Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression

Jeffrey F. Cohn
Department of Psychology
University of Pittsburgh
4015 O'Hara Street, Pittsburgh PA 15260
jeffcohn@vms.cis.pitt.edu

Adena J. Zlochower
Department of Psychology
University of Pittsburgh
4015 O'Hara Street, Pittsburgh, PA 15260
adena@vms.cis.pitt.edu

James J. Lien
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, PA 15260
jjlien@cs.cmu.edu

Takeo Kanade
Departments of Computer Science and
Electrical Engineering, Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
tk@cs.cmu.edu

Abstract

Current approaches to automated analysis have focused on a small set of prototypic expressions (e.g., joy or anger). Prototypic expressions occur infrequently in everyday life, however, and emotion expression is far more varied. To capture the full range of emotion expression, automated discrimination of fine-grained changes in facial expression is needed. We developed and implemented an optical-flow based approach (feature point tracking) that is sensitive to subtle changes in facial expression. In image sequences from 100 young adults, action units and action unit combinations in the brow and mouth regions were selected for analysis if they occurred a minimum of 25 times in the image database. Selected facial features were automatically tracked using a hierarchical algorithm for estimating optical flow. Images sequences were randomly divided into training and test sets. Feature point tracking demonstrated high concurrent validity with human coding using the Facial Action Coding System (FACS).

1. Introduction

The face is an important source of information about human behavior. Facial displays express emotion [7], influence interpersonal behavior [5], reflect brain function and pathology [15] and reveal changes with development in children (e.g., [20]). To make use of the information afforded by facial expression, reliable, valid, and efficient methods of measurement are critical.

Computer-vision based approaches to facial expression analysis (e.g., Black and Yacoob [2]; Mase [14]) discriminate between a small set of emotions. This focus follows from the work of Darwin [6] and more recently Ekman [7], who proposed that "basic emotions" (i.e., joy, surprise, anger, sadness, fear, and disgust) each have a prototypic facial expression, involving changes in facial features in multiple regions of the face, which facilitates analysis. In everyday life, prototypic expressions may occur relatively infrequently, and emotion more often is communicated by changes in one or two discrete features, such as tightening the lips, which may communicate anger [4]. To capture the subtlety of human emotion and non-verbal communication, automated discrimination of fine-grained changes in facial expression is needed.

The Facial Action Coding System (FACS) [8] is a human-observer-based system designed to detect subtle changes in facial features. FACS consists of 44 anatomically based "action units," which individually or in combinations can represent all visibly discriminable expressions. Seminal work by Mase [14], Pentland, and

Essa [10] suggested that FACS action units could be detected from differential patterns of optical flow. Essa and Pentland [10] found increased flow in muscle regions associated with action units in the brow and in the mouth region. The specificity of optical flow to action unit discrimination, however, was not tested. Discrimination of facial expression remained at the level of emotion prototypes rather than the finer level action units. Bartlett et al. [1] discriminated between action units in the brow and eye regions. The number of subjects was small (10 each in the training and test samples), and the image data received extensive manual pre-processing.

Current methods of estimating optical flow may lack sensitivity to subtle motion, which is needed to discriminate expressions at this more fine-grained level. Relatively large feature regions (e.g., mouth or cheeks) are used, and flow direction is changed to conform to the plurality of flow [2, 17, 19] or average flow [13, 14] within the region. Black and colleagues [2, 3] also assign parameter thresholds to their classification paradigm. Information about small deviations is lost when the flow pattern is removed or thresholds are imposed. As a result, the recognition ability and accuracy of the systems for subtle feature changes may be reduced.

We developed and implemented an optical-flow-based approach that overcomes these difficulties. Closely spaced facial feature points within 13x13 pixel windows are tracked by optical flow. Feature points are selected based on two criteria. They are in regions of moderate to high texture and represent underlying muscle activation of closely related actions. The sensitivity and specificity of feature point tracking was evaluating by comparing system performance with that of human FACS coders.

2. Methods

2.1 Image acquisition

Facial behavior was recorded in 100 adults (65% male and 15% African American or Asian, ages 18 to 35 years). Subjects were situated 90 degrees from the image plane of a camera and performed a series of facial expressions that included single action units (e.g., AU 12, or smile) and combinations of action units (e.g., AU 1+2, or brow raise). Each expression sequence began from a neutral face. Six of the expressions were based on descriptions of prototypic emotions. Action units were coded by a certified FACS coder. Inter-observer reliability was quantified with coefficient kappa [21], which corrects for chance agreement. Mean κ was 0.86.

Table 1. FACS action units (AU).

Action Unit	Description
Brows	
AU 1+2	Inner and outer portions of the brows are raised
AU 1+4	Medial portion of the eyebrows is raised and pulled together
AU 4	Brows are lowered and drawn together
Eyes	
AU 5	Upper eyelids are raised, which produces a widening of the eyes
AU 6	The lower-eye and infra-orbital furrows are raised and deepened and the eye opening is narrowed
AU 7	Lower eyelids are tightened, which narrows the eye opening
Mouth	
AU 27	Mouth is stretched open and mandible extended
AU 26	Lips are relaxed and parted; mandible lowered
AU 25	Lips are relaxed and parted; mandible not lowered
AU 12	Lip corners are pulled up and backward
AU 12+25	AU 12 with mouth opening
AU 20+25	Lips are parted, pulled back laterally, and may be slightly raised or pulled down.
AU 15+17	Lip corners are pulled down and stretched laterally (AU 15), and chin boss is raised, which pushes up the lower lip (AU 17).
AU 17+23+24	AU 17 and the lips are tightened, narrowed, and pressed together (AU 23+24)
AU 9+17±25	The infra-orbital triangle and center of the upper lip are pulled upwards (AU 9) with AU 17. In 25% of cases, AU 9+17 occurred with AU 25.

Action units in the brow and mouth regions were selected for analysis if they occurred a minimum of 25 times. When an action unit occurred in combination with other action units that may modify its appearance, the combination rather than the single action unit was the unit of analysis. The action units and action unit combinations

within each facial region that met this criterion are shown in Table 1. The action units we analyzed represent key components of emotion and related paralinguistic displays. In each region, the actions chosen included similar appearance changes (e.g., brow narrowing due to AU 1+4 versus AU 4 and mouth widening due to AU 12 versus AU 20.).

2.2 Image alignment

Image sequences (from neutral to peak expression) were digitized into 640 by 490 pixel arrays (mean duration ~ 20 frames at 30 frames per second). Because subjects produced little out of-of-plane motion, an affine transformation was adequate to normalize face position and maintain face magnification invariance. The position of all feature points was normalized by automatically mapping them to a standard face model based on three facial feature points: the medial canthus of both eyes and the uppermost point of the philtrum (Figure 1).
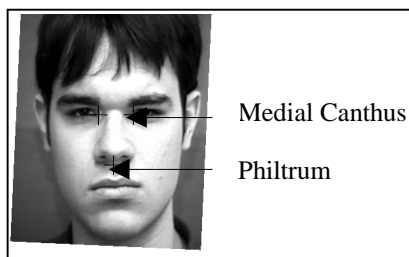
Figure 1. Standard face model.

2.3 Facial feature point tracking

In the first digitized frame, key feature points were manually marked with a computer-mouse around facial landmarks (Figure 2).
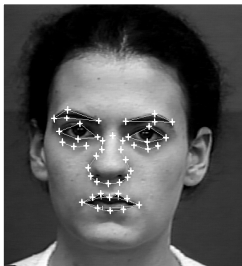

Figure 2. Feature point marking.

Each point is the center of a 13x13-flow window that includes horizontal and vertical flows. A hierarchical optical flow method [12] is used to automatically track feature points in the image sequence. The displacement of

each feature point is calculated by subtracting its normalized position in the first frame from its current normalized position. The resulting flow vectors (6 horizontal and vertical dimensions in the brow region, 8 horizontal and vertical dimensions in the eye region, 6 horizontal and vertical dimensions in the nose region, and 10 horizontal and vertical dimensions in the mouth region) are concatenated to produce a 12 dimensional displacement vector in the brow region, a 16-dimensional displacement vector in the eye region, a 12 dimensional displacement vector in the nose region, and a 20 dimensional vector in the mouth region (Figure 3).
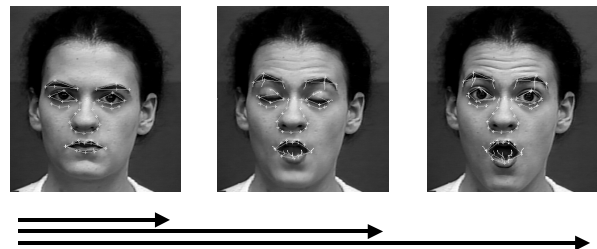

Figure 3. Feature point displacement.

Figure 3 show a sequence in which the subject's expression changes from neutral (AU 0) to brow raise, eye widening, and mouth stretched wide open (AU 1+2+5+27), which is characteristic of surprise. The feature points are precisely tacked across the image sequence. Lines trailing from the feature points represent change in the location of feature points due to expression. The length of the lines corresponds to strength of action unit intensity. As the action units become more extreme, feature point displacement as indicated by line length becomes greater.

2.4 Action unit discrimination

The database consisted of 504 image sequences containing 872 action units or action unit combinations from 100 subjects. Separate discriminant function analyses (DFA) were conducted within each facial region. Predictors were feature point displacement between the initial and peak frames in the image sequence. *A priori* probabilities of actions units were assumed to be equal. Because the primary goal was classification and not evaluating the relative importance of individual feature point displacements, direct entry of predictors was used. Separate group variance-covariance matrices were used for classification or recognition. Data were randomly divided into training and cross-validation or test sets. Classification accuracy was quantified with kappa (κ) coefficients, which correct for chance agreement.

Classification accuracy did not vary with subjects' race or gender.

With three action units or action unit combinations in the brow region (AU 1+2, AU 1+4, and AU 4), there were two possible discriminant functions. Wilk's lambda and both discriminant functions were highly significant (λ = .06, $p < .001$, canonical correlations = .94 and .69, $p < .001$). In the test set (Table 2), 92% were correctly classified (κ= 0.86); accuracy ranged from 86% for AU 1+4 to 91% and 95% for AU 4 and AU 1+2, respectively.

Table 2. Proportion agreement in the brow region.

Human Coding		Feature Point Tracking			
		N	AU 1+2	AU 1+4	AU 4
AU 1+2		43	**.95**	.02	.02
AU 1+4		22	.05	**.86**	.09
AU 4		65	.02	.08	**.91**
κ = .86					

With three action units in the eye region (AU 5, AU 6, and AU 7), there were two possible discriminant functions. Wilk's Lambda and both functions were highly significant (λ = 0.07, $p < .001$; canonical correlations = .92 and .74, $p < .001$). In the test set (Table 3), 88% were correctly classified (κ = 0.81). Disagreements that occurred were between AU 6 and AU 7.

Table 3. Proportion agreement in the eye region.

Human Coding		Feature Point Tracking			
		N	AU 5	AU 6	AU 7
AU 5		33	**.97**	.00	.03
AU 6		36	.06	**.81**	.14
AU 7		20	.00	.15	**.85**
κ = .81					

Nine action units were analyzed in the nose and mouth regions. Wilk's Lambda, with five significant discriminant functions, was 0.0003 (canonical correlations = .95, .93, .89, .79, and .66, $p < .001$). In the test set 83% were correctly classified (κ = 0.81). With the exception of AU 26, accuracy for action units ranged from 78% to 100% (Table 4).

3. Discussion

Previous studies have used optical flow to recognize facial expression [10, 18]. Sample sizes in these studies have been small, and with the exception of Bartlett et al. [1], this work has focused on the recognition of molar expressions, such as positive or negative emotion or emotion prototypes (e.g., joy, surprise, fear). We developed and implemented an optical flow based approach that is sensitive to subtle motion in facial displays. Feature point tracking was tested in images sequences from 100 subjects and achieved a level of precision that was as high as or higher than that of previous studies and comparable to that of human coders.

Accuracy in the test sets was 92% in the brow region, 88% in the eye region, and 83% in the nose and mouth regions. The one previous study to demonstrate accuracy for discrete facial actions [1] used extensive pre-processing of image sequences and was limited to upper face action units in 10 subjects. In the present study, pre-processing was limited to manual marking with a pointing device in the initial digitized image, facial behavior included action units in both the upper and lower face, and the large number of subjects, which included African-Americans and Asians in addition to Caucasians, provided a sufficient test of how well the initial training analyses generalized to new image sequences. Feature point tracking demonstrated moderate to high concurrent validity with human FACS coding.

The level of inter-method agreement for action units was comparable to that achieved in tests of inter-observer agreement in FACS. Moreover, the inter-method disagreements that did occur were generally ones that are common in human coders, such as the distinction between AU 1+4 and AU 4 and AU 6 and AU 7. These findings suggest that the two methods are at least equivalent for the type of image sequences and action units analyzed here. In future work, feature point tracking will be tested in longer image sequences and ones involving spontaneous displays of emotion.

One reason for the lack of 100% agreement at the level of action units is the inherent subjectivity of human FACS coding, which attenuates the reliability of human FACS codes. Two other possible reasons were the restricted number of optical flow feature windows and the reliance on a single computer vision method. We have not yet implemented optical flow estimation in many feature windows, including the forehead, glabella, infra-orbital furrow, cheeks, the area above the lips, and the chin boss. Preliminary findings [11] suggest that algorithms for dense flow [18] optimize recognition accuracy in these regions.

Many action units involve changes in transient features, such as lines or furrows, that may occur or vary across an image sequence. "Crows-feet" wrinkles, for instance, form at the eye corners from contraction of the orbicularis oculi in AU 6, and recognition of AU 5 is assisted by identifying increases in the sclera above the eyeball. These features are represented by intensity gradients in the image sequence and are quantified by the computer vision method of edge detection. For some action units, the use of edge detectors should prove essential. To discriminate between AU 25 and AU 26, FACS specifies a requisite distance between upper and lower teeth, which is readily detected by edge detectors but not by optical flow. By increasing the number of search regions and supplementing optical flow estimation with edge detection, further improvement in facial feature analysis can be achieved.

In human communication, the timing of a display is an important aspect of its meaning. For example, the duration of a smile distinguishes between felt and false positive emotion. Smiles that last too long may communicate aggression. To evaluate the validity of feature point tracking in the spatio-temporal domain, we have begun using Hidden Markov Models (HMM) [16] which are widely used in speech and gesture recognition. Preliminary results based on 36 test sequences in the brow region and 60 test sequences in the mouth region are consistent with those of the DFA.

In summary, feature point tracking was sensitive to subtle changes in facial features and discriminated facial expression at the fine-grained level of individual action units. The focus of current work is to incorporate convergent methods of quantifying facial expression, increase the number of action units and action unit combinations that can be recognized, and increase the generalizability of the system to a wide range of image orientations

Acknowledgement

References

[1] M.S. Bartlett, P.A. Viola, T.J. Sejnowski, B.A. Golomb, J. Larsen, J.C. Hager, and P. Ekman, "Classifying Facial Action,". In D. Touretski, M. Mozer, and M. Hasselmo (Eds.), Advances in Neural Information Processing Systems 8 (pp. 823-829). MIT, Cambridge, MA, 1996.

[2] M.J. Black and Y. Yacoob, "Recognizing Facial Expressions under Rigid and Non-Rigid Facial Motions," International Workshop on Automatic Face and Gesture Recognition, pp. 12-17, Zurich, 1995.

[3] M.J. Black, Y. Yacoob, A.D. Jepson, and D.J. Fleet, "Learning Parameterized Models of Image Motion, Computer Vision and Pattern Recognition, 1997.

[4] J.M. Carroll and J.A. Russell, "Facial Expressions in Hollywood's Portrayal of Emotion," Journal of Personality and Social Psychology, 72:164-176, 1997.

[5] J.F. Cohn and M. Elmore, "Effect of Contingent Changes in Mothers' Affective Expression on the Organization of Behavior in 3-Month-Old Infants, " Infant Behavior and Development, 11:493-505, 1988.

[6] C. Darwin, "The Expression of Emotion in Man and Animals, " University of Chicago, 1872/1965.

[7] P. Ekman, "Facial Expression and Emotion," American Psychologist, 48:384-392, 1993.

[8] P. Ekman and W.V. Friesen, "Facial Action Coding System," Consulting Psychologist Press, Palo Alto, CA, 1978.

[9] I.A. Essa, "Analysis, Interpretation and Synthesis of Facial Expressions," Perceptual Computing Technical Report 303, MIT Media Laboratory, February 1995.

[10] I.A. Essa and A. Pentland, "A Vision System for Observing and Extracting Facial Action Parameters," IEEE Computer Vision and Pattern Recognition, 1994.

[11] J.J. Lien, A.J. Zlochower, J.F. Cohn, C.C. Li, and T. Kanade, "Automatically Recognizing Facial Expressions in the Spatio-Temporal Domain," Proceedings of the Workshop on Perceptual User Interfaces, pp. xxx-xxx, 1997.

[12] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique With an Application in Stereo Vision," Seventh International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.

[13] K. Mase and A. Pentland, "Automatic Lipreading by Optical-Flow Analysis," Systems and Computers in Japan, 22(6), 1991.

[14] K. Mase, "Recognition Of Facial Expression from Optical Flow," IEICE Transactions, E74:3474-3483, 1991.

[15] W.E. Rinn, "The Neuropsychology of Facial Expression: A Review of the Neurological and Psychological Mechanisms for Producing Facial Expressions," Psychological Bulletin, 95:52-77, 1984.

[16] L.R. Rabiner, "An introduction to Hidden Markov Models," IEEE ASSP Magazine, 4-16, January 1986.

[17] M. Rosenblum, Y. Yacoob, and L.S. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture," Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects, Austin, TX, November 1994.

[18] Y.T. Wu, T. Kanade, J.F. Cohn, and C.C. Li., "Optical Flow Estimation Using Wavelet Motion Model," Proceedings of the International Conference on Computer Vision (ICCV), 1998.

[19] Y. Yacoob and L. Davis, "Computing Spatio-Temporal Representations of Human Faces," In Proc. Computer Vision and Pattern Recognition, pp. 70-75, Seattle, WA, June 1994.

20] R.N. Emde, T.J. Gaensbauer, & R.J. Harmon, "Emotional Expression in Infancy: A Biobehavioral Study," Psychological Issues, 10: 37, 1996.

[21] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, 20: 37-46, 1960.

Table 4. Proportion agreement in the nose and mouth regions.

Human		Feature Point Tracking								
	N	27	26	25	12	12+25	20+25	15+17	17+23+24	9+17
27	29	**.83**	.10	.03	.00	.00	.03	.00	.00	.00
26	24	.25	**.54**	.21	.00	.00	.00	.00	.00	.00
25	22	.00	.05	**.86**	.00	.00	.00	.09	.00	.00
12	18	.00	.00	.00	**.78**	.22	.00	.00	.00	.00
12+25	35	.00	.00	.03	.00	**.86**	.11	.00	.00	.00
20+25	31	.00	.00	.00	.00	.16	**.81**	.03	.00	.00
15+17	36	.00	.00	.00	.00	.00	.00	**.94**	.06	.00
17+23+24	12	.00	.00	.00	.00	.00	.00	.08	**.92**	.00
9+17	17	.00	.00	.00	.00	.00	.00	.00	.00	**1.00**
$\kappa = .81$										