Recognizing Facial Expressions by Spatio-Temporal Analysis

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. Recognizing six facial expressions, as well as eye blinking, are demonstrated on a collection of image sequences.

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.

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 (see Figure 1).

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

2 Overview of our approach

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

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.

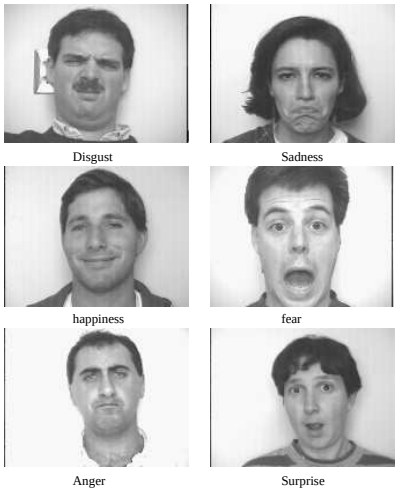

Figure 1: Six universal expressions

- The face is viewed from a near frontal view throughout the sequence. This allows us to avoid the increase in ambiguity of expression interpretation as the face moves from the frontal view.
- 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.
- We consider only the six universal emotions and blinking.

We focus on the motions associated with the edges of the mouth, nose, eyes, and eyebrows. These edges allow us to refer to the face features using natural linguistic terminology.

Figure 2 describes the flow of computation of our facial expression system. The first two components as well as the psychological basis for recognition of face expressions are given in [4,5]. The other components are explained in the rest of this paper.

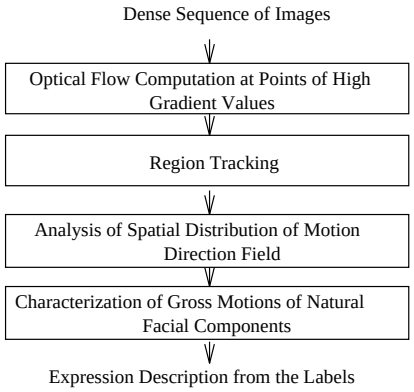

Figure 2: The flow of the facial analysis algorithm

3 Computing motion representations

3.1 A dictionary for facial dynamics

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. Similar tables were created for the eyes and the eyebrows.

Comp.	Basic Act.	Motion Cues
upper lip	raising	upward motion of w's upper part
	lowering	downward motion of w's upper part
	contract.	horizontal shrinking of w upper part
	expansion	horizontal expansion of w upper part
lower lip	raising	upward motion of w's lower part
	lowering	downward motion of w's lower part
	contract.	horizontal shrinking of w lower part
	expansion	horizontal expansion of w lower part
L. corner	raising	upward motion of w's left part
	lowering	downward motion of w's left part
R. 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
	compact.	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.

3.2 Computing basic actions

The approach we use for optical flow computation is correlation-based and was 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 requantized into eight principle directions.

The optical flow vectors are filtered using both spatial and temporal procedures that improve their coherence and continuity, respectively.

For example, we show the computation of the motion of the mouth by considering a set of vertical and horizontal partitions of its surrounding rectangle ([4]). 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.

Confidence measurements (see [4]) 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.

4 Recognizing facial expressions

We have designed a rule based system that combines some of the expression descriptions from [3] and [2]. We divide every facial expression into three temporal parts: the *beginning*, *epic* and *ending*. Figure 3 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 3 indicates that the detection of mouth corner motions might not occur at the same frame 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. In general, motions ending a facial action are not necessarily the reverse of the motions that begin it.

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,

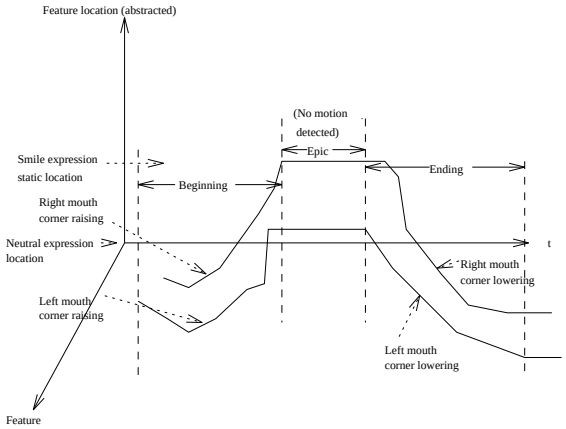

Figure 3: The temporal model of the 'smile'

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. Then, it identifies the first frame, f_3, in which an action 'lowering mouth corners' is detected and, 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 compact.
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. Our database of image sequences includes sequences of 32 different faces. We recorded short and long sequences (about 8 seconds and 16 seconds, respectively), containing 2-3, and 3-5 expressions, respectively. We requested each subject to display the emotions (as they usually would) in front of the video camera while minimizing 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 system detects such rigid motion and marks the respective frames as unusable for analysis.

On a sample of 46 image sequences of 32 subjects 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%. Table 3 shows the details of our results. Occurrences of fear, disgust and sadness are less frequent than 'happiness,' 'surprise' and 'anger' since the former were harder to stimulate in the subjects of our experiments. Some confusion of expressions occurred between the following pairs: 'fear' and 'surprise,' 'sadness' and 'disgust,' and 'sadness' and 'anger.' These distinctions rely on subtle coarse shape and motion information that were not always accurately detected.

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

Table 3: Facial expression recognition results

References

[1] M. Abdel-Mottaleb, R. Chellappa, and A. Rosenfeld, "Binocular motion stereo using MAP estimation", *IEEE CVPR*, 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," *Journal of Personality and Social Psychology*, Vol. 37, 2049-2059, 1979.

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

[4] Y. Yacoob, and L.S. Davis, *Computing Spatio-Temporal Representations of Human Faces,* Technical Report CAR-TR-706, Center for Automation Research, Univ. of Maryland, College Park, 1994.

[5] Y. Yacoob, and L.S. Davis, "Computing Spatio-Temporal Representations of Human Faces" *IEEE CVPR*, 70-75, 1994.