Computing Spatio-Temporal Representations of Human Faces

Yaser Yacoob & Larry Davis
Computer Vision Laboratory
University of Maryland
College Park, MD 20742

Abstract

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

1 Introduction

Human visual communication has been extensively studied in the social and psychology literature, mainly as a means to describe the emotional state of the subject [3,5,11,15]. Research in psychology has indicated that at least six emotions are universally associated with distinct facial expressions. Several other emotions, and many combinations of emotions, have been studied but remain unconfirmed as universally distinguishable. The six principle emotions are: happiness, sadness, surprise, fear, anger, and disgust (Figure 1).

Most psychology research on facial expression has been conducted on "mug-shot" pictures that capture the subject's expression at its peak [15]. These pictures allow one to detect the presence of static cues (e.g., wrinkles) as well as the position and shape of the facial features. Few studies have directly investigated the influence of the motion and deformation of facial features on the interpretation of facial expressions. Bassili [2] suggested that motion in the image of a face would allow emotions to be identified even with minimal information about the spatial arrangement of features. The subjects of his experiments viewed image sequences in which only white dots on the dark surface of the person displaying the emotion are visible.

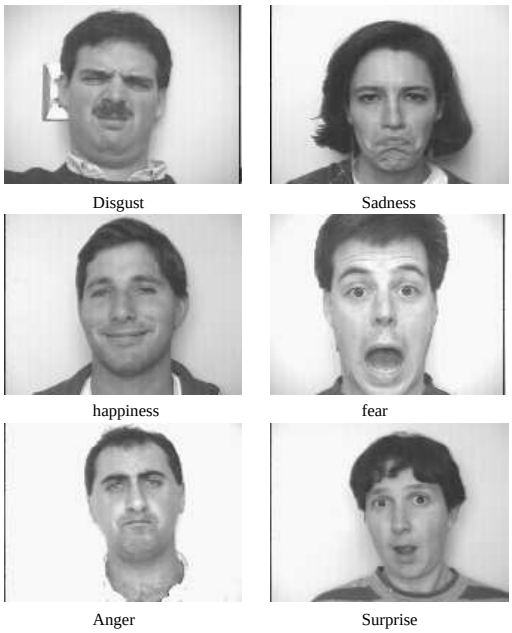

Figure 1: The six universal facial expressions

We provide an expression classifier that employs a representation of facial feature actions. It is based on the descriptions of the epic of facial expressions from static pictures as suggested by Ekman and Friesen in [6], and the descriptions of motion patterns of the face as proposed by Bassili in [2].

We chose not to model or analyze facial muscle actions, setting our work apart from [9,10,12] as well as not to use models for muscle actions [7]. Instead, we focus on the motions associated with the edges of the mouth, eyes, and eyebrows (the relevant considerations appear in [14]).

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

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.

computational methods that relate such motions as *cues* for action recovery.

The following constitute 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 any two consecutive frames.
- The non-rigid motions that are the result of face deformations are spatially bounded, in practice, by an nxn window between any two consecutive frames. The image sequence is densely sampled in time.
- We consider only the six primary emotions - happiness, sadness, anger, fear, disgust and surprise- and eye blinking.

Figure 2 describes a high level flow of computation of our facial expression system.

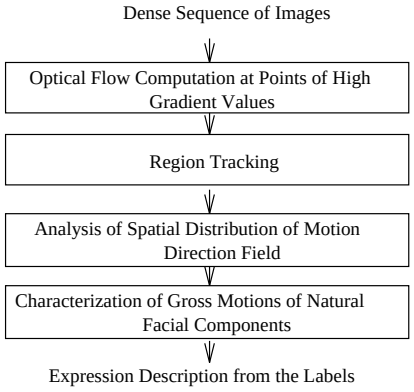

Figure 2: The flow of the facial analysis algorithm

2 Tracking face regions

Approaches for localization of facial features were proposed in [4,13,16], and for facial feature tracking in [12].

Our approach to tracking 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. The following are computed from frame i, f_i, after placing the rectangles from frame $i-1$, f_{i-1}, over the image in f_i:

- The centroid $({C^i}_x, {C^i}_y)$ of the points having a high gradient value within each rectangle in f_i.
- The window $W = ({WX^i}_{min} - 2, {WY^i}_{min} - 2, {WX^i}_{max}+2, {WY^i}_{max}+2)$ which encloses those high gradient values and leaves a buffer, 2 pixels deep, that allows the detection of window expansion during subsequent iterations.

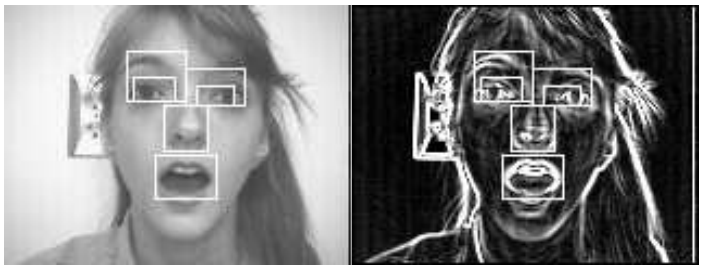

Figure 3: Feature localization and tracking

The centroid's location determines the translation of the rectangle from the previous frame. The window W determines the scaling of the rectangle. The translation and scaling of the rectangles are limited between consecutive images.

In order to enhance the tracking we incorporated the optical flow results in refining the scaling and translations computed by the gradient magnitude change. 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).

3 Computing motion representations

3.1 Psychological basis

A summary of the results of Ekman and Friesen [6] on the universal cues for recognizing the six principle emotions appears in [14]. These cues describe the peak of each expression and thus they provide a human interpretation of the static appearance of the facial feature. For example, a description such as "brows are raised" means that the viewer's interpretation of the location of the brows relative to other facial features indicates they are not in a neutral state but higher than usual. The viewer uses many cues to deduce such information from the image, among these are: the appearance of wrinkles in certain parts of the face, effect of the hypothesis of a high brow on the shape of the eyes (i.e., state of eyelids), etc. Unfortunately, the performance of humans in arriving at such descriptions is far better than what can be achieved, currently, by computers if only static images are considered.

Some of the linguistic expressions used to describe these cues appear very hard to model computationally-"upper lid is tense, tension or stress in the mouth, lip trembling" etc.

Figure 4 summarizes the observations of Bassili [2] on motion based cues for facial expressions. Recall that the experiments of Bassili were intended to explore only the role of motion in recognizing facial expressions; therefore the face features, texture and complexion were unavailable to the experiment subjects. As illustrated in Figure 4, Bassili's conclusions are relatively simple to describe; he identified principle facial motions that provide powerful cues to the subjects to recognize facial expressions.

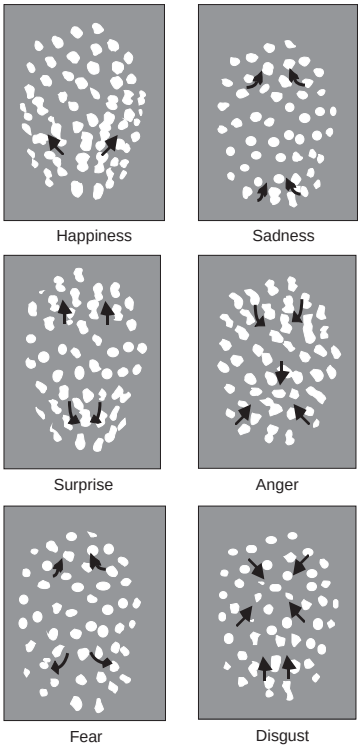

Figure 4: Motion cues for facial expression [2]

Bassili's results do not explicitly associate the motion patterns with specific face features or muscles since such information was unavailable to the experiment subjects.

3.2 A dictionary for facial dynamics

In this subsection we develop the dictionary of facial feature actions. The dictionary borrows from the facial cues of universal expression descriptions proposed in [6], and from the motion patterns of expression proposed in [2]. As a result, we arrive at a dictionary that is a *motion-based feature description of facial actions.*

The dictionary we propose 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 based on motion detection by optical flow within these regions.

Table 1 shows the components, basic actions, and cues that model the mouth (W denotes the window around the feature). Similar tables were created for the eyes and the eyebrows.

Comp.	Basic Action	Motion Cues
upper lip	raising	upward motion of W's upper part
	lowering	downward motion of W's upper part
	contraction	horizontal shrinking of W's upper part
	expansion	horizontal expansion of W's upper part
lower lip	raising	upward motion of W's lower part
	lowering	downward motion of W's lower part
	contraction	horizontal shrinking of W's lower part
	expansion	horizontal expansion of W's lower part
left corner	raising	upward motion of W's left part
	lowering	downward motion of W's left part
right corner	raising	upward motion of W's right part
	lowering	downward motion of W's right part
mouth	raising	upward motion throughout W
	lowering	downward motion throughout W
	compaction	overall shrinkage in mouth's size
	expansion	overall expansion in mouth's size

Table 1: The dictionary for mouth motions

The cues in Table 1 are not mutually exclusive. For example, the raising of a corner of the mouth can be a byproduct of raising of the upper or lower lip. Therefore, we introduce a ranking of actions according to interpretation precedence. Lip actions have higher interpretation precedence than mouth corners actions and whole mouth actions have the highest interpretation precedence.

3.3 Computing mid-level representations

The dictionary allows us to convert local directional motion patterns within a face region into a linguistic, per-frame, mid-level representation for facial actions. In addition to the basic actions the mid-level representation includes:

- Region actions: Basic actions within a rectangle surrounding a feature are combined to construct a region action. For example, the simultaneous raising of the upper lip and the lowering of the lower lip produce "mouth opening."

- Coordinated actions: Region actions that occur simultaneously at symmetric (i.e., the eyes, eyebrows, and cheeks) features can be combined to construct a coordinated action. For example, the raising of the right and left eyebrows produce a "raising brows" coordinated action.

A temporal consistency procedure is applied to the mid-level representation to filter out errors due to noise or illumination changes (see [14]).

3.4 Computing basic action cues

The approach we use for optical flow computation is one recently proposed by Abdel-Mottaleb et al. [1]. The flow magnitudes are first thresholded to reduce the effect of small motions probably due to noise. The motion vectors are then re-quantized into eight principle directions. The optical flow vectors are filtered using both spatial and temporal procedures that improve their coherence and continuity, respectively.

The largest set of motions is associated with the mouth since it has the most degrees of freedom at the anatomic and musculature levels (i.e., the independent mobility of the jaw in the upward/downward directions and left/right directions, the independence of the lips motion deformations and mutual interactions, and the separate control of the mouth corners).

We measure the motion of the mouth by considering a set of vertical and horizontal partitions of its surrounding rectangle (see Figure 5). 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 statistical measurements of these partitions (for details see [14]) 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.

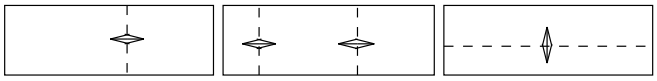

Figure 5: The vertical and horizontal partitions

4 Recognizing facial expressions

We describe in this section how the mid-level representation is used by a rule-based system to recognize facial expressions.

We divide every facial expression into three temporal parts: the *beginning*, *epic* and *ending*. Figure 6 shows the temporal parts of a smile model. Since we use the outward-upward motion of the mouth corners as the principle cue for a smile motion pattern, these are used as the criteria for temporal classification also. Notice that Figure 6 indicates that the detection of mouth corner motions might not occur at the same frames in both the beginning and ending of actions, and that we require at least one corner to start moving to label a frame with a "beginning of a smile" label, while the motions must completely stop before a frame is labeled as an epic or an ending. Notice that in general, motions ending a facial action are not necessarily the reverse of the motions that begin it.

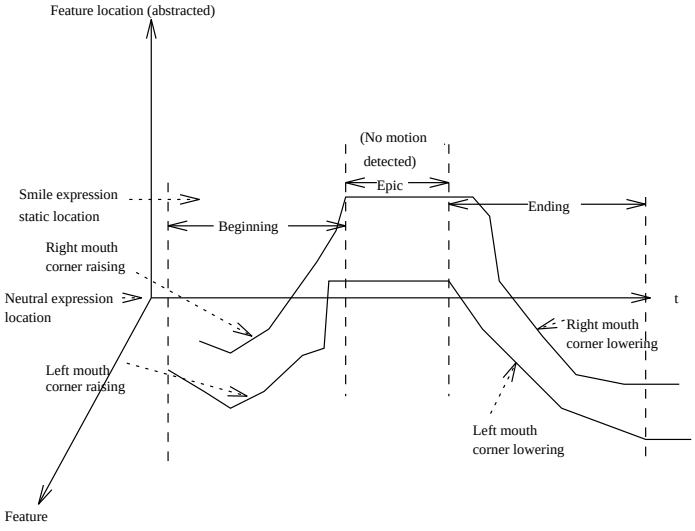

Figure 6: The temporal model of the "smile"

Similar temporal models are defined for each facial expression or action. These temporal models are only a first approximation to the huge repertoire of temporal facial actions that can occur. Furthermore, our overall modeling approach of characterizing expressions using a beginning-epic-ending trilogy has some fundamental limitations that need to be addressed in the future (e.g., when expressions overlap in time).

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 best demonstrated by an example- detection of a happiness expression. The system locates 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: 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", "epic", and "ending" of a smile.

Expr.	B/E	Satisfactory actions
Anger	B	inward lowering brows & mouth compaction
Anger	E	outward raising brows & mouth expansion
Disgust	B	upward nose motion & mouth expanded/opened lowering of brows
Disgust	E	downward nose motion & raising of brows
Happiness	B	raising mouth corners or mouth opening with its expansion
Happiness	E	lowering mouth corners or mouth closing with its contraction
Surprise	B	raising brows& lowering of lower lip (jaw)
Surprise	E	lowering brows& raising of lower lip (jaw)
Sadness	B	lowering mouth corners & raising mid mouth & raising inner parts of brows
Sadness	E	lowering mouth corners & lowering mid mouth & lowering inner parts of brows
Fear	B	slight expansion and lowering of mouth & raising inner parts of brows
Fear	E	slight contraction and raising of mouth & lowering inner parts of brows

Table 2: The rules for classifying facial expressions (B=beginning, E=ending)

5 Experiments

Our experimental subjects were asked to display emotion without additional directions. As a result, we acquired a variety of presumably similar facial expressions; some were consistent with [2] and [6] while others varied. The 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 express emotions they were not feeling at the time (fear and sadness were hard to induce).

We recorded short and long sequences (about 8 seconds and 16 seconds, respectively, taken at 30 frames per second, 120x160 pixels each) containing 2-3, and 3-5 expressions, respectively.

We requested each subject to display the emotions in front of the video camera while minimizing his/her head motion. Nevertheless, most subjects inevitably moved their head during a facial expression. As a result, the optical flow at facial regions was sometimes overwhelmed by the overall head rigid motion. The facial expression system we developed detects such rigid motion and marks the respective frames as unusable for analysis.

On a sample of 46 image sequences of 30 subjects (Figure 7) displaying a total of 105 emotions, the system achieved a recognition rate of 86% for smile, 94% for surprise, 92% for anger, 86% for fear, 80% for sadness, and 92% for disgust. Blinking detection success rate was 65%.

Figure 7: Tweleve subjects (out of more than 30)

Table 3 shows the details of our results. Occurrences of fear, and sadness are less frequent than happiness, surprise and anger. 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 detected.

Expression	Correct	False Alarm	Missed	Confused	Rate
Happiness	32	-	5	-	86%
Surprise	29	2	1	2	94%
Anger	22	1	2	2	92%
Disgust	12	2	1	2	92%
Fear	6	-	1	3	86%
Sadness	4	-	1	1	80%
Blink	68	11	38	-	65%

Table 3: Facial expression recognition results

Figure 8 shows four frames, the gap between each two frames being four frames. The upper left quarter shows the intensity image, the upper right quarter shows the gradient image, the rectangle in between displays the classification of facial expression, the lower left quarter shows the optical flow results, the rectangles around the face regions of interest and the mapping of colors into directions, and the lower right quarter shows the mid-level descriptions. Figure

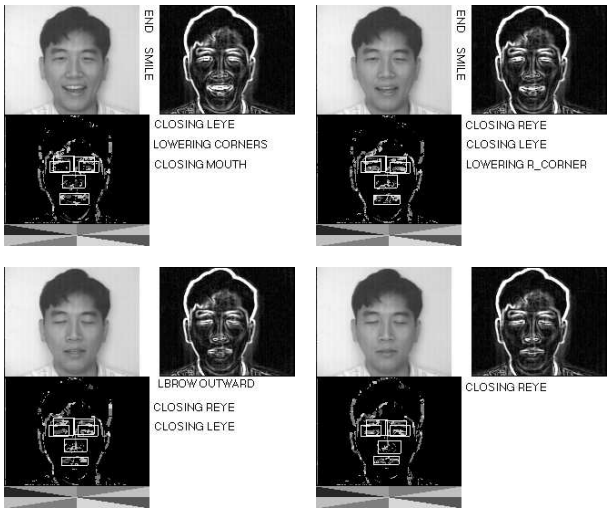

Figure 8: Four analyzed frames

8 shows the detection of an eye closing sequence. It also shows the closing of the mouth after a smile, the description of the mouth corner motions indicating the ending of a smile expression.

6 Summary

An approach for analyzing and classifying facial expressions from optical flow was proposed. This approach is based on qualitative tracking of principle regions of the face and flow computation at high intensity gradients points. A mid-level representation is computed from the spatial and the temporal motion results. The representation is linguistically motivated, following research in psychology in [2,6].

We have carried out experiments on thirty subjects in a laboratory environment and achieved good classification of facial expressions. Further study of the system's components is being performed as well as expanding its capability to deal with non-emotion facial messages. Our research suggests that models for analyzing and representing the dynamics of facial expression from images are needed.

Acknowledgements The authors would like to thank Gregory Baratoff, Pedja Bogdanovich, Tomas Brodsky, David Doermann, Claudio Esperanca, Jean-Yves Herve, Jennifer Kampfe, Cecilia Kullman, Vojislav Lalich-Petrich, Chih-Lung Lin, Hong-Che Liu, Carlos Morimoto, Kilaudia Rodriguez, Mark Rosenblum, Stephen Sickels, Saad Sirohey, Avital Sivan, Aya Soffer, Ansel Teng, Scott Thompson, Sebastian Toelg, Ting Wu, Yi-Sheng Yao and Min Zhou for subjecting themselves to the intrusive experiments on their facial expression. Thanks are due to Mohamed Abdel-Mottaleb for his assistance in implementing the optical flow algorithm.

References

[1] M. Abdel-Mottaleb, R. Chellappa, and A. Rosenfeld, "Binocular motion stereo using MAP estimation", *IEEE CVPR*, 1993, 321-327.

[2] J.N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of Personality and Social Psychology*, Vol. 37, 1979, 2049-2059.

[3] V. Bruce, *Recognizing Faces*, Lawrence Erlbaum Assoc., London, 1988.

[4] G. Chow and X. Li, "Towards a system for automatic facial feature detection," *Pattern Recognition*, Vol. 26, No. 12, 1993, 1739-1755.

[5] P. Ekman, (Edited) *Darwin and Facial Expression*, Academic Press, Inc. 1973.

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

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

[9] H. Li, P. Roivainen, and R. Forcheimer, "3-D motion estimation in model-based facial image coding," *IEEE PAMI*, Vol. 15, No. 6, 1993, 545-555.

[10] K. Mase, "Recognition of facial expression from optical flow," *IEICE Transactions,* Vol. E 74, No. 10, 1991, 3474-3483.

[11] K.R. Scherer and P. Ekman (Edited), *Approaches to Emotion*, Lawrence Erlbaum Associates, Inc., 1984.

[12] D. Terzopoulos, and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," IEEE PAMI, Vol. 15, No. 6, 1993, 569-579.

[13] Y. Yacoob, and L.S. Davis, "Labeling of human face components from range data," *IEEE CVPR,* 1993, 592-593.

[14] Y. Yacoob, and L.S. Davis, *Recognizing Facial Expressions*, in preparation.

[15] A.W. Young and H.D. Ellis (Edited), *Handbook of Research on Face Processing*, Elsevier Science Publishers B.V., 1989.

[16] A.L. Yuille, D.S. Cohen, and P.W. Hallinan, "Feature extraction from faces using deformable templates," *IEEE CVPR*, 1989, 104-109.