Jeffrey F. Cohn

Advances in Behavioral Science Using Automated Facial Image Analysis and Synthesis

The face conveys information about a person's age, sex, background, and identity; what they are feeling, thinking, or likely to do next. Facial expression regulates face-to-face interactions, indicates reciprocity and interpersonal attraction or repulsion, and enables intersubjectivity between members of different cultures. Facial expression indexes neurological and psychiatric functioning and reveals personality and socioemotional development. Not surprisingly, the face has been of keen interest to behavioral scientists.

About 15 years ago, computer scientists became increasingly interested in the use of computer vision and graphics to automatically analyze and synthesize facial expression. This effort was made possible in part by the development in psychology of detailed coding systems for describing facial actions and their relation to basic emotions, that is, emotions that are interpreted similarly in diverse cultures. The most detailed of these systems, the Facial Action Coding System (FACS) [1], informed the development of the MPEG-4 facial animation parameters for video transmission and enabled progress toward automated measurement and synthesis of facial actions for research in affective computing, social signal processing, and behavioral science.

INTRODUCTION

Early work focused on expression recognition between closed sets of posed facial actions. More recently, investigators have focused on the twin challenges of action unit (AU) detection in naturalistic settings, in which low base rates, partial occlusion, pose variation, rigid head motion, and lip movements associated with speech complicate detection, and real-time synthesis of photorealistic avatars that are accepted as live video by naïve participants. This article reports key advances in behavioral science that are becoming possible through these developments. Before beginning, automated facial image analysis and synthesis (AFAS) is briefly described.

Digital Object Identifier 10.1109/MSP.2010.938102

AUTOMATED FACIAL IMAGE ANALYSIS AND SYNTHESIS

A number of approaches to AFAS have been proposed. A leading one used in most of the research described below is referred to as either a morphable model (MM) [2] or an active appearance model (AAM) [3]. The terms MM and AAM can be used interchangeably, as discussed by [4], so for the purposes of this article we shall refer to them collectively as an AAM. The AAM is a statistical shape (rigid and nonrigid) and appearance model that describes a holistic representation of the face [4]. Given a predefined linear shape model with linear appearance variation, AAMs align the shape model to an unseen image containing the face and facial expression of interest. The shape $\mathbf{s}$ of an AAM is described by a triangulated mesh. The coordinates of the mesh vertices define the shape $\mathbf{s}$. These vertex locations correspond to a source appearance image, from which the shape is aligned. Since AAMs allow linear shape variation, the shape $\mathbf{s}$ can be expressed as a base shape $\mathbf{s}_0$ plus a linear combination of m shape vectors

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i,$$

where the coefficients $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$ are the shape parameters. Additionally, a global normalizing transformation (e.g., a geometric similarity transform) is applied to $\mathbf{s}$ to remove variation due to rigid head motion. Given a set of training shapes, Procrustes alignment is employed to normalize these shapes and estimate the base shape $\mathbf{s}_0$, and principal component analysis (PCA) is used to obtain the shape and appearance basis eigenvectors $\mathbf{s}_i$ (Figure 1).

SYNTHESIS

An important advantage of AAMs is that the models are approximately invertible. Synthetic images that closely approximate the source video can be generated from the model parameters. An example can be seen in Figure 1(d)–(f), which shows appearance synthesized directly from an AAM. In some of the examples described below, we exploit the synthesis capabilities of AAMs to investigate human social dynamics. AAMs have made possible for the first time to experimentally disambiguate static cues (sex, age, and so on) from biological motion (such as expression and gesture).

EXPRESSION DETECTION

In many applications, it is of interest to know what facial actions have occurred and their intensity. Support vector machine (SVM) classifiers, as an example, may be trained from video that has been labeled for FACS AUs, emotion-specified expressions, or other descriptors. SVMs attempt to find the hyperplane that maximizes the margin between positive and negative observations for a specified class. Accuracy typically is quantified as A', which is the area under the receiver operating characteristics (ROC) curve. A' values can range between .5 (chance) and 1 (perfect agreement).

FACIAL IMAGE ANALYSIS FOR BEHAVIORAL SCIENCE

Standard methods for facial expression analysis are manual annotation

1053-5888/10/$26.00©2010IEEE

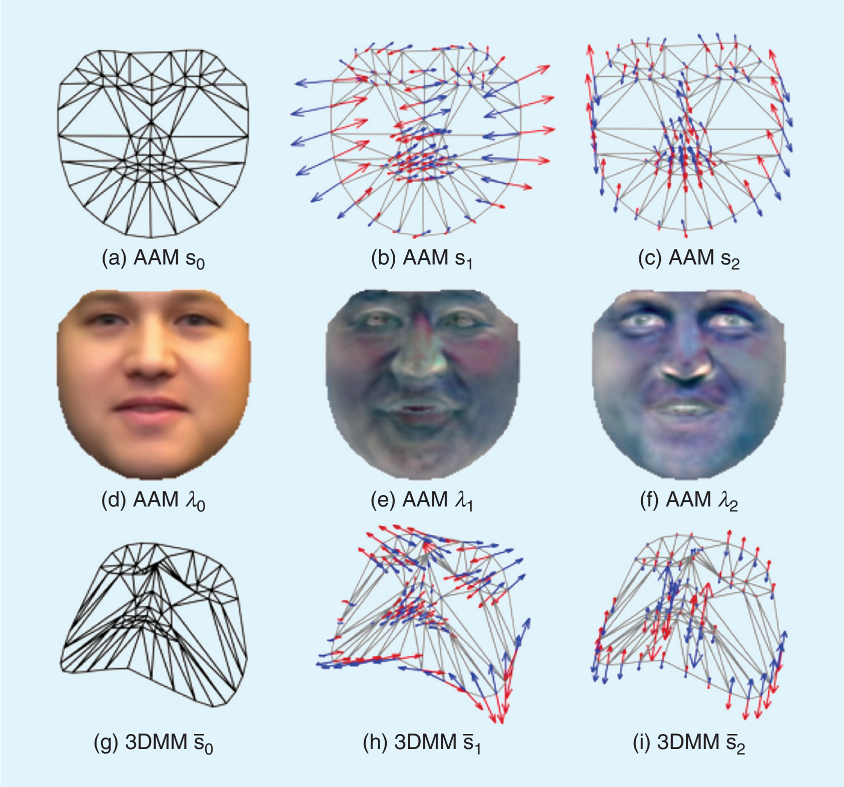
(referred to as coding in behavioral science), perceptual judgments, and facial electromyography (EMG). The FACS [1], which is the most comprehensive manual method, defines over 44 anatomic AUs that individually or in combinations can describe nearly all possible facial movements. FACS itself makes no inferences about the meaning of AUs (e.g., emotion). As with all manual methods, FACS is inherently subjective, labor intensive, and difficult to employ consistently across laboratories and over time. Perceptual judgment methods can reveal the meaning of events and shifts in event categories but are less able to inform what information people use when making judgments. Manual coding and perceptual judgment methods can be complementary when used in combination.

Facial EMG enables quantitative, automated measurement of muscle contractions that underlie facial actions but has several limitations. The number of possible sensors is limited; they lack specificity, may inhibit facial expression, and cannot distinguish between observable and occult facial actions. Motion capture provides more coverage and potential specificity but is expensive, time consuming, and reactivity effects are of particular concern. AFAS makes automated, quantitative measurement of the timing of specific observable facial actions possible without use of sensors or motion-capture technology that may inhibit spontaneous movement.

My colleagues and I have used AFAS to detect FACS AUs and expressions; evaluate its concurrent validity with manual measurement of AU intensity; investigate the timing and configuration of facial actions in relation to observers' judgments of facial expression; study physical pain and clinical depression; assess reciprocity between mothers and infants; and investigate human social dynamics using a videoconference paradigm.

AUTOMATIC DETECTION OF FACIAL AUs

Motivated by our interest in emotion expression and social interaction, we have used several databases to focus on AUs that are most common in these contexts.



(a) AAM s_0 (b) AAM s_1 (c) AAM s_2

(d) AAM λ_0 (e) AAM λ_1 (f) AAM λ_2

(g) 3DMM $\bar{s}_0$ (h) 3DMM $\bar{s}_1$ (i) 3DMM $\bar{s}_2$

[FIG1] An example of the computation of AAM shape and appearance. The figure shows the mean and first two modes of variation of (a)–(c) two-dimensional AAM shape and (d)–(f) appearance variation and (g)–(i) first three three-dimensional shape modes. (Image used with permission from IEEE Computer Society).

Rutgers University-FACS (RU-FACS) [5] consists of interviews with young adults. Head pose is frontal with small to moderate head motion and speech. The Group-Formation-Task (GFT) database [6] includes unstructured conversations between groups of three young adults over the course of a half hour. Partial occlusion, nonfrontal pose, moderate head rotation, and speech are common. The Spectrum database [7] includes serial symptom interviews with depressed outpatients over the course of their treatment. Pose is nonfrontal, AUs have lower intensity than in the other databases, and, like them, head motion, occlusion, and speech are common.

Available FACS codes for each database vary. In RU-FACS, approximately ten AU occurred with sufficient frequency to train and tune classifiers. In GFT, the investigators were especially interested in AU 6 and AU 12 as indices of positive emotion. AU 6 refers to tightening of the sphincter muscle around the eye (which lifts the cheeks and causes crow's feet wrinkles to form) and AU 12 refers to the zygomatic major muscle that lifts the lip corners obliquely in a smile. These two AUs in combination have been described as the *Duchenne smile*.

Previous research most often has trained and tested classifiers in separate subsets of the same database. A more rigorous approach is to train and test in separate databases collected under different conditions. We trained and tuned classifiers in RU-FACS and tested them in Spectrum and GFT.

Accuracy was high in both. For the ten most frequent AUs and AU combinations in Spectrum, A′ averaged 79.4. In GFT, A′ for both AUs was .90 or greater in absence of partial occlusion, and .90 and .75 (for AU 6 and AU 12, respectively) when partial occlusion occurred [6]. These results were robust to head pose and head motion. Within plus/minus 20° pitch and yaw, accuracy ranged from .84 to .98. Together, these findings from depression interviews and group interaction tasks suggest that classifiers independently trained in one data set can accurately detect socially significant actions in completely independent data sets.

COMPARISON WITH CRITERION MEASURES OF FACIAL DYNAMICS

Expression intensity and change in intensity over time modulate an expression's message value. Smiles perceived as insincere, for instance, have more rapid onset to peak intensity than those judged sincere. Polite smiles typically have lower intensity than those of enjoyment [8]. If automated facial image analysis can detect facial actions reliably, how well can it detect their intensity? The correlation between automated and manual measurement of AU intensity has been evaluated several ways.

One is to compare measures of physical displacement. Facial EMG is a gold standard for measurement of facial muscle activity. AFAS and zygomaticus major EMG were compared for lip corner motion (AU 12). Not surprisingly, facial EMG had greater sensitivity to occult muscle contractions, but the two measures were highly consistent for observable smiles (r = .95).

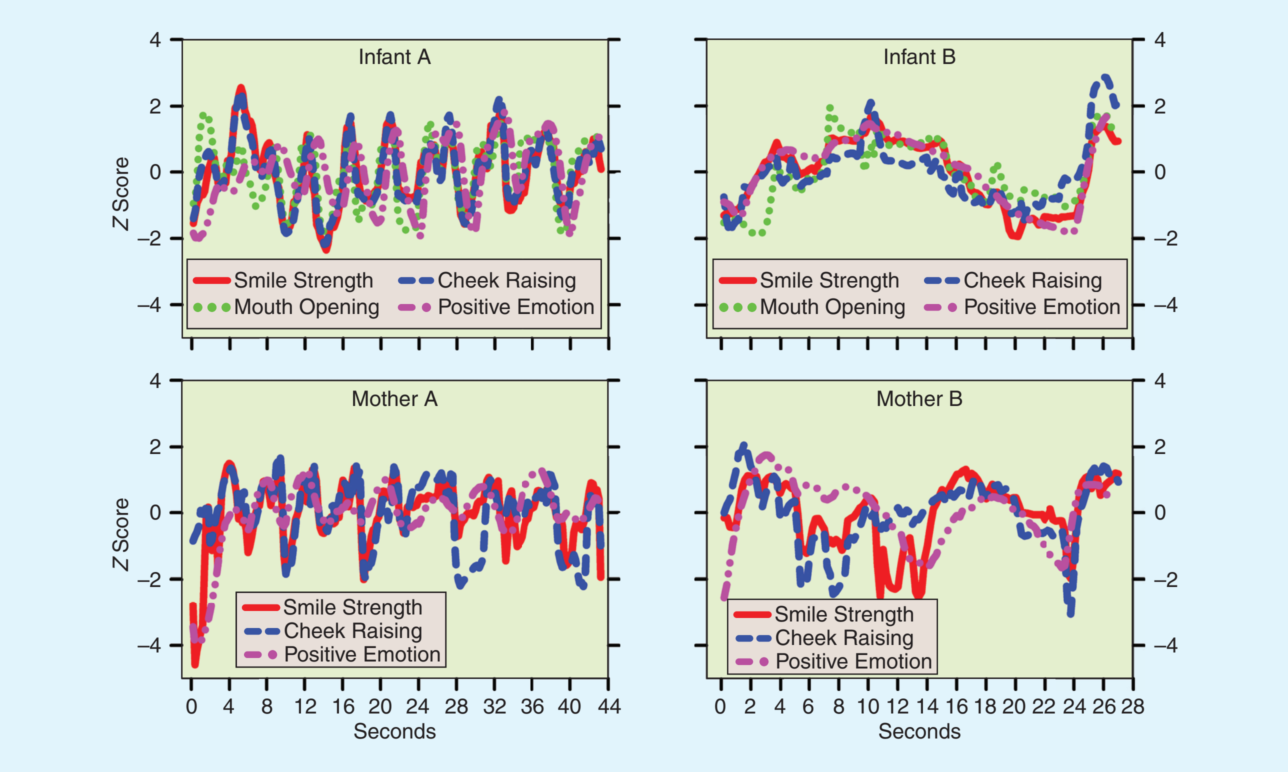
A second is to ask whether movement is consistent with perceptual judgments of an expression's message value, such as its emotional valence. In [9], naïve observers used a joystick device to measure the perceived valence of mothers' and infants' expressions during face-to-face interaction. A portion of the resultant time series for perceived valence and FACS coded intensity of AU 6, AU 12, and AU 26 (mouth opening) are plotted in Figure 2. Intersystem consistency was high, which supports the use of automated facial image analysis to measure the intensity of perceived positive emotion.

SYNCHRONY

The time series (Figure 2) reveal synchrony within and between partners. Within partners, eye constriction and smiling (AU 6 and AU 12, respectively) were coupled, which is consistent with the hypothesis that the Duchenne smile is a temporally integrated event. Between partners, cycles of eye constriction and smiling were loosely coupled with moderate cross correlations between time series.

CONFIGURATION AND TIMING OF SMILES IN RELATION TO THEIR PERCEIVED MEANING

Smiles are the most frequent facial expressions and can communicate diverse meanings. What accounts for the differences in meaning that may occur? All smiles involve an oblique pull of the lip corners (AU 12 in FACS), but what influences whether a smile is perceived as one of delight, embarrassment, or politeness? The authors in [8] used a combination of manual and automated measurement to answer this question. They selected unposed smiles that occurred during the recording of facial action tasks. Smiles were considered unposed if they occurred at times other than when participants were requested to perform a smile or other



[FIG2] Smile parameters and rated positive emotion over time. Infant graphs show the association of automated measurements of smile strength (AU 12 intensity), eye constriction (AU 6 intensity), mouth opening (AU 25-27), and rated positive emotion. Mother graphs show the association of automated measurements of corresponding parameters. Positive emotion is offset by 0.6 s to account for rating lag. (Image used with permission from Taylor & Francis Group, LLC.)

action. These unposed smiles then were shown to groups of naïve observers to classify into one of several categories: amusement, polite, and embarrassed or nervous. Morphological characteristics, such as presence or absence of AU 6 were derived from manual FACS coding. Dynamic characteristics, such as maximum velocity of the lip corners during smile onsets and offsets, head nods and head turns, were measured automatically.

The three types of smiles varied with respect to multiple features. Relative to perceived polite smiles, perceived amused smiles had larger amplitude, longer duration, more abrupt onset and offset, and more often included AU 6, open mouth, and smile controls. Relative to those perceived as embarrassed/nervous, perceived amused smiles were more likely to include AU 6 and have less downward head movement. Relative to those perceived as polite, perceived embarrassed/nervous smiles had greater amplitude, longer duration, more downward head movement, and were more likely to include open mouth. These findings begin to answer the question about what characteristics influence the perceived meaning of facial actions and exemplify how automated and manual measurements may be used in combination.

PAIN DETECTION

Pain is difficult to assess and manage. Pain is fundamentally subjective and is typically measured by patient self-report. Using a visual analog scale (VAS), patients indicate the intensity of their pain by marking a line on a horizontal scale, anchored at each end with words such as "no pain" and "the worst pain imaginable." This and similar techniques are popular because they are convenient, simple, satisfy a need to attach a number to the experience of pain, and often yield data that confirm expectations.

Self-report measures, however, have several limitations. They are idiosyncratic, depending as they do on preconceptions and past experience; are susceptible to suggestion, impression management, and deception; and lack utility with young children, individuals with certain types of neurological impairment, many patients in postoperative care or transient states of consciousness, and those with severe disorders requiring assisted breathing, among other conditions.

Pain researchers have made significant progress toward identifying facial actions indicative of pain. These include brow lowering (AU 4), orbital tightening (AU 6 and 7), eye closure (AU 43) and nose wrinkling, and lip raise (AU 9 and 10). Previous work suggested that these actions could be identified automatically. This led us to ask whether AFAS could replicate expert ratings of pain.

Participants with a history of shoulder injury (e.g., torn rotator cuff) were recorded while manipulating their affected and unaffected shoulders. Their facial behavior was FACS coded and pain was measured using a composite of AUs associated with pain and with self-report. AFAS successfully detected each of the key AUs and precisely identified episodes of pain [10], [11].

> AN IMPORTANT CLINICAL IMPLICATION IS THAT AUTOMATED PAIN DETECTION IN BIOMEDICAL SETTINGS APPEARS FEASIBLE AND READY FOR TESTING.

Two related findings also emerged. First, pain could be detected with comparable accuracy either directly from AAM features fed to a classifier or by a two-step classification in which core AUs were first detected and they in turn were given to a classifier to detect pain. This finding suggests that classifier design may be simplified in related applications and has implications for our understanding of the face of pain. Second, adequate results could be achieved from relatively coarse ground truth in place of frame-by-frame behavioral coding. Both findings have implications for pain detection and related detection tasks, especially those for which relatively longer behavioral states (e.g., pain or depression) are of interest rather than fast facial actions. An important clinical implication is that automated pain detection in biomedical settings appears feasible and ready for testing.

DEPRESSION SEVERITY

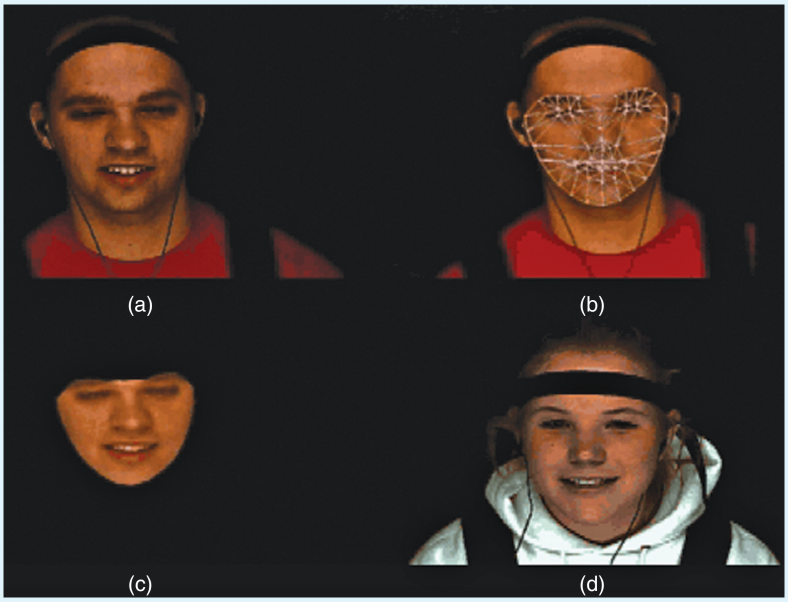
Diagnosis and assessment of symptom severity in psychopathology are almost entirely informed by what patients, families, or caregivers report. Standardized procedures for incorporating facial and related nonverbal expression are lacking. This is especially salient for depression, for which there are strong indications that facial expression and other nonverbal communication may be powerful indicators of disorder severity and response to treatment. In comparison with nondepressed individuals, depressed individuals look less at conversation partners, gesture less, show fewer Duchenne smiles, more sadness and suppressor movements, and less facial animation. Depressed mothers are slower and more variable to respond to their infants [7].

To learn whether such effects might be detected using AFAS, we studied patients enrolled in a clinical trial for treatment of depression. Each was seen on up to four occasions over 20 weeks. Sessions were divided into those with clinically significant severity and those meeting criteria for remission. Automated measurement was 79% accurate in discriminating these two types [7]. More recent findings suggest that lowered head pitch and head turns away from the clinician along with more variable vocal turn-taking were especially discriminative.

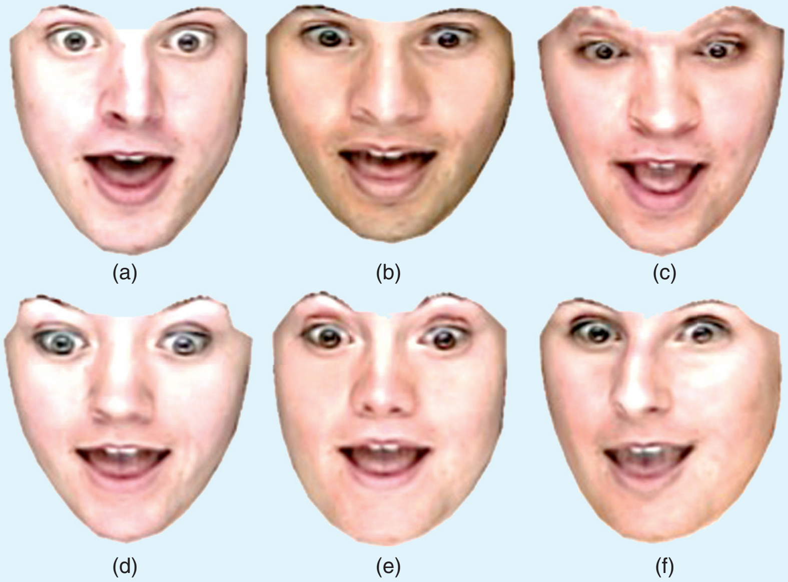
These findings are particularly informative in that all participants initially met criteria for depression. Almost all previous work has compared only depressed and nondepressed comparison participants. Because depression is moderately to highly correlated with personality (especially neuroticism) comparisons between depressed and nondepressed participants lack specificity for depression. This study is one of very few to find within a clinical sample that depression severity is revealed by expressive behavior and to do so using automated measurement.

FACIAL IMAGE SYNTHESIS FOR BEHAVIORAL SCIENCE

In conversation, expectations about another person's identity are inseparable from their actions. Even over the telephone, when visual information is unavailable, we make inferences from the sound of the voice about the other person's gender, age, and background. To what extent do we respond to whom we think we are talking to rather than to the dynamics of their behavior? This question had been unanswered because it is difficult to separately manipulate expectations about a person's identity from their actions. An individual has a characteristic and unified appearance, head motions, facial expressions, and vocal inflection. For this reason, most studies of person perception and social expectation are naturalistic or manipulations in which behavior is artificially scripted and acted. But scripted and natural conversations have different dynamics. AFAS provides a way out of this dilemma. For the first time, static and dynamic cues become separable.



[FIG3] Illustration of the videoconference paradigm. (a) Video of the source person; (b) AAM tracking of the source person; (c) AAM reconstruction that is viewed by the naive participant; d) video of the naive participant. (Image reused from Figure 5 in [12] with permission from the Royal Society.)



[FIG4] Applying expressions of a male to the appearances of other persons. In (a), the avatar has the appearance of the person whose motions were tracked. In (b) and (c), the avatars have the same–sex appearance. Parts (d)–(f) show avatars with opposite–sex appearance. (Image courtesy of the APA.)

Pairs of participants had conversations in a video-conference paradigm. One was a confederate for whom an AAM had previously been trained. Unbeknownst to the other participant, a resynthetized avatar was substituted for the live video of the confederate (Figure 3) [12]. The avatar had the face of the confederate or another person of same or opposite sex. All were animated by the actual motion parameters of the confederate (Figure 4).

The apparent identity and sex of a confederate was randomly assigned and the confederate was blind to the identity and sex that they appeared to have in any particular conversation. The manipulation was believable in that, when given an opportunity to guess the manipulation at the end of experiment, none of the naive participants was able to do so. Significantly, the amplitude and velocity of head movements were influenced by the dynamics (head and facial movement and vocal timing) but not the perceived sex of the partner.

These findings suggest that gender-based social expectations are unlikely to be the source of reported gender differences in head nodding between partners. Although men and women adapt to each other's head movement amplitudes it appears that adaptation may simply be a case of people (independent of sex) adapting to each other's head movement amplitude. A shared equilibrium is formed when two people interact.

These results are consistent with a hypothesis of separate perceptual streams for appearance and biological motion. Head movements generated during conversation respond to dynamics but not appearance. In a separate perceptual study, we found that judgments of sex were influenced by appearance but not dynamics. Judgments of masculinity and femininity were more complicated; appearance and dynamics each contributed. This dissociation of the effects of appearance and dynamics is difficult to explain without independent streams for appearance and biological motion.

CONCLUSION AND FUTURE DIRECTIONS

Significant progress has been made in developing and applying AFAS to behavioral science applications. In several relatively challenging data sets, we have detected socially relevant AUs, expressions, and behavioral states (e.g., pain and depression) of theoretical, developmental, and clinical import. Beginning with the Spectrum database, clinical trials have begun.

An ongoing effort is to make "automated" measurement more automated and to use it to extend human performance. As mentioned previously, AAMs can be used for both analysis (i.e., estimating nonrigid shape and appearance) and synthesis. The analysis process is commonly referred to as "tracking" in the computer vision literature as it is estimating shape and appearance measurements from a temporal image sequence. The AAM tracking approach we have exploited requires that about 3% of video be hand annotated to train person-specific models. While this is feasible for research use, clinical and many affective computing and social signal processing applications seek approaches that can work "out of the box." While person-independent (i.e., generic) alternatives to AAM tracking have been proposed (most notably constrained local models (CLMs) [13], they lack the relatively precise shape estimation possible with person-specific AAM. When face-shape tracking loses precision, the information value of shape and many appearance features degrades ungracefully. An exciting development is registration invariant representations (e.g., Gabor magnitudes) to provide a way out of this dilemma. In combination with real-time tracking via more generic but imprecise trackers (e.g., CLM), "AAM-like" expression detection results can be obtained in the presence of noisy dense registration through the use of registration invariant features [14]. These results are limited so far to posed facial actions. Current work is applying CLM and registration invariant features to the more challenging video databases described above.

A fruitful approach is to use AFAS to extend the reach and yield of manual efforts. The research on types of smiles, for instance, illustrates the advantages that accrue by using a combination of new and traditional tools for video analysis. Our group has shown that manual coding time can be reduced by over half when AFAS is used in combination with manual FACS coding. FACS coders code only AU peaks; AFAS then automatically finds the onsets and offsets of each AU. Systems that enhance user expertise and efficiency by automated facial video analysis are coming online.

The real-time, veridical animation capabilities of AFAS are an especially exciting research direction. As noted above, for the first time it is possible to experimentally separate appearance and motion and to manipulate temporal dynamics as well. The finding that head nodding is regulated by dynamics rather than by the partner's evident gender informs theory in embodied cognition.

Correlational evidence suggests that specific facial expressions (e.g., the Duchenne smile) have unique signal value. Until now, experimental tests of this hypothesis have not been possible. Using the communication avatar, facial displays might be manipulated on the fly to test their functional role. Initial work toward this goal has already begun. We found that attenuating facial expression and head nods in the avatar caused a reciprocal increase in expressiveness by the partner. Age and ethnicity are among other variables to explore with this approach, as well as extensions to video conferencing and electronic entertainment.

In summary, AFAS is well advanced, making contributions to a range of topics in emotion, developmental, and social psychology, pain, and human social dynamics.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. National Institutes of Health MH 51435, U.S. National Science Foundation BCS05 27397, and the Canadian Institutes of Health Research CIHR145703. The author thanks Fernando De la Torre and Simon Lucey for their comments on an earlier version of this article.

AUTHOR

Jeffrey F. Cohn (jeffcohn@cs.cmu.edu) is a professor of psychology at the University of Pittsburgh and adjunct faculty at the Robotics Institute at Carnegie Mellon University.

REFERENCES

[1] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: Research Nexus*. Salt Lake City, UT: Network Research Information, 2002.

[2] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 25, no. 9, pp. 1063–1074, Sept. 2003.

[3] T. F. Cootes, G. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 23, no. 6, pp. 681–685, 2001.

[4] I. Matthews, J. Xiao, and S. Baker, "2D vs. 3D deformable face models: Representational power, construction, and real-time fitting," *Int. J. Comput. Vis.,* vol. 75, no. 1, pp. 93–113, 2007.

[5] M. Frank, J. Movellan, M. S.Bartlett, and G. Littlewort in RU-FACS-1 Database, Machine Perception Lab., Univ. California, San Diego.

[6] J. F. Cohn and M. A. Sayette, "Spontaneous facial expression in a small group can be automatically measured: An initial demonstration," *Behav. Res. Methods,* to be published.

[7] J. F. Cohn and M. A. Sayette, "Detecting depression from facial actions and vocal prosody," in *Proc. Affective Computing and Intelligent Interaction (ACII'09),* Sept. 2009, pp. 1–7.

[8] Z. Ambadar, J. F. Cohn, and L. I. Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *J. Nonverbal Behav.,* vol. 33, no. 1, pp. 17–34, 2009.

[9] D. S. Messinger, M. H. Mahoor, S. M. Chow, and J. F. Cohn, "Automated measurement of facial expression in infant-mother interaction: A pilot study," *Infancy,* vol. 14, no. 3, pp. 285–305, 2009.

[10] P. Lucey, J. F. Cohn, S. Lucey, S. Sridharan, and K. Prkachin "Automatically detecting pain using facial actions," in *Proc. Affective Computing and Intelligent Interaction (ACII'09),* 2009, pp. 12–18.

[11] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, K. M. Prkachin, and P. Solomon, "The painful face: Pain expression recognition using active appearance models," *Image Vis. Comput.,* vol. 27, no. 12, pp. 1788–1796, 2009.

[12] S. M. Boker, J. F. Cohn, B. J. Theobald, I. Matthews, J. Spies, and T. Brick, (2009). "Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars," *Philos. Trans. B Roy. Soc.*, vol. 364, pp. 3485–3495.

[13] S. Lucey, Y. Wang, M. Cox, S. Sridharan, and J. F. Cohn, "Efficienct constrained local model fitting for non-rigid face alignment," *Image Vis. Comput. J.,* vol. 27, no. 12, pp. 1804–1813, 2009.

[14] S. Lucey, P. Lucey, and J. F. Cohn, "Registration invariant representations for expression detection," in *Proc. Int. Conf. Digital Image Computing: Techniques and Applications (DICTA),* 2010.