Recognizing Human Facial Expressions From Long Image Sequences Using Optical Flow

Yaser Yacoob and Larry S. Davis

Abstract—An approach to the analysis and representation of facial dynamics for recognition of facial expressions from image sequences is presented. The algorithms utilize optical flow computation to identify the direction of rigid and nonrigid motions that are caused by human facial expressions. A mid-level symbolic representation motivated by psychological considerations is developed. Recognition of six facial expressions, as well as eye blinking, is demonstrated on a large set of image sequences.

Index Terms—Face expression recognition, non-rigis motion analysis, optical flow, tracking.

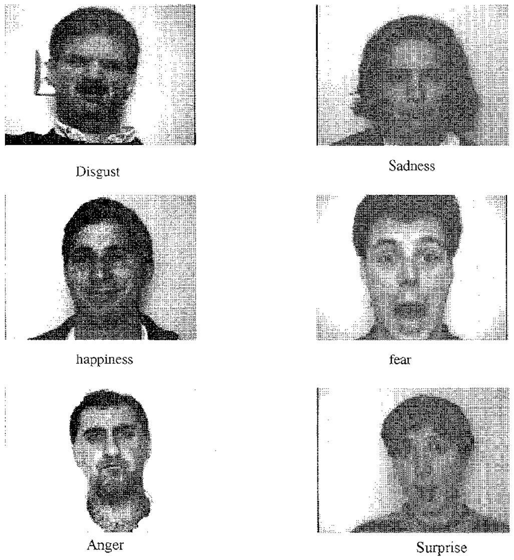
1 Introduction

VISUAL communication plays a central role in human communication and interaction. This paper explores methods by which a computer can recognize visually communicated facial actions–facial expressions. Such methods can contribute to human-computer interaction and to applications such as: video facial image queries and low-bandwidth transmission of facial data. In human-computer interaction, recognition of mood, concentration and mood-accompanied speech will be required for keyboardless computing and user affect-dependent software (e.g., educational software). Video facial image queries can be useful when querying content by actions of people in the image stream. Low bandwidth transmission of facial data can be made more efficient by using mid- and high-level visual representation of the facial actions (e.g., send a smile and a few parameters that determine the mouth actions involved).

Visual communication has been extensively studied in the psychology literature, mainly as a means of describing the emotional, cognitive and physical states of subjects and the role they play in social interactions [17]. Research in psychology has indicated that at least six expressions of emotion are universally associated with distinct facial expressions. Several other expressions, and many combinations of expressions, have been studied but remain unconfirmed as universally distinguishable. The six principal expressions of emotion are: happiness, sadness, surprise, fear, anger, and disgust (see Fig. 1).

Ekman and Friesen [8] developed the most comprehensive system for synthesizing facial expressions based on what they call Action Units (the Facial Actions Coding System-FACS). Each AU may correspond to several muscles that together bring about a certain facial action. The FACS model has been used to synthesize images of facial expressions, but only limited exploration of its use in analysis has been performed [10], [11], [12], [15].

Most psychological research on facial expressions has been conducted on static "mug-shot" pictures that capture the subject's expression at its peak [17]. Few studies have directly investigated the influence of the motion and deformation of facial features on the recognition of facial expressions. Bassili [2] showed that motion in the image of a face would allow expressions to be identified even with minimal information about the spatial arrangement of features. His subjects viewed image sequences in which only white dots on the darkened surface of the face displaying the expression were visible (see Fig. 2). Notice that the face features, texture and complexion were unavailable to the subjects. The figure also shows the principal facial motions that provide cues to the recognition of facial expressions. Bassili's experimental results indicate that facial expressions were more accurately recognized from dynamic images than from a single static "mug-shot" image.



Disgust

Sadness

happiness

fear

Anger

Surprise

Fig. 1. Six universal (i.e., pancultural) expressions expressed by six faces.

The problem of recognizing facial expressions has recently attracted attention in the computer vision community [3], [9], [11], [12], [13], [14], [15]. The work of [9], [12], [14] is most closely related to ours since they use optical flow computation for recognizing and analyzing facial expressions. Mase approached facial expression recognition from both the top-down and bottom-up directions. In both cases, the focus was on computing the motions of facial *muscles* rather than the motions of facial *features*. Four facial expressions were studied: surprise, anger, happiness, and disgust.

Essa and Pentland [9] and Essa [10] recently proposed a physically based approach for modeling and analyzing facial expressions. They proposed extending the FACS model to the temporal dimension (thus calling it FACS+) to allow combined spatial and temporal modeling of facial expressions. They assumed that a mesh is originally overlayed on the face, then tracked all its vertices based on the optical flow field throughout the sequence. The emphasis is on accuracy of capturing facial changes, which is most essential to synthesis. Recognition results were reported in [10] on six subjects displaying four expressions and eyebrow raising.

2 The Proposed Architecture

Before proceeding, we introduce some terminology that is needed in the paper. Face region *motion* refers to changes in images of facial features caused by facial *actions* corresponding to physical feature deformations on the 3D surface of the face. The goal is to

- *The authors are with the Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, MD 20742-3275. E-mail: yaser@cs.umd.edu and lsd@umiacs.umd.edu.*

Manuscript received Oct. 27, 1994; revised Dec. 11, 1995. Recommended for acceptance by R.M. Bolle.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P96026.

0162-8828 /96$05.00 ©1996 IEEE

develop computational methods that use face region motions as *cues* for action recovery.

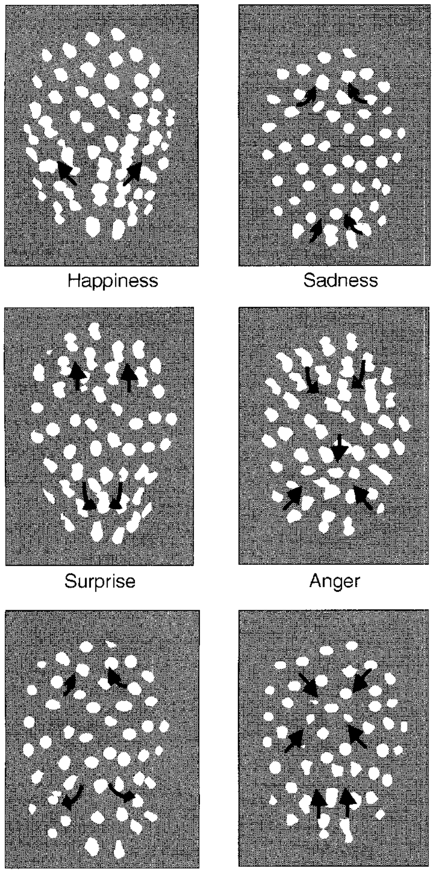


Fig. 2. The cues for facial expression as suggested by Bassili [2].

The following is the framework within which our approach for analysis and recognition of facial expressions is developed:

- The face is viewed from a near frontal view throughout the sequence.
- The overall rigid motion of the head is small between consecutive frames.
- The nonrigid motions resulting from face deformations are spatially bounded, in practice, by an $n \times n$ window between any two consecutive frames.
- We consider only the six universal expressions—happiness, sadness, anger, fear, disgust and surprise—and blinking.

Fig. 3 describes the flow of computation of our facial expression system.

3 Tracking Face Regions

The approach we use for optical flow computation is a correlation approach recently proposed by Abdel-Mottaleb et al. [1]. Our choice stems from its reliability when only small subpixel motions occur, and employing minimal assumptions on the structure of the image and the motions occuring.

Locating the facial features in an image has been extensively addressed in the past five years [5], [16], [18]. An extensive review of this subject appears in [6]. In our work we simply assume that rectangles enclosing the facial features have been initially placed before the onset of tracking.

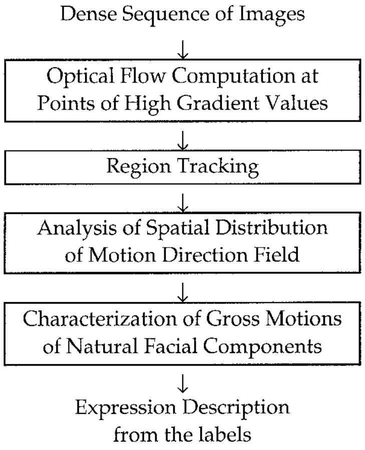

Dense Sequence of Images
Optical Flow Computation at Points of High Gradient Values
Region Tracking
Analysis of Spatial Distribution of Motion Direction Field
Characterization of Gross Motions of Natural Facial Components
Expression Description from the labels

Fig. 3. The flow of the facial analysis algorithm.

Since our approach is based on a statistical characterization of the motion patterns in specified regions of the face, and not in tracking individual point motions, we develop a region tracker for rectangles enclosing the face features. Each rectangle encloses one feature of interest, so the flow computation within the region is not "contaminated" by the motions of other facial features. To simplify the modeling of the eyebrows, we define the rectangles to include the eyes, and then subtract the rectangle of the eye from the combined rectangle.

The tracking algorithm integrates spatial and temporal information at each frame. The former is based on the gradient magnitude of the intensity image and the latter is based on the optical flow field. The spatial tracking of the face regions is based on computing two sets of parameters at the points with high gradient values within the rectangle that encloses each feature. We assume that the gradient image is relatively stable between any two images (a reasonable assumption considering that the face is sampled at 30Hz and the lighting is stable). The following are computed from frame i after placing the rectangles from frame $i - 1$ over the image in frame i:

- The centroid $\left(C_x^i, C_y^i\right)$ of the points having a high gradient value within each rectangle in frame i.
- The window

$$W = \left(WX_{min}^i - 2,\ WY_{min}^i - 2,\ WX_{max}^i + 2,\ WY_{max}^i + 2\right)$$

  which encloses those high gradient values and leaves a buffer, two pixels deep, that allows the detection of window expansion during subsequent iterations.

The centroid's location determines the translation of the rectangle from the previous frame. The window W determines the scaling of the rectangle.

In order to enhance the tracking the statistics of the motion directions within a rectangle are used to verify translation of rectangles upward and downward (by measuring significant similar optical flow) and verify scaling of the rectangles (by measuring motions that imply scaling).

In spite of the simplicity of the tracking algorithms, it was quite robust, routinely tracking face features for hundreds of frames with no manual intervention. In more recent work [4], Black and Yacoob report a new robust approach for tracking facial features while undergoing significant rigid and nonrigid motions. Their approach models the motion of the head as a plane moving in 3D, while allowing each feature to undergo extended affine deformations.

TABLE 1
THE DICTIONARY FOR MOUTH MOTIONS

Component	Basic Action	Motion Cues
upper lip	raising	upward motion of an upper part of window
	lowering	downward motion of an upper part of window
	contraction	horizontal shrinking of an upper part of window
	expansion	horizontal expansion of an upper part of window
lower lip	raising	upward motion of a lower part of window
	lowering	downward motion of a lower part of window
	contraction	horizontal shrinking of a lower part of window
	expansion	horizontal expansion of a lower part of window
left corner	raising	upward motion of a left part of window
	lowering	downward motion of a left part of window
right corner	raising	upward motion of a right part of window
	lowering	downward motion of a right part of window
whole mouth	raising	upward motion throughout window
	lowering	downward motion throughout window
	compaction	overall shrinkage in mouth's size
	expansion	overall expansion in mouth's size

4 COMPUTING LOCAL MOTION REPRESENTATIONS

Our dictionary of facial feature actions borrows from the facial cues of universal expression descriptions proposed by Ekman and Friesen [7], and from the motion patterns of expression proposed by Bassili [2]. As a result, we arrive at a dictionary that is a *motion-based feature description of facial actions*.

The dictionary is divided into *components, basic actions of these components*, and *motion cues*. The components are defined qualitatively and relative to the rectangles surrounding the face regions, the basic actions are determined by the component's visible deformations, and the cues are used to recognize the basic actions using optical flow within these regions. Table 1 shows the components, basic actions, and cues that model the mouth. Similar models were created for the eyes and the eyebrows.

Basic actions cues are computed from the optical flow field as follows. The flow magnitudes are first thresholded to reduce the effect of small motions due to noise. The motion vectors are then requantized into eight principal directions.

The optical flow vectors are filtered using both spatial and temporal procedures that improve their coherence and continuity, respectively. The spatial procedure examines the neighborhood of each point and performs a voting among all neighbors and enforces coherence of its direction label. The temporal procedure follows the spatial procedure; it uses a fixed temporal window to determine the plurality's flow direction, again changing the flow direction at the center of a temporal window if it disagrees with the plurality.

Statistical analyses of the resulting flow directions within each face region window provide indicators about the general motion patterns that the face features undergo. The statistical analyses differ from one feature to another, based on an allowable set of motions. The largest set of motions is associated with the mouth since it has the most degrees of freedom at the anatomic and musculature levels. we use it as an example to illustrate the procedure.

We measure the motion of the mouth by considering a set of vertical and horizontal partitions of its surrounding rectangle. The horizontal partitions are used to capture vertical motions of the mouth. These generally correspond to independent motions of the lips. The two types of vertical partitions are designed to capture several mouth motions. Single vertical partitions capture mouth horizontal expansions and contractions when the mouth is not completely horizontal. The two vertical partitions are designed to capture the motion of the corners of the mouth.

The partitions use free-sliding dividers. For each possible partition, *P*, and for every side of the divider of *P* we define the following parameters:

- m—the total number of points on this side of the divider.
- c^q—the number of points having a motion vector direction $q(q = 1, 2, ..., 8)$ on this side of the divider.
- $p^q = c^q / m$—the percentage of points with motion vectors in direction q.

p^q indicates the degree of clustering of the motion across the divider, while c^q indicates the "strength," in count, of the motion in direction q. The confidence measure of the motion's homogeneity and strength that a divider creates in a direction q is given by (applicable to any type of divider):

$$H^q = c^q \cdot p^q \tag{1}$$

Within each type of partition, partitions are ranked according to the values H^q.

These measurements are used to construct the mid-level representation of a region motion. The highest ranking partition in each type is used as a pointer into the dictionary of motions (see Table 1), to determine the action that may have occurred at the feature. The set of all detected facial actions is used in the following section for recognizing facial expressions.

5 RECOGNIZING FACIAL EXPRESSIONS

5.1 Temporal Considerations for Recognizing Expressions

In the following, we assume that the face's initial expression is neutral.[1] We divide every facial expression into three temporal

[1] The assumption of a neutral expression is not limiting. In an online environment (or when extended viewing is available), it is reasonable to assume that the portion of time a person expresses emotion is short, so the system could examine the relatively long periods of no motion and designate them as neutral.

segments: the *beginning*, the *apex*, and the *ending*. Fig. 4 shows the temporal segments of a smile model. Since the outward-upward motion of the mouth corners are the principal cues for a smile motion, these are used as the criteria for temporal classification also. Notice that Fig. 4 indicates that the detection of mouth corner motions might not occur at the same frames at either the beginning or ending of actions, but it is required that at least one corner starts moving to label a frame with a "beginning of a smile" label, while the motions must terminate before a frame is labeled as an "apex" or an "ending."

We have designed a rule based system that combines some of the expression descriptions from [7] and [2]. Table 2 shows the rules used in identifying the onsets of the "beginning" and the "ending" of each facial expression. These rules are applied to the mid-level representation to create a complete temporal map describing the evolving facial expression. This is demonstrated by an example of detecting a happiness expression. The system identifies the first frame, f_1, with a "raising mouth corners" action, and verifies that the frames following f_1 show a region or basic action that is consistent with this action (in this case, it can be one of the following: right or mouth corner raised, or mouth expansion with/without some opening). It then locates the first frame f_2 where the motions within the mouth region stop occurring (verified with later frames, as before). Then, it identifies the first frame, f_3 , in which an action "lowering mouth corners" is detected and verifies it as before. Finally, it identifies the first frame, f_4 , where the motion is stopped and verifies it. The temporal labeling of the smile expression will have the frames (f_1 ... f_2 − 1), (f_2 ... f_3 − 1), and (f_3 ... f_4) as the "beginning," "apex," and "ending" of a smile, respectively.

Due to errors in the optical flow, some temporal cues need "assistance" from the spatial cues such as the size and shape changes in the rectangles tracking the facial features. Such changes are currently computed only for the mouth since the accuracy of measuring size changes of other face features was not sufficient at the spatial resolution employed. For example, an "anger" expression is characterized by inward lowering motion of the eyebrows and by compaction of the mouth. The compaction may be hard to detect from the optical flow due to noise, aperture, or tracking inaccuracies. We measure the aspect ratios of the window surrounding the mouth during the hypothesized start of an expression, and verify that some compaction in the window size occurred.

5.2 Resolving conflicts between expressions

The system is designed to identify (i.e., to detect an occurrence of an expression) and recognize facial expressions from long video clips (in this case, clips including three to six expressions varied in duration and with neutral expressions in-between). Fig. 5 illustrates the high level architecture of the system. We simplified the behavior model of our subjects by asking them to display one expression at a time, and include an arbitrary duration of a neutral state between expressions. Since the six expression classifiers operate in parallel on the whole sequence the system may create conflicting hypotheses as to the occurrence of facial expressions.

Conflicts may arise when an ending of one expression is confused as the beginning of another expression (quite a likely event in a long sequence containing several expressions). For example, the "anger" recognition module may consider the lowering of the eyebrows during the ending of a "surprise" expression as the beginning of an "anger" expression. To resolve such conflicts, we employ a memory-based process that gives preference to the expression that started earlier, and we also assume that initially (i.e., the first frame) a neutral expression is displayed.

6 Experiments

Our experimental subjects were asked to display expressions without additional directions (in fact, we asked the subjects to

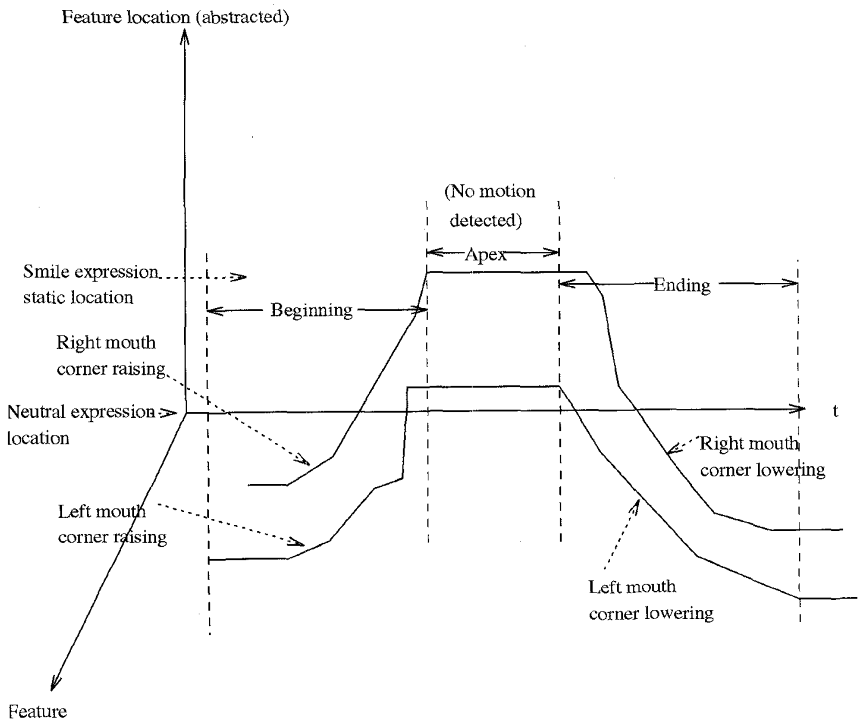

Feature location (abstracted)
(No motion detected)
Apex
Smile expression static location
Beginning
Ending
Right mouth corner raising
Neutral expression location
t
Right mouth corner lowering
Left mouth corner raising
Left mouth corner lowering
Feature

Fig. 4. The temporal model of the "smile" expression.

TABLE 2
THE RULES FOR CLASSIFYING FACIAL EXPRESSIONS

Expression	Beginning or Ending	Satisfactory Actions
Anger	Beginning	inward lowering brows and mouth compaction
Anger	Ending	outward raising brows and mouth expansion
Disgust	Beginning	upward nose motion, mouth expanded/opened and lowering of brows
Disgust	Ending	downward nose motion and raising of brows
Happiness	Beginning	raising mouth corners or mouth opening with its expansion
Happiness	Ending	lowering mouth corners or mouth closing with its contraction
Surprise	Beginning	raising brows and lowering of lower lip (jaw)
Surprise	Ending	lowering brows and raising of lower lip (jaw)
Sadness	Beginning	lowering mouth corners, raising mid mouth and raising inner parts of brows
Sadness	Ending	lowering mouth corners, lowering mid mouth and lowering inner parts of brows
Fear	Beginning	slight expansion and lowering of mouth and raising inner parts of brows
Fear	Ending	slight contraction and raising of mouth and lowering inner parts of brows

choose any subset of expressions and display them in any order and as naturally as they possibly could). As a result, we acquired a variety of presumably similar facial expressions; some were consistent with Ekman and Friesen's dictionary [7] for static images and Bassili's [2] dictionary for motion images, while others varied considerably. This variance can be attributed to the real variance in dynamics and intensities of expressions of individuals as well as to the artificial environment in which the subjects had to develop facial expressions that indicate emotions they were not feeling at the time

Our database of image sequences includes 32 different faces (see Fig. 6). For each face several expressions were recorded each lasting between 15-120 frames of 120 x 160 pixels (at 30 frames per second), some expressions recurring. We requested each subject to display the expressions of emotion in front of the video camera while minimizing his/her head motion. Nevertheless, subjects inevitably moved their head during a facial expression.

In our sample of 46 image sequences of 32 subjects, we had a total of 116 expressions and 106 blinkings (65 percent recognition rate for the latter). The recognition is relative to a ground truth recognition that we established by viewing the video clips and visually identifying the expressions based on the psychology lit-

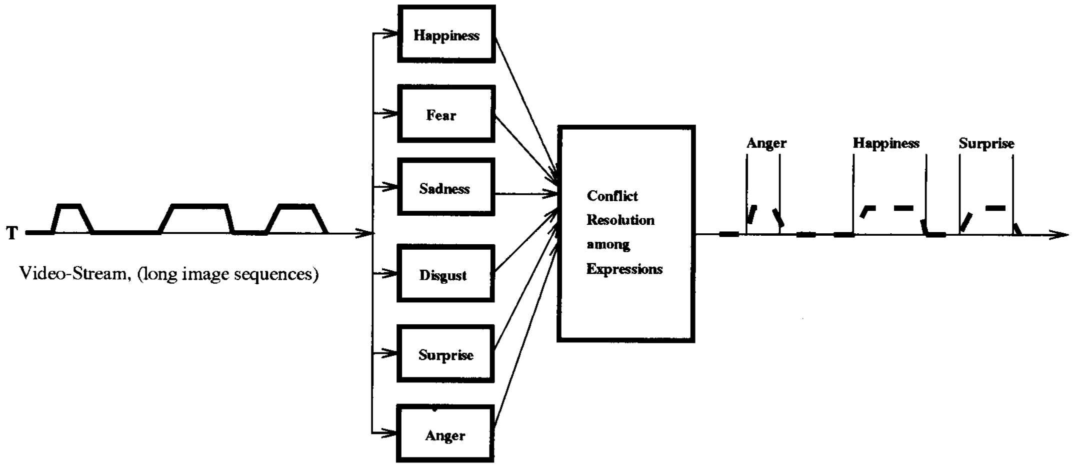

Happiness
Fear
Sadness
Disgust
Surprise
Anger
T
Video-Stream, (long image sequences)
Conflict Resolution among Expressions
Anger
Happiness
Surprise

Fig. 5. The high level flow of the facial analysis system.

TABLE 3
CONFUSION MATRIX FOR FACIAL EXPRESSION RECOGNITION RESULTS

Visual Automatic	Smile	Anger	Surprise	Disgust	Fear	Sadness	Neutral
Smile	32	0	0	0	0	0	0
Anger	0	22	0	0	0	0	0
Surprise	0	0	29	0	2	0	1(RM)
Disgust	1(RM)	0	0	12	0	0	0
Fear	0	0	2	0	6	0	0
Sadness	0	0	0	0	0	4	0
Neutral	5(S, RM, AP, FE)	2(S, FE)	2(RM)	1(FE)	0	1(RM)	

Key: S = Subtle expression, RM = Rigid Motion, AP = Aperture dominated motion, FE = Following closely an expression

erature descriptions (the top horizontal line of Table 3). This ground truth recognition also identified the 'beginning' and 'ending' of expressions, which were found to vary slightly from the automatic system's detected 'beginnings' and 'endings' (measureable optical flow turned out to occur earlier than our visual recognition of such motion—possibly due to different levels of thresholds).

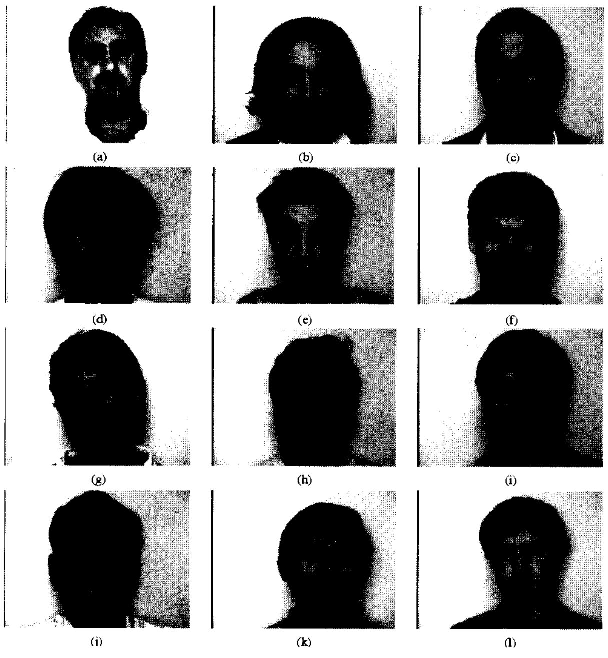


Fig. 6. Twelve faces (out of about 32) used in experiments.

Table 3 shows the details of our results in the form of a confusion matrix (the left column indicates the automatic system recognition and the upper line indicates the ground truth). Occurrences of fear and sadness are less frequent than happiness, surprise and anger. Notice that a 'neutral' category was added (as a "reject" category). Some occurrences of expressions are classified as 'neutral' (i.e., rejected) due to reasons listed in the table.

Some confusion of expressions occurred between the following pairs: fear and surprise, anger and disgust, and sadness and surprise. These distinctions rely on subtle coarse shape and motion information that were not always accurately measured.

Fig. 7 shows four images, the gap between each two images being four frames. For Figs. 7a-7d, the upper left quadrant shows the intensity image, the upper right quadrant shows the gradient image, the rectangle in between displays the classification of facial expression, the lower left quadrant shows the optical flow field, the rectangles around the face regions of interest and the mapping of colors into directions, and the lower right quadrant shows the mid-level descriptions that were computed.

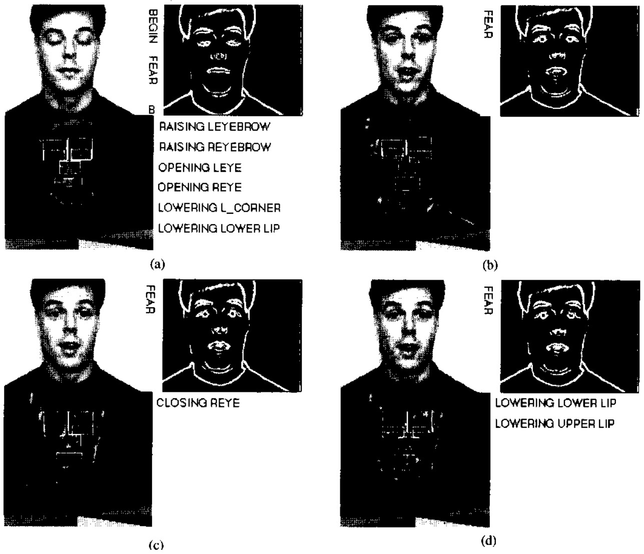


Fig. 7. Four frames analyzed by the facial expression system.

Fig. 7 shows the detection of the beginning of a 'fear' expression; the main cues are the inward raising of the eyebrows and the opening of the mouth. The detection of the inward motion of the eyebrows takes place by analysis of the optical flow field in the rectangles.

With the exception of the computation of optical flow, the algorithms operate at near frame-rates. Since the correlation-based optical flow performs an exhaustive search, its computational cost is high.

7 SUMMARY AND CONCLUSIONS

An approach to analyzing and classifying facial expressions from optical flow was proposed. This approach is based on qualitative tracking of principal regions of the face and flow computation at high intensity gradient points. A mid-level representation is computed from the spatial and the temporal motion fields. The representation is motivated by the psychology literature [2], [7].

We carried out experiments on over 30 subjects in a laboratory environment and achieved good classification of facial expressions on a very large database.

Further study of the system's components will be carried out as well as expanding its capability to deal with nonemotion facial messages. Specifically, the richness of facial expression requires developing more sophisticated representations and capabilities at all levels. At the lowest level, the optical flow and tracking have to be improved to process real video clips that involve rigid, articulated and nonrigid motions of the subject (see [4]). At the mid-level, the representation of actions, spatially and temporally, can be enhanced to capture more of facial behaviors. The incorporation of shape in addition to motion analysis may be useful in refining the representation. At the highest level, there is a need to develop more complex models that are able to capture the diversity of facial actions both from a single subject and across subjects.

Acknowledgments

The support of the Defense Advanced Research Projects Agency (ARPA Order No. 6989) and the U.S. Army Topographic Engineering Center under Contract DACA76-92-C-0009 is gratefully acknowledged.

References

[1] M. Abdel-Mottaleb, R. Chellappa, and A. Rosenfeld, "Binocular Motion Stereo Using MAP Estimation," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 321-327, 1993.

[2] J.N. Bassili, "Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face," *J. Personality and Social Psychology*, vol. 37, pp. 2,049-2,059, 1979.

[3] D. Beymer, A. Shashua, and T. Poggio, "Example Based Image Analysis and Synthesis," M.I.T. A.I. Memo No. 1431, 1993.

[4] M.J. Black and Y. Yacoob, "Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using Local Parametric Models of Image Motions," *Proc. Int'l Conf. Computer Vision*, Boston, pp. 374-381, 1995.

[5] R. Brunelli and T. Poggio, "Face Recognition: Features Versus Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1,042-1,052, 1993.

[6] R. Chellappa, S.A. Sirohey, C.L. Wilson, and C.S. Barnes, "Human and Machine Recognition of Faces: A Survey," Center for Automation Research, Univ. of Maryland Technical Report CAR-TR-731, Aug. 1994.

[7] P. Ekman and W. Friesen, *Unmasking the Face.* Prentice Hall, 1975.

[8] P. Ekman and W. Friesen, *The Facial Action Coding System.* San Francisco: Consulting Psychologists Press, 1978.

[9] I.A. Essa and A. Pentland, "A Vision System for Observing and Extracting Facial Action Parameters," *Proc. IEEE CVPR*, pp. 76-83, 1994.

[10] I.A. Essa, "Analysis, Interpretation, and Synthesis of Facial Expressions," M.I.T. Media Laboratory, Perceptual Computing Group Report No. 303, 1994.

[11] H. Li, P. Roivainen, and R. Forcheimer, "3D Motion Estimation in Model-Based Facial Image Coding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, pp. 545-555, 1993.

[12] K. Mase, "Recognition of Facial Expression from Optical Flow," *IEICE Trans.*, vol. E 74, pp. 3,474-3,483, 1991.

[13] K. Matsuno, C. Lee, and S. Tsuji, "Recognition of Human Facial Expressions Without Feature Extraction," *Proc. ECCV*, pp. 513-520, 1994.

[14] M. Rosenblum, Y. Yacoob, and L.S. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture," *IEEE Workshop Motion of Non-Rigid and Articulated Objects*, Austin, Texas, pp. 43-49, Nov. 1994.

[15] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, pp. 569-579, 1993.

[16] Y. Yacoob and L.S. Davis, "Labeling of Human Face Components from Range Data," *CVGIP: Image Understanding*, vol. 60, no. 2, pp. 168-178, 1994.

[17] *Handbook of Research on Face Processing*, A.W. Young and H.D. Ellis, eds. Elsevier Science Publishers, 1989.

[18] A.L. Yuille, D.S. Cohen, and P.W. Hallinan, "Feature Extraction from Faces Using Deformable Templates," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 104-109, 1989.