Facial and Body Feature Extraction for Emotionally-Rich HCI

Kostas Karpouzis
National Technical University of Athens, Greece

Athanasios Drosopoulos
National Technical University of Athens, Greece

Spiros Ioannou
National Technical University of Athens, Greece

Amaryllis Raouzaiou
National Technical University of Athens, Greece

Nicolas Tsapatsoulis
National Technical University of Athens, Greece

Stefanos Kollias
National Technical University of Athens, Greece

INTRODUCTION

Emotionally-aware Man-Machine Interaction (MMI) systems are presently at the forefront of interest of the computer vision and artificial intelligence communities, since they give the opportunity to less technology-aware people to use computers more efficiently, overcoming fears and preconceptions. Most emotion-related facial and body gestures are considered to be universal, in the sense that they are recognized along different cultures; therefore, the introduction of an "emotional dictionary" that includes descriptions and perceived meanings of facial expressions and body gestures, so as to help infer the likely emotional state of a specific user, can enhance the affective nature of MMI applications (Picard, 2000).

As a general rule, our intuition of what a human expression represents is based on trying to mimic the way the human mind works, while making an effort to recognize such an emotion. This means that even though image or video input is necessary for this task, this process cannot come to robust results without taking into account features like hand gestures or body pose. These features are able to convey messages in a much more expressive and definite manner than mere wording, which can be misleading or ambiguous. Sometimes, a simple hand action, such as placing one's hands over the ears, can pass on the message that you've had enough of what you are hearing more expressively than any spoken phrase.

BACKGROUND

Emotion Representation

Most emotion analysis applications attempt to annotate video information with category labels that relate to emotional states. However, since humans use an overwhelming number of labels to describe emotion, we need to incorporate a higher level and continuous representation that is closer to our conception of how emotions are expressed and perceived.

Activation-emotion space (Cowie, 2001) is a simple representation that is capable of capturing a wide range of significant issues in emotion. It rests on a simplified treatment of two key themes:

- Valence: The clearest common element of emotional states is that the person is influenced by feelings that are "valenced" (i.e., they are centrally concerned with positive or negative evaluations of people, things, or events).
- Activation Level: Research has recognized that emotional states involve dispositions to act in certain ways. Thus, states can be rated in terms of the associated activation level (i.e., the strength of the person's disposition to take some action rather than none).

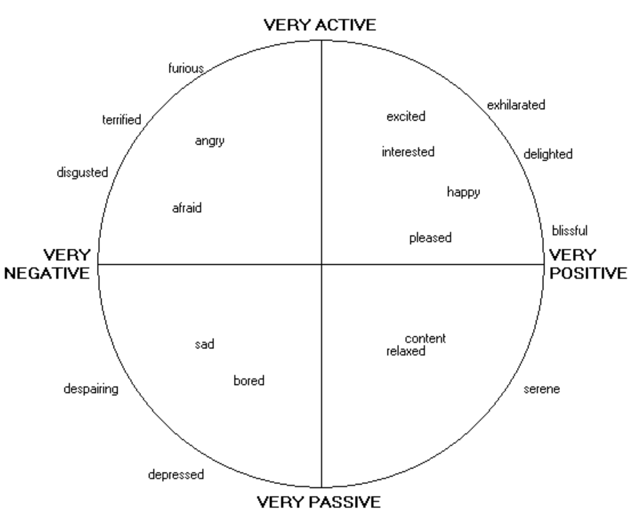
The axes of the activation-evaluation space reflect those themes, with the vertical axis showing activation level, while the horizontal axis represents evaluation. This scheme of describing emotional states is more tractable than using words and still can be translated into and out of verbal descriptions. Translation is possible because emotion-related words can be thought of as positions in activation-emotion space.

A surprising amount of emotional discourse can be captured in terms of activation-emotion space. Perceived full-blown emotions are not evenly distributed in activation-emotion space; instead, they tend to form a roughly circular pattern. In this framework, the center can be thought of as a natural origin, thus making emotional strength at a given point in activation-evaluation space proportional to the distance from the origin. The concept of a full-blown emotion can then be translated roughly as a state where emotional strength has passed a certain limit. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation. A related extension is to think of primary or basic emotions as cardinal points on the periphery of an emotion circle. Plutchik (1980) has offered a useful formulation of that idea—the emotion wheel—(see Figure 1).

Facial Expression Analysis

There is a long history of interest in the problem of recognizing emotion from facial expressions (Ekman, 1978), and there have been extensive studies on face perception during the last 20 years (Davis, 1975; Ekman, 1973; Scherer, 1984). The salient issues in emotion recognition from faces are parallel in some respects to the issues associated with voices, but divergent in others. In most cases, these studies attempt to define the facial expression of emotion

Figure 1: The activation-emotion space



in terms of qualitative patterns capable of being displayed in a still image. This usually captures the apex of the expression (i.e., the instant at which the indicators of emotion are most noticeable). More recently, emphasis has switched towards descriptions that emphasize gestures (i.e., significant movements of facial features).

In the context of faces, the task almost always has been to classify examples of the six emotions considered to be universal: joy, sadness, anger, fear, disgust, and surprise (Ekman, 1978). More recently, morphing techniques have been used to probe states that are intermediate between archetypal expressions. They reveal effects that are consistent with a degree of categorical structure in the domain of facial expression, but they are not particularly large, and there may be alternative ways of explaining them—notably by considering how category terms and facial parameters map onto activation-evaluation space (Karpouzis, 2000).

Analysis of the emotional expression of a human face requires a number of pre-processing steps that attempt to detect or track the face; to locate on it characteristic facial regions, such as eyes, mouth, and nose; and to extract and follow the movement of facial features, such as characteristic points in these regions or model facial gestures using anatomic information about the face. Facial features can be viewed (Ekman, 1975) either as static (i.e., skin color), slowly varying (i.e., permanent wrinkles), or rapidly varying (i.e., raising the eyebrows) with respect to time evolution. Detection of the position and shape of the mouth, eyes (particularly eyelids), wrinkles, and extraction of features related to them are the targets of techniques applied to still images of humans. However, in Bassili's (1979) experiments, expressions were recognized at above chance levels when based on image sequences, whereas only happiness and sadness were recognized at above chance levels when based on still images. Techniques that attempt to identify facial gestures for emotional expression characterization face the problems of locating or extracting the facial regions or features, computing the spatio-temporal motion of the face through optical flow estimation, and introducing geometric or physical muscle models describing the facial structure or gestures.

Body Gesture Analysis

The detection and interpretation of hand gestures have become an important part of human computer interaction in recent years (Wu, 2001). To benefit from the use of gestures in MMI, it is necessary to provide the means by which they can be interpreted by computers. The MMI interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and other parts of the human body, be measurable by the machine. First attempts to address this problem resulted in me-

Related Content