example, research has shown that only one type of smile accompanies the experience of positive emotion – the enjoyment smile (Ekman & Friesen, 1982; Frank & Ekman, 1993). This enjoyment smile features not only zygomatic major action (moving the lip corners upward), but also orbicularis oculi action (producing crow's feet around the eyes). The presence of these enjoyment smiles on the part of a person who has survived the death of their romantic partner predicts successful coping with that traumatic loss (Bonnano & Keltner, 1997). Infants show enjoyment smiles to the presence of their mothers, but not to strangers (Fox & Davidson, 1988). Mothers do not show as many enjoyment smiles to their difficult compared to their non-difficult children (Bugental, 1986). Research based upon FACS has shown that facial expressions have similar utility in predicting the onset and remission of depression, schizophrenia, and other psychopathology (Ekman & Rosenberg, 1997). Even in the domain of relationships, the facial expressions of disgust or contempt, but not anger, predicts marital divorce (Gottman, 1994).

Second, interpersonal deception research has begun to note that these unbidden facial expressions of emotion – typically fear and distress, but also disgust, contempt, or even enjoyment - can occur for very brief flashes, called "microexpressions", that under certain circumstances can betray deception (Ekman, 1985; Ekman, O'Sullivan, Friesen, & Scherer, 1991; Frank & Ekman, 1997).

These facial expressions will have utility in any situation in which it is important to ascertain a person's emotional state. This is useful for any health/psychological care worker, police officer, intelligence or counter intelligence officer, customs inspector, security personnel, judge, and even business negotiator, job interviewer, insurance investigator, and so forth. As much as the perception of emotion is key to successful human interaction, an automatic evaluation of human emotion is key to building a new generation of human-machine interfaces that is more pleasant, convenient and powerful. This booming area of research and technology will also greatly benefit from techniques to estimate automatically the emotional state of humans from facial expressions.

The Facial Action Coding System (FACS) – a brief history

The Facial Action Coding System (FACS), is a technique developed by Paul Ekman and Wallace Friesen in the mid 1970's (Ekman & Friesen, 1978). FACS is a comprehensive system that measures all visible facial muscle movements, and not just those presumed to be related to emotion. FACS also scores head and eye movements. When learning FACS, a coder learns the characteristic pattern of bulges, wrinkles, and movements for each facial Action Unit (AU). Ekman and Friesen determined these through close observation of others, and by inserting electrodes into their own faces to stimulate individual muscles. These AUs approximate individual facial muscle movements, but are not necessarily so. For example, the frontalis muscle (which covers the forehead) does not always move as one unit. When it moves as one unit, it gives the appearance of both eyebrows moving upward. However, the medial portion of the frontalis muscle can move independently from the lateral portion of the muscle – and gives the appearance of just the inner corners of the eyebrows raising upward. This particular facial action is often found in the emotion of sadness/distress. Thus, the medial and lateral portions of the frontalis muscle are scored as separate AUs

(AU 1 for medial, AU 2 for lateral). One would look for raised inner corner of eyebrows, and horizontal wrinkles in the middle of the forehead, to score this medial frontalis action. Likewise, some facial actions always co-occur, as in the 3 muscles that produce the nose wrinkle (AU 9).

FACS also scores the intensity of each facial action, on an A to E scale. An A intensity score means that not all the specific criteria for an AU is present, but the coder sees some movement; B to E represent an ordinal grading of intensity where all the criteria for scoring the AU is present, with E representing the maximum possible movement of a particular AU.

It takes approximately 100 hours for a coder to learn FACS, and each coder must take a final test to insure the reliability of his or her codes (at rates equal or greater than 80% agreement). Coders are permitted to score videotapes with a listing of all AUs present (i.e., it is always an "open book" type test or scoring). FACS requires slow motion, back and forth examination, of particular facial landmarks in order to infer the action of these subcutaneous facial muscles. To score all possible head, eye, and facial movements, it can take up to one or even two hours to score one minute of behavior. This high ratio of effort to outcome has always been one of the largest impediments to research in emotion and the use of FACS in general. However, FACS has the advantage of being a non-obtrusive way to measure emotion. It has been used to verify the physiological presence of emotion in a number of studies, with high (over 75%) reliability (e.g., Ekman, Friesen, & Ancoli, 1980; Ekman, Levenson, & Friesen, 1983; Ekman, Davidson, & Friesen, 1990; Levenson, Ekman, & Friesen, 1990; Ekman, Friesen, & O'Sullivan, 1988). Because it is comprehensive, FACS also allows for the discovery of new patterns related to emotional or situational states. For example, using FACS Ekman et al (1990) and Davidson et al (1990) found that smiles that featured both orbicularis oculi (AU6), as well as zygomatic major action (AU12), were correlated with self-reports of enjoyment, as well as different patterns of brain activity, than were smiles that featured only zygomatic major (AU12). FACS has also been used to examine for discrete negative emotions (e.g., see Ekman, Friesen, and Ancoli, 1980), and can be used to identify disgust, fear, and embarrassment (Keltner, 1995). FACS has also been able to identify patterns of facial activity involved in alcohol intoxication that observers not trained in FACS failed to note (Sayette, Smith, Breiner, & Wilson, 1992). Finally, it has discovered various patterns reliably related to deception (Ekman et. al., 1988; Ekman, O'Sullivan, Friesen, & Scherer, 1991; Frank & Ekman, 1997).

State of the art of computational facial analysis

Interest in the computational interpretation of human appearance and actions has increased significantly in the last decade. Recent research can be broadly divided into analysis of static and dynamic aspects of human appearance. Static analysis focuses on the enduring aspects of human appearance such as identity, age, gender, height, weight and wearables (dress style, eyeglasses, beards, etc.), while dynamic analysis focuses on the transient or instantaneous aspects such as facial expression, gaze, posture and body movement. Early research has focused on face detection, tracking and recognition and is primarily motivated by applications of surveillance and access control. This research evolved into estimation and interpretation of facial dynamics and full body movements. While significant progress has been made in detection of faces and face-features in a scene the automatic analysis of the dynamics of faces is still in its infancy. We give a brief review here.
Research areas on facial and head dynamics are:

1. Facial actions and expressions (see [Pantic et al., 2000] for a review). Dynamic face analysis can be coarsely divided into several stages: detection/initialization, tracking and spatio-temporal classification.

   Detection involves locating the image region that belongs to the face and labeling image regions that identify face features. These features can correspond to what human may perceive as features such as mouth, eyes, eyebrows, nose, cheeks etc. or computational features defined by image properties (Gabor-based features, edges, principal components, etc.). Detection can be sought in 2D or 3D space. Significant progress has been made in detection using cues such as skin color, motion, learned templates, etc. The difficulty of the problem depends on the quality of the data and the amount of information that is known a-priori. It is fairly well-solved in the case of frontal views of the face in standard lighting against a stationary background. The problem remains

challenging when the number of people in the scene is large, motion clutter is present, or the faces take up a small part of the image.

Tracking involves both face tracking and face feature tracking between consecutive frames. It can be accomplished be repeating the detection process on each frame and thus ignoring the temporal continuity of image information or by employing motion models that represent the changes that the face and face features can undergo between frames. The latter approach usually assumes that the overall face is moving rigidly and the face feature movements are non-rigid.

Temporal classification employs the parameters that reflect the face and features states at each frame as measured during tracking to derive abstracted states such as action unit activation or face expressions. While FACS is used to capture the apex of expression, classification generates an on-line output that reflects the abstracted states reflecting instantaneous and past states. The above three stages have been approached in diverse ways by researchers. Depending on assumptions and design choices sub-tasks arise and configurations change.

Estimation of facial deformation has been pursued for single images [Donato et al., 1999, Lanitis, 1997, Lyons, 1999] or image sequences [Black & Yacoob, 1997, DeCarlo & Metaxas, 2000, Essa & Pentland, 1997, Mase, 1991, Tian et al., 2001, Terzopoulos & Waters, 1991]. The former posed a pattern classification problem in which the appearance of the face is analyzed. Template-based [Edwards, 1998] or feature-based approaches [Donato et al., 1999] have been employed to maximize the differentiation between face states. The latter approaches generally extract cues to facial deformation by employing models of non-rigid motion over the whole face or parts of it and then perform pattern classification on the temporal parametric representation. Template [Black & Yacoob, 1997, Mase, 1991], 3D-model [DeCarlo & Metaxas, 2000], and feature-based approaches [Tian et al., 2001, Lanitis et al., 1997] have been used to capture the characteristics of facial movement.

Comparative performance of feature extraction techniques that use data-driven spatial filters and temporal models was shown in Donato et al. 1999. Specifically, a comparison between feature extraction approaches based on Principal Component Analysis, Independent Component Analysis, Gabor Wavelets and optical flow was conducted. While some researchers focused on estimating the facial movements during expressions (i.e., tracking) [DeCarlo & Metaxas, 2000, Terzopoulos & Waters, 2000] others also addressed the interpretation problem by generally focusing on the six basic expressions [Black & Yacoob, 1997, Essa & Pentland, 1997, Mase, 1991].

Interpretation of facial expressions Most approaches that sought interpretation of facial expressions used FACS implicitly or explicitly. Implicit use generally translated the Action Units movements of each expression into the parameters of the chosen estimation model [Black & Yacoob, 1997,Essa & Pentland, 1997, Lanitis, 1997, Mase, 1991, Yacoob & Davis, 1997]. Explicit use involved estimating the Action Units movements' as defined by FACS [Donato et al., 1999, Pantic et al., 2000, Tian et al, 2001] and employing these activations for interpretation. The reported classification performance was high; however, experiments were mostly carried out on small databases, focused on isolated expressions (of the primary six) and mostly used staged face-expressions. The classification approaches included rule-based approaches [Black & Yacoob, 1997], discriminant function [Cohn et al., 1998], nearest neighbor [Essa & Pentland,, 1997] and neural networks [Padgett & Cottrell, 1996, Rosenblum et al., 1996].

2. Eye-gaze estimation. Accurate gaze estimation of unconstrained subjects using environmental cameras has received only little attention. Infrared illumination is often used [Sheng-Wen et al., 2000, Morimoto et al., 2000] to enhance iris detection. Also, multiple cameras are used to derive 3D information [Sheng-Wen et al., 2000, Matsumoto & Zelinsky, 2000]. An appearance-based approach has been shown to provide qualitative estimates [Pappu & Beardsley, 1998].

3. Head movements and gestures. While estimation of head movement has been accomplished using diverse approaches [Black & Yacoob, 1997, Essa & Pentland, 1997, DeCarlo & Metaxas, 2000] there has been remarkably little research into interpretation of head movement. Moreover, its use in expression and communication has received little attention. Both 2D and 3D approaches have been employed successfully for head tracking and can be used for interpretation.

4. Speech understanding by lip-reading. Computational speech recognition has been found to improve if visual information about mouth motions can be estimated and used to disambiguate phonemes. Several methods for estimation and interpretation of mouth motion have been proposed [Basu et al., 1998, Bregler et al., 1997, Fleet et al., 2000]. While estimation of mouth motion has been studied there has been very little research on the interpretation of mouth motion in realistic speech.

5. Facial manipulation by the hands. The hands are regularly used to manipulate or hide parts of the face or their movements. There is currently no computational research to uncover this information.

6. Facial tone/color appearance change. These properties include appearing pale, turning red, etc. While humans have a developed ability to judge these underlying phenomena (especially for familiar faces) there is currently no computational research to estimate these properties.

The first steps in computational interpretation of facial movements have been made. Much remain to be done to bring capabilities to an impact-making level. Firstly, recent work has focused primarily on face expression tracking and classification and only marginally on other aspects of facial dynamics. Secondly, the experimentation is not realistic since the number of subjects was usually small, most data is of posed subjects who act emotions and the expression instances are isolated. Progress in the low-level processing of facial perception has not been accompanied by progress in high-level modeling and interpretation. Up until now work has been carried out by computational teams without including psychologists, hence classification performance has not advanced beyond the classification of six basic expressions. Moreover, the ambiguity that is associated with facial expressions has not been adequately modeled.

Data for training and testing

The current database

The database used in this initial study was created by Frank & Ekman in the early 1990's to measure behaviors and judgments of behaviors that occur during interpersonal deception. In this experiment, 20 subjects engaged in two different deception situations. In the first situation – which was the one used in the experiments evaluated in this report - they were instructed to enter a room and search for a briefcase. Inside of this briefcase may or may not be $50. Some were instructed to choose whether to take the money, whereas others were assigned to do so. They were then to return and were interrogated by an expert lie catcher (Ekman). They were instructed that if they took the money, they were to lie and deny that they took the money. Those who did not take the money were instructed to be truthful. All subjects were instructed that there would be consequences and benefits to their behaviors. If the subject lied about taking the money and was able to fool the interrogator, he was told he could keep the $50. If the subject lied about taking the money and was not able to fool the interrogator, was told he would have to return the $50, not receive any reimbursement for participation in the experiment, and has to face anywhere from 10 to 40 110 dbl blasts of white noise for an hour after the experiment. If the subject told the truth about not taking the money, and was believed by the interrogator, he was paid an additional $10. If the subject told the truth and was not believed by the interrogator, he faced the same punishment as the liar who did not fool the interrogator. Note that no subject was punished, and all were paid a subject fee independent of the interrogator's judgment. What was important psychologically to this experiment was that the subject knew that there were these high stakes during the interrogation. A more detailed report of the procedure can be found in Frank & Ekman (1997).

These subjects were videotaped in facial close-up and in full body. This report featured the facial close-up video. This meant typically the shot extended from the bottom of the subject's neck to the top of his head. All subjects were males between 18 and 28 years of age. Six subjects wore glasses, two had beards, and ethnically there were 3.5 Asians, 2.5 Africans (i.e., one subject was ½ Vietnamese and ½ African American), and 14 Caucasians. Each segment analyzed for this report was approximately one minute in length.

Thus, this database featured spontaneous, real human interaction between the interrogator and the subject. These subjects were not coached, asked to act or behave, in any way. They were simply told to try to convince the interrogator of their truthfulness.

Using human coders, Frank & Ekman (1997) found that these subjects were feeling spontaneous emotions. They reported fearing the punishments, and approximately 76% of the subjects could be accurately classified as lying or truthful by using the facial expressions of distress, fear, and disgust, as derived from the FACS scoring. This meant they were feeling strong emotions during this experiment. These subjects were initially scored using FACS event coding rules, with 80% agreement. Event coding rules meant that one did not rescore an individual AU if it did not change more than two degrees of intensity.

In order to facilitate the use of these data for the two teams, the database was rescored by human coders to include stop and start times of individual AUs, without regard to FACS event coding rules. This was done at a high rate of agreement. Moreover, the initial scoring did not include head or eye movements, and so the AUs for head and eye movements were also added in the rescore.

Groups' achievement and comparative performance

A primary objective and contribution of Pittsburgh and San Diego teams has been to develop and test techniques for estimation of action unit activations in spontaneous face movement where subjects are unaware of the presence of the video camera. Therefore, the teams tested the performance of their existing techniques and developed new techniques to handle the particular challenges that stem from the data set. These challenges included: rapid and extensive head motion and occlusion, non-frontal face poses, wearables, head gestures and spontaneously generated facial expressions – typically less accentuated than acted expressions.

Teams' chosen approaches

CMU-Pittsburgh

The CMU-Pittsburgh team built on a solid track record focusing on feature tracking and temporal analysis as a primary vehicle for face expression representation and recognition (see the respective final report as well as the publication list). A particular strength of the team is in model-based visual analysis where representation and image processing techniques combine to solve facial motion problems.

The team assumes that the location of the face is known in the first frame (manually or automatically delineated). Then, the face is tracked using a 3D rigid motion model of a cylinder taking into account perspective projection. The cylinder motion is estimated between consecutive frames and therefore the amount of motion is generally small. Each subsequent face region is warped backward to the initial frame of the face based on the computed and accumulated motions as well as re-registering current pose with respect to reference pose when feasible to totally align the face regions. As a result, the image of head is stabilized to a frontal view of the face throughout the image sequence so that non-rigid motion is more easily measurable.

Four points are manually marked around the eye and eyebrows in the first frame and their region remains fixed throughout the registered image sequence. As a result, changes in the state of the eye are reflected by changes in the region defined by the points at the initial frame. The eye region (i.e., rectangle) is divided into lower and upper parts and the average illumination intensity in these two parts when viewed over time indicates whether an eye closure occurred. Simple hand-crafted rules on the temporal evolution of the illumination curves of the upper and lower eye regions were used to discriminate between a blink, multiple blinks and no-blinks.

Three points marking the right, center and left locations of the eyebrow were manually marked in the first frame. These points were used to recover the contour marking the brow upper area with respect to the skin. At the same time the region directly above the brow is analyzed for wrinkle detection by employing edge

detection. Two sources of information, wrinkle creation and disappearance and contour motion are used by a rule-based system to label brow motion into upward, downward and no-motion.

UCSD

The UCSD team built on a solid track record of developing machine learning techniques for classification of acted facial expressions in frontal views (see the respective final report as well as the publication list). In the area of face analysis members of the team have focused, prior to the current work, on issues of representation of faces and face features with focus on biologically inspired models. The UCSD's team technical approach is based on three main modules:

Image normalization: The image is warped and grayscale-normalized in order to compensate for head rotation and lighting changes. Warping is achieved by estimating the 3D pose of the head (from manually acquired features), projecting image texture onto a 3D head model, re-projecting image texture onto a frontal-view image plane. Grayscale normalization is carried out by histogram normalization.

Feature extraction: vectors of image features are computed by means of multi-resolution multi-orientation filters (Gabors/wavelets). As an alternative front-end the group experiments with the raw grayscale pixel values.

Classification: The expression in each image is classified using binary-decision support vector machines (SVMs), each of which discriminates between any two pairs of image expressions. The results of all such classifications serve as input to a number of Hidden Markov Models (HMMs), each trained on a specific FACS. The latter steps serve to integrate information on multiple frames in time and reach a final decision.
The technical approach is overall sound and well thought-out. The key technical contribution is validating previous work on frontal views of acted AUs in the more challenging scenario of free-head spontaneous AUs. An equally important contribution is providing an end-to-end prototype of an automatic FACS classification system.

Evaluation and Comparative Performance

The teams are the first to quantitatively analyze spontaneous face expressions to estimate Action Units. Objective challenges (illumination, occlusions, imager quality, and diversity of the subjects) mandated significant research effort and slowed down the progress of their research. The teams achieved a remarkable level of performance and automation in the analysis of a small number of AUs and their reported research can inspire and motivate continuing work on this topic.

Following is a comparison of the reported approaches:

1. The teams differed in the mix of automatic versus manual processing of images. CMU/Pittsburgh constrained the manual involvement to the first frame in which points are located around the eyes and eyebrows. UCSD allowed for more manual intervention marking the location of 8 points on the face for every image in the sequence. This difference is not very significant as recent progress in face tracking algorithms warrants the assumption that points can be located on the face automatically with a high degree of success. In both cases this falls short of full-automation.

2. Both teams developed techniques for 3D registration of a subject's head in an image sequence. The CMU/Pittsburgh team employed a cylindrical face model for estimating head motion while the UCSD team employed a 3D mask that can be warped to match the general outline of the face of the user. The benefit of registration is that it cancels out the rigid motion of the head and thus the registered sequence includes only non-rigid facial deformation, i.e., the signal. The pitfall of registration is that unless it is accurate it may introduce spurious facial motions that are difficult to model and may exceed in magnitude the actual signal. Moreover, evaluating the accuracy of registration is tricky in the absence of exact 3D head models as well as a good model of the illumination of the scene. No measurement is yet available on which registration algorithm performed better on the current database.

3. The differences between the teams are most apparent in the recognition of Action Unit activations. CMU/Pittsburgh employ a rule-based system that takes as input a number of hand-crafted measurements on the facial features. UCSD base their process on classifiers (SVMs and HMMs) which are learned automatically from the training data.

Conclusion. Both approaches appear to perform well on the chosen benchmarks. Using a consensus of human coders as the criterion, both systems classified accurately AUs 1+2 and AU 45 at rates of 90% or better. Note that although we use the word accurately here, technically in all instances we are really referring to agreement with human coders, as we did not measure these subcutaneous muscle movements directly via electrodes to establish the ground truth of a particular muscle movement. Regardless, this level of agreement between these systems and the human experts on these AUs is on par with the level of agreement amongst highly trained independent human observers. The consultants agree that the performance of both teams is excellent considering the difficulty of the task at hand, and ought to encourage optimism in the overall feasibility of a system that can detect and classify AUs automatically. Moreover, although not intended, it appears that these systems have also been able to generate information about the dynamic actions of the muscles involved in these facial expressions. This is important because research has suggested that these dynamic patterns may be critical to distinguishing genuine from falsified emotions (e.g., Frank, Ekman, & Friesen, 1993).

For a balanced evaluation of the results consider a few caveats:

1. The AUs examined by the two teams, the eyebrow movements (AU 1 + 2) and the eyeblink (AU 45) were fairly basic and are those on which human observers tend to show high agreement. They are significant facial movements however. Yet in research on deception of concealed emotion, often much more subtle facial movements must be observed. For example, the AU 1 + 2 is often seen in communication, but adding AU 4 to that combination makes that brow movement a significant predictor of the emotion of fear. Yet the visible difference between AU1+2 and AU1+2+4 can be very subtle. Due to a limited sample size, it was not possible to ascertain how well the two systems would do in differentiating these two AU patterns, but the ability to do so is crucial to future research in human behavior. This is another reason why future research requires a larger data set.

2. In future research on more complex behavioral and communicative actions, such as interpersonal deception, other characteristics, such as body posture, paralinguistic measures such as voice tone, and analysis of the speech characteristics, need to be added to the FACS scoring to possibly make more accurate distinctions between truth and deception. Although deception was not the focus of this project, in order to take maximum advantage of this computer based facial scoring other complementary scoring techniques may need to be developed, which ultimately may enhance the performance of automatic AU classification.

3. The current experiments involved three-way AU classification. Many more AUs are relevant to the task of recognizing emotion at a basic level, and detecting deception at a more complex level. When confronted with a larger number of classes, AU classification performance will typically deteriorate. Yet, some AUs are not as important as others in the classification of emotion, so errors in reading those AUs may not compromise overall performance. However, miscoding even a few AUs would negate one of the strengths of FACS, in that it allows for discovery of unpredicted relationships in facial movement.

4. A good metric of AU classification performance is comparing with both trained and untrained human classifiers (see UCSD report, table 1 for non-spontaneous AUs). Such a metric cannot be applied to the current experiments since ground truth was established by consensus of two experts. Therefore we do not know whether 90% agreement performance is the maximum achievable or better performance could be expected.

5. The systems required manual intervention: both in order to select facial features and in order to select the relevant segments of video to be classified. We are not yet looking at fully automatic

systems. A fully automated system may work either worse (useful information coming from human operators is withheld) or better (at times useful features, unknown to researchers, are discovered by unsupervised learning techniques) than a manually assisted one.

6. These computer based systems involving identifying points on the face that are invariant, so as to measure movements in relation to those invariant points. However, some of those points proposed to be invariant in earlier versions of these systems (e.g., the ridge under the eye and at the top of the cheek) are not actually invariant, and can change with AUs such as 6, 12 or 13. Thus, future research should insure that any facial invariants are truly invariant.

New generation FACS database

While the current databases (the one used by the two teams, described above, as well as others circulating in the research community) have been key to the development of the two systems, it is obvious that they have major shortcomings. First and foremost: current databases are too small. They do not contain a sufficient number of each type of AUs for training classifiers. Other shortcomings have emerged amongst which the main ones are: (1) the pixel resolution is sufficient for coarse classification, but not for detecting and measuring subtle changes such as pupil dilation; (2) it would be useful to monitor the posture of the body as well as the face; (3) the point of view changes by as much as 30-40 degrees from AU to AU, thus making training and generalization even more difficult.

If research and development of systems for automatic classification of AUs are to make substantial progress from the current state it is concluded that new databases need to be collected. The paradigm used by Ekman and Frank for collecting the existing database (i.e. interviews with subjects who have been put through a controlled crafted scenario to induce specific intents and emotions) is still valid. What need to be improved are both the quality and the quantity of the data. The main cost in collecting the database is the time taken to gather the subjects, run the experiments and label the AUs. Capturing the same scene from different viewpoints at higher resolution and using a number of different sensors increases this cost moderately.

We make here a number of recommendations and observations in order to guide the design and collection of such databases and estimate the cost of the process:

Resolution: 100-120Hz and 1 Mega pixels. This will allow researchers to obtain sufficient time sampling for quick motions (blink) and fine details (pupil dilation).

Cameras: at least three viewpoints (frontal, 30 degrees on left and right) for the face, to allow for normal head motions and reliable 3D pose estimation as well as obtaining more data for training classifiers that are viewpoint-invariant. Additionally: one or two wide-angle cameras for monitoring the body posture. Color should be recorded and possibly a thermal camera could be used as well. The signal should ideally be digitized on the fly and stored temporarily on computer hard drives.

Audio & other sensors: high quality audio, with microphone placement such that the sound is not influenced by head pose. Also consider: microwave cardiac measurement, passive breathing pattern, skin conductance. We feel that EMG is probably not useful because it would draw subjects' attention to their faces, thus artificially affecting their supposed spontaneous facial reactions. Moreover, there is some controversy about whether simple skin electrodes are measuring just the muscle they have been purported to measure (see Ekman, Fridlund, & Oster, 1987, for more details).

Ground truth: provided by professionals trained and supervised by Ekman and Frank. It is advisable to hire ad-hoc personnel rather than rely on graduate students. Each sequence would ideally be labeled independently by at least two different people.

Subjects: Ideally at least 200 subjects, each filmed for 2-4 minutes. This sample would generate a sufficient number of most AUs both for training and testing. Many of these would be head movements and eye movements/blinks; however, one could reasonably expect - based on previous work - at least 1/3rd of these AUs would be facial movements. Note that some of these AUs are quite rare, both in the laboratory as well as in the real world (e.g., AU 11, or AU 13), and thus we would expect a limited sample of these AUs. In fact, their scarcity suggests it may not be critical to distinguish them. Finally, the subjects should be diverse in sex, age, ethnic group, and wearables.

Scenarios: There are a variety of deception scenarios – false opinion, mock theft, concealed information, etc., that have been quite successful in generating spontaneous emotion. The experiments could also involve a scenario where subjects are shown films designed to elicit both positive and negative emotions. Finally, the experiments can use the marital dispute scenario, where couples discuss problem areas in their relationship, which has also been quite successful in generating spontaneous emotions.

Restrictions of use and consents from subjects: Any new data set should stipulate that the subjects have the option of allowing the full use of their behaviors by not only the research team, but in the future to allow access by other groups of credible scientists. There should also be no "destroy-by" date. This particular procedure allows the most comprehensive study and building upon findings to advance progress in this area of research. The rules of the Agency funding data collection should allow the use of the data to credible scientists and not stipulate destruction of the data. Ideally the subjects would consent to distribution of data at least to select groups of scientists. Careful planning of the consent forms is vital to make the database available to as wide a number of scientists and engineers as possible.

Cost: Approximately 2-3 man-years of technician time for double-labeling 200 sequences (this includes training technicians). One half man-year of senior scientist's time for running the experiments, training and supervising labeling technicians. Two years support for graduate student to help in coordinating experiments and documenting the process. Cost of the hardware is on the order of $70-100K. Total cost of the effort is in the order of $500K including overhead.

Support for the data: Tens of DVDs could be used for storing the data for dissemination. Minimally compressed master copies should be kept on hard-drives or other appropriate support.

Availability of the proposed database: This sort of database should be a joint effort between behavioral and computer scientists. It may even be converted to some sort of national archive, that by the nature of its availability maximize the impact of this work for not only basic behavioral scientists but in particular computer science/computer vision researchers. This might encourage other collections examining other spontaneous human interactions and reactions as well.

Potential users of these new generation databases: A large group of basic and applied researchers would be the beneficiaries of such new data sets. Behavioral scientists would be very interested in the basic emotion process, and the relationships between expressive behaviors and internal physiological processes. This goes beyond facial expression, but also color changes (e.g., researching embarrassment), head movements, eye movements, and so forth. Psychologists as well as the intelligence, immigration, customs, border patrol, and law enforcement community in general would be interested in these emotions and other actions in relation to how they predict deception, or falsification of information, possession of contraband, the status of interpersonal relationships, or true versus fabricated attitudes and opinions. This database would allow for more efficient and cost-effective testing of various hypotheses to uncover the patterns and actions associated with not only these nefarious acts, but also day-to-day significant interpersonal interactions. The computer vision community would benefit by using a realistic database that is supported by a credible ground truth measurements.

Recommendations

The realization of a system for automatic detection of deception and, in general, classification of emotions as expressed by the human face appears to be a reachable goal given the work of the two teams. The current achievements, while impressive, fall short of achieving the full potential of understanding human communications. The reviewed research is the first to address spontaneous expressions computationally. We feel that it is imperative to continue this research.

The laborious nature of coding reliably facial expressions and other human behaviors has been *the main impediment* to progress in much of the social sciences (Frank, 2002). An automated coding system is the key to unlocking this research. For example, researchers have noted reliable indicators of treatment outcomes in psychopathology via facial expressions, yet to take the time to learn FACS and apply it has in all likelihood slowed this research considerably (See Ekman & Rosenberg, 1997, for a review). Likewise, research on human emotion in stressful interpersonal interactions, be they problem marriages (Gottman, 1994), child abuse (Bugental, 1986), or deception (Frank & Ekman, 1997), have all shown the utility of facial expressions for diagnostics and/or treatment. Yet researchers examining these topics note that the laborious nature of the facial coding by humans has limited examining the face and other nonverbal markers to a smaller circle of researchers who have invested the many hours learning FACS. This is despite the fact that the study of facial behavior has been the driving force involved in bringing biological principles back into social science.

The two main recommendations we make are:

a) A better database needs to be collected
b) Funding for further research on the topic should be awarded in a way that encourages the formation of multidisciplinary teams composed of engineers and psychologists.

Research infrastructure

A realistic database is vital to guide researchers in both communities (psychology and vision) to develop credible research that leads to tangible findings. Technical and legal concerns made it difficult to obtain data. While recent technological progress makes it easy to acquire large volumes of data even in a clandestine situation such as the one provided to the teams, privacy concerns are an obstacle to wide distribution. The priority at this point is the collection of a larger and better database on which to train and test existing and new approaches. Current work is data-limited and further improvement in high-level interpretation of facial motion is unlikely without a new database. Once such a database is available a number of short-term gains may be obtained in fine-tuning and evaluating existing systems. Most notably multi-class discrimination experiments with more than 3AUs and comparison with human observers may be carried out. We make a concrete proposal for the creation of such a database earlier in this document. Furthermore, the existence of a better database would have two key effects: encourage interest in this problem of more research teams, thus enabling faster progress and better experimental validation; moreover, allow psychologists to extend the FACS codes to full facial behaviors, as well as the development of complementary scoring systems to measure body motion and speech behavior.

Basic research topics

In order to achieve a new generation of systems that are better able to interpret human emotion a number of research issues needs to be addressed. These include:

1. Interpretation and data-mining of facial and upper body movements. A sufficiently large dataset enables researchers to develop and employ algorithms to uncover behaviors and correlations that can be tedious to detect through a hypothesize-and-test approach.

2. Multi-modal analysis of human expressions and communications. Face expression is one aspect of the multi-dimensional space of human communication and expression. Other visual, verbal, acoustic and contextual issues are integral to human and computational interpretation. However, facial expressions are mostly studied in isolation and as a result significant ambiguity and qualification need to be taken into account. Arguably, as automatic facial movement estimation

improves it will become possible to study the complex of human communication as a whole. Automatic extraction and analysis of cues such as head and body posture and speech as well as tongue movements, swallowing and gaze need to be addressed.

3. FACS plays a fundamental role in the computational interpretation of facial expressions since its quantitative properties lend themselves to mathematical formulation. However, FACS is first and foremost a representation that enables psychologists to communicate observations and develop hypotheses about facial interpretations. In the absence of competing quantitative models it is necessary to harmonize FACS with the fundamental capabilities of computational processing. Specifically, a joint team may propose a representation that uses FACS as well as computational considerations to bridge the current gap between the vision and psychology communities and thus arrive at an automation-viable representation that is psychology-compatible as well. A related effort has been in progress in the video compression community (MPEG-7) where representations that are designed to measure and encode facial deformation where developed to take into account computational as well as facial appearance quality factors.

4. Developing metrics for evaluating performance of the critical stages of face analysis enables progress in the field to be solid and more rapid. The research conducted so far by the Pittsburgh and San Diego teams is the first step in this direction.

5. In-depth analysis of deception detection by human observers. Discover additional cues, and explore methods for aggregating multiple-cue information. Study the aspects of voluntary and involuntary expression recognition.

6. Exploration of direct learning-based approaches which bypass AU modeling (this may require one order of magnitude more training data than currently envisioned). AUs are just a means to describe facial expressions in order to measure human emotion. It is possible that a machine vision system could be trained directly to recognize emotion from image sequences.

7. Moving beyond just presence/absence of facial movements into automatic coding of the dynamic flow of the facial expression – e.g., smoothness of onset, synchrony amongst AUs, duration of movements – as these dynamic features have been found to distinguish experienced emotion from falsified emotion (Frank, Ekman, & Friesen, 1993). This dynamic information facilitates a number of research programs investigating the role of dynamics in perception and interpretation of facial expressions.

8. The expansion of facial analysis work to include the face viewed under arbitrary angles, expression, lighting conditions, occlusions, wearables and imaging-variant conditions.

Research organization strategy

In the near term, it is suggested to enhance the realism and quality of data available to the psychology and computer vision communities and encourage close interaction and exchange of findings. Also it is suggested that further study of communication aspects that are closely related to face expressions be undertaken to improve the basic capability of measuring and interpreting their occurrence. In the long term integrated studies in which the multidimensional nature of human expression and expression perception will be studied computationally, thereby enabling researchers in behavioral studies to utilize quantitative data in research to uncover and characterize human communication and expression.

There are two intertwined components that need to be addressed: psychological/social studies leading to better understanding of human expressive modalities and computational models to support automatic analysis of these expressions. This inter-dependence suggests that it is most beneficial to achieve tandem progress if at all feasible. Psychologists, who have so far painstakingly proceeded by hand-analysis of motion sequences will see their efforts augmented by the ability to process large volumes of data automatically and will benefit from the rigor required by engineering-level system design. Conversely,

engineers will need expert knowledge in human psychology in order to be successful at the task for building machines that interact successfully with humans.

For example, behavioral scientists should focus on developing scenarios that produce a large number of spontaneous facial expressions and movements. This should be done in consultation with computer vision scientists to insure the technical aspects of any data collection are adequate for computer study. As the data is collected, the original research teams should have both human coding, as well as computer vision coding. Once the teams establish the patterns of accuracy and agreement, then these data should be released to other teams of scientists to insure reliability as well as potential discovery of new or improved techniques for measuring human facial behavior.

Involvement of more research groups would benefit the field by creating a community with some level of critical mass. The problem should be broadened from detection of deception to visual and auditory classification of emotional states and intention. Care should be taken in involving both engineers and psychologists and in promoting collaborative interaction both in analysis and in the design of engineered systems.

References

Basu, S., Oliver, N., Pentland, A. (1998). 3D modeling and tracking of human lip motions. Sixth International Conference on Computer Vision, 1998, 337 –343.

Black, M.J., and Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion, Int. Journal of Computer Vision, 25(1), 1997, 23-48.

Bonanno, G.A., and Keltner, D. (1997). Facial expressions of emotion and the course of conjugal bereavement. Journal of Abnormal Psychology, 106, 126-137.

Bregler, C., Covell, M., and Slaney, M. (1997). Video Rewrite: Driving Visual Speech with Audio.Proc. ACM SIGGRAPH 97, in Computer Graphics Proceedings, Annual Conference Series, (1997).

Bugental, D.B. (1986). Unmasking the "polite smile": Situational and personal determinants of managed affect in adult child interaction. Personality and Social Psychology Bulletin, 12, 7-16.

Cohn, J.F., Zlochower, A.J., Lien, J.J., Kanade, T. (1998). Feature-point tracking by optical flow discriminates subtle differences in facial expression. Proceedings of the International Automatic Face and Gesture Recognition, 1998, 396 –401.

Darwin, C. (1872/1998). The expression of the emotions in man and animals. New York: Oxford. (3rd Edition, w/ commentaries by Paul Ekman).

Davidson, R.J. (1984). Affect, cognition and hemispheric specialization. In C. E. Izard, J. Kagan, and R. Zajonc (Eds.) Emotion, cognition, and behavior. New York: Cambridge University Press, 320-365.

Davidson, R.J. (1992). Emotion and affective style: Hemispheric substrates. Psychological Science, 3, 39-43.

Davidson, R. J., Ekman, P., Saron, C., Senulius, J., and Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology I. Journal of Personality and Social Psychology, 58, 330-341.

Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., and Sejnowski, T.J. (1999). Classifying facial actions. IEEE Transactions on Pattern Analysis and Machine Intelligence, (21) 10, Oct. 1999, 974 –989.

DeCarlo, D., and Metaxas, D. (2000). Optical flow constraints on deformable models with applications to face tracking. International Journal of Computer Vision, July 2000, 38(2), pp. 99-127.

Edwards, G.J., Taylor, C.J., and Cootes, T.F. (1998). Interpreting face images using active appearance models**.** Proceedings of the International Conference on Automatic Face and Gesture Recognition, 1998, 300 –305.

Ekman, P. (1972). Universal and cultural differences in facial expressions of emotion. In J K Cole (Ed.) Nebraska symposium on motivation 1971 Lincoln, NE: University of Nebraska Press, 207-283.

Ekman, P. (1985). Telling lies: Clues to deceit in the marketplace, politics, and marriage. New York: Norton.

Ekman, P. (1994). Strong evidence for universals in facial expression: A reply to Russell's mistaken critique. Psychological Bulletin, 115, 268-287.

Ekman, P., and Davidson, R.J. (1993). Voluntary smiling changes regional brain activity. Psychological Science, 4, 342-345.

Ekman, P., Davidson, R. J., and Friesen, W.V. (1990). The Duchenne smile: Emotional expression and brain physiology II. Journal of Personality and Social Psychology, 58, 342-353.

Ekman, P., and Friesen, W.V. (1978). The facial action coding system. Palo Alto: Consulting Psychologists Press.

Ekman, P., and Friesen, W.V. (1982). Felt, false, and miserable smiles. Journal of Nonverbal Behavior, 6, 238-252.

Ekman, P., Friesen, W.V., and Ancoli, S. (1980). Facial signs of emotional experience. Journal of Personality and Social Psychology, 39, 1125-1134.

Ekman, P., Friesen, W.V., and O'Sullivan, M. (1988). Smiles when lying. Journal of Personality and Social Psychology, 54, 414-420.

Ekman, P., Huang, T.S., Sejnowski, T.J., and Hager, J.C. (1992). Final Report to NSF of the Planning Workshop on Facial Expression Understanding. 1992.

Ekman, P., Levenson, R.W., and Friesen, W.V. (1983). Autonomic nervous system activity distinguishes among emotions. Science, 221, 1208-1210.

Ekman, P., O'Sullivan, M., Friesen, W.V., and Scherer, K. (1991). Invited article: Face, voice, and body in detecting deceit. Journal of Nonverbal Behavior, 15, 125-135.

Ekman, P., and E.L. Rosenberg (Eds.) (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System. (217-242). New York: Oxford.

Essa, I.A., Pentland, A.P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, (19) 7, July 1997, 757 –763.

Fleet, D. J., Black, M. J., Yacoob, Y., and Jepson, A..D. (2000). Design and use of linear models for image motion analysis, Int. Journal of Computer Vision, 36(3), 2000, 171-193.

Fox, N.A., and Davidson, R.J. (1988). Patterns of brain electrical activity during facial signs of emotion in 10-month old infants. Developmental Psychology, *24*, 230-236.

Fridlund, A. J., Ekman, P., Oster, H. (1987). Facial expressions of emotion. In A. W. Siegman, and S. Feldstein, (Eds). (1987). Nonverbal behavior and communication (2nd ed.). (143-223). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Frank, M.G. (2002). Facial expressions. In N. Eisenberg (Ed.) International Encyclopedia of the Social and Behavioral Sciences. (in press). Oxford: Elsevier.

Frank, M. G. and Ekman, P. (1993). Not all smiles are created equal: The differences between enjoyment and other smiles. Humor: The International Journal for Research in Humor, 6, 9-26.

Frank, M.G., and Ekman, P. (1997). The ability to detect deceit generalizes across different types of high stake lies. Journal of Personality and Social Psychology, 72, 1429-1439.

Frank, M. G., Ekman, P., and Friesen, W.V. (1993). Behavioral markers and recognizability of the smile of enjoyment. Journal of Personality and Social Psychology, 64, 83-93.

Gottman, J. (1994). Why marriages succeed or fail. New York: Fireside.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. Psychological Bulletin,115, 288-299.

Keltner, D. (1995). The signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. Journal of Personality and Social Psychology, 68, 441-454.

La Cascia, M., Sclaroff, S., and Athitsos, V. (2000). Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models. IEEE Transactions on Pattern Analysis and Machine Intelligence, (22) 4, April 2000, 322 –336.

Lanitis, A., Taylor, C.J., Cootes, T.F. (1997). Automatic interpretation and coding of face images using flexible models. IEEE Transactions on Pattern Analysis and Machine Intelligence, (19) 7 , July 1997, 743 –756.

Levenson, R.W., Carstensen, L.L., Friesen, W.V., and Ekman, P. (1991). Emotion, physiology, and expression in old age. Psychology and Aging, *6*, 28-35.

Levenson, R.W., Ekman, P., and Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. Psychophysiology, 27, 363-384.

Levenson, R.W., Ekman, P., Heider, K., and Friesen, W.V. (1992). Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra. Journal of Personality and Social Psychology, 62, 972-988.

Lyons, M.J., Budynek, J., Akamatsu, S. (1999). Automatic classification of single facial images**.** IEEE Transactions on Pattern Analysis and Machine Intelligence, (21)12 , Dec. 1999, 1357 –1362.

Mase, K. (1991). Recognition of Facial Expression from Optical Flow. IEICE Transactions, E 74, 10, 1991, 3474-3483.

Matsumoto, Y., Zelinsky, A. (2000). An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement**.** Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2000 499 –504.

Morimoto, C.H., and Flickner, M. (2000). Real-Time Multiple Face Detection Using Active Illumination. Proc. of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, Grenoble, France, March 2000, 8-13.

Padgett, C., and Cottrell, G.W. (1996). Representing face images for emotion classification. Proceedings Conference on Advances in Neural Information Processing System, 1996, 894-900.

Pantic, M., and Rothkrantz, L.J.M. (2000). Automatic analysis of facial expressions: the state of the art**.** IEEE Transactions on Pattern Analysis and Machine Intelligence, (22) 12, Dec. 2000 Page(s): 1424 –1445.

Pappu, R., and Beardsley, P.A. (1998). A qualitative approach to classifying gaze direction. Proceedings of the International Conference on Automatic Face and Gesture Recognition, 1998, 160 –165.

Rosenblum, M., Yacoob, Y., and Davis, L. (1996). Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture. IEEE Trans. on Neural Networks, 7(5), 1996, 1121-1138.

Sayette, M. A., Smith, D. W., Breiner, M.J., and Wilson, G. T. (1992). The effect of alcohol on emotional response to a social stressor. Journal of Studies on Alcohol, 53, 541-545.

Shih, S., Wu, Y., and Liu, J. (2000). A calibration-free gaze tracking technique. Proceedings. 15th International Conference on Pattern Recognition, Volume: 4, 2000, 201–204.

Terzopoulos, D., and Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. IEEE Transactions on Pattern Analysis and Machine Intelligence, (15) 6, June 1993, 569 –579.

Tian, Y.I., Kanade, T., and Cohn, J.F. (2001). Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, (23) 2 , Feb. 2001, 97 –115.

Yacoob, Y., Davis, and L.S. (1996). Recognizing human facial expressions from long image sequences using optical flow**.** IEEE Transactions on Pattern Analysis and Machine Intelligence, (18) 6, June 1996, 636 –642.