Enhancing Nonverbal Human Computer Interaction with Expression Recognition

Kostas Karpouzis, Nicolas Tsapatsoulis, Amaryllis Raouzaiou, George Moshovitis and Stefanos Kollias

Image Processing, Video and Multimedia Laboratory
Computer Science Division, Electrical and Computer Engineering Department
National Technical University of Athens
Heroon Polytechnic 9 GR 157 73 Zographou, Greece

email address : kkarpou@softlab.ece.ntua.gr

Abstract:- This paper describes an integrated system for human emotion recognition, which is used to provide feedback about the relevance or impact of the information that is presented to the user. Other techniques in this field extract explicit motion fields from the areas of interest and classify them with the help of templates or training sets; the proposed system, however, compares indication of muscle activation from the human face to data taken from similar actions of a 3-d head model. This comparison takes place at curve level, with each curve being drawn from detected feature points in an image sequence or from selected vertices of the polygonal model. The result of this process is identification of the muscles that contribute to the detected motion; this conclusion can then be used in conjunction with the Mimic Language, a table structure that maps groups of muscles to emotions. This method can be applied to either frontal or rotated views, as the curves that are calculated are easier to rotate in 3-d space than motion vector fields. The notion of describing motion with specific points is also supported in MPEG-4 and the relevant encoded data can be used in the same context, to eliminate the need to use machine vision techniques.

1. Introduction

The most widely employed method of human-computer interaction is the use of a keyboard and a mouse. Thus, users with low awareness or expertise with computer machinery are strongly impeded and their input is reduced to answering predetermined software prompts and following linear paths in menus and wizards. Moreover, the system receives little or no input about the relation of what the users get with respect to what they expected. This is critical in web-based applications where the user has to go through scores of data in order to reach that particular piece that they want. Usually, the users' first reaction to what they receive from an information mining system is a clear indication of whether it matched their expectation or not [7], [9]. Using such a piece of information, a system can present more relevant data to the user, in the case of a positive reaction, or start a new search, possibly formulated in a different manner.

The problem of facial expression recognition and coding is being tackled with various approaches that may be divided in two categories: static and motion dependent. In static approaches, recognition of a facial expression is performed using a single image of a face, while in the latter case, one extracts temporal information from two or more instances of a face in the same emotional state. Fully dynamic approaches use several frames (generally more than two and less than 15) of video sequences of a facial expression, typically lasting 0.5 to 4 seconds. This case seems to be the most promising and has up to now involved data analysis in the form of optical flow energy estimation and templates, as well as region tracking and deformation estimation [1], [3]. In our approach, automatic feature point extraction and motion estimation upon the extracted points is performed. The goal of this is to observe the temporal movement of key facial points that reveal the type of expression that takes place in a video sequence. The comparison is based on synthetic 3-d generated prototypes for the corresponding points. Matching real and synthetic results, and thus classification, is accomplished through the use of neural networks.

2. Facial muscles and emotions

Facial muscles' actuation, which results in the expression of emotions, has been the focus of attention of many scientists. Researchers during the 19th century divided facial muscles into groups, w.r.t. the emotions during which they are activated. Although this mapping does not conform with modern anatomical studies, it has been utilized to minimize the perceptive nature of a human emotion into discrete and countable features.

The FACS system [2] improved this mapping, by introducing the notion of an action unit (AU), i.e. the recognizable result of the flexing of a single or a small group of facial muscles. The 66 atomic action units that can be identified are combined into facial expressions. In some cases, more that ten action units can be recognized in a single movement or expression, while in others only a single unit is involved. This is a result of some motions being difficult to classify, based on mere visual data; in such cases, FACS defines an individual AU that involves all the muscles in the region or includes the same muscle in different units.

The FACS tables were derived by anatomical and physiological studies of the human face. This knowledge can help one understand and encode facial actions and reduce them to features and symbols. One can suggest that the movement of the facial bones and skin is a result of muscle contraction. The reverse may also be implied: the visual detection of motion in a human face can be connected with some muscle movement. Thus, if we recognize a change in a human face, we can safely deduce that at least a single muscle has flexed.

The muscles that are related to expressions are superficial and work collectively; as a result, it is difficult to separate the margins between the areas of influence of distinct muscles. However, each area of the face is mainly affected by a single group's contraction. In addition to that, the deformation of the surface of the face is not standard throughout the face, as a result of the different orientation of the facial muscles and the way they flex. Thus, one can assume that the observation of motion in the skin of the face and the tracking of its path can deduce the flexing of specific facial muscles. Let us presuppose that we have detected motion in the inner eyebrow area; the anatomy of the face informs us that this location moves under the influence of three distinct muscles, the frontalis, in the forehead, the depressor supercilii, under the medial end of the eyebrow and the depressor glabelle, which runs parallel between the eyebrows. All these muscles are linear, so we expect the motion to follow a parallel path to one of them or a linear combination, if more than one is activated. The position of these muscles can help us conclude that the median part of the eyebrow moves upward as a result of the flexing of the frontalis muscle, while any descending motion occurs when the depressor supercilii flexes. If the motion detected in this area is not parallel to the coronal plane of the head but is an aggregation of upward and lateral movement, then both the frontalis and the depressor glabelle are flexing.

2.1 Utilization of a 3-d model

In order to classify the motion of the different areas of the face, we use a 3-d model of a human head. The assumption that any conceivable motion in the area of the face occurs as a result of muscle flexing and the knowledge of facial anatomy help us extend the results from a generic model to the vast majority of the human faces. Most people smile in different ways and with distinct visual results; in any occasion, though, this expression is a result of the contraction of the same muscles to the same track.

Areas of vertices in the model are grouped w.r.t. the facial feature to which they correspond and the muscles responsible for their transformation. This mapping is a result of anatomy surveys and is not based on any mathematical models or measurements. The nature of each muscle helps us model the deformation of the overlying surface; for example, the flexing of linear muscles results in the forming of a furrow shape in the surrounding area.

The outcome of the modeling process is a “library” of possible muscle actions, grouped into emotions. Some of these are termed universal [8], as humans of different cultures can describe these fundamental emotions in a standard manner. For example, the eyebrows of a scared person are slightly raised and pulled to the medial plane, while the inner part is translated upward. Similar ideas and notions have been classified into the look-up tables of the Mimic Language [8].

3. Feature point extraction

3.1 Template matching

An interesting approach in the problem of automatic facial feature extraction is a technique based on the use of template prototypes, which are portrayed on the 2-d space in grayscale format. This is a technique that is, to some extent, easy to use, and also effective. It uses correlation as a basic tool for comparing the template with the part of the image that we wish to recognize. An interesting question that arises is the behavior of recognition with template matching in different resolutions. To examine this matter, gaussian pyramids were used, involving multi-resolution representations. The experiments proved that not very high resolutions are needed for template matching recognition. For example, the use of templates of 36¥36 pixels proved sufficient. This fact shows us that template matching is not as computationally complex as we originally imagined.

3.2 Automated Facial Feature Extraction

In our approach, as far as the frontal images are concerned, the fundamental concept upon which the automated localization of the predetermined points is based, consists of two steps: the hierarchic and reliable selection of specific blocks of the image and subsequently the use of a standardized procedure for the detection of the required benchmark points.

In order for the former of the two processes to be successful, the need of a secure method of approach emerged. The detection of a block describing a facial feature relies on a previously, effectively detected feature. By adopting this reasoning, the choice of the most significant characteristic -the ground of the cascade routine- had to be made. The importance of each of the commonly used facial features, regarding the issue of face recognition, has already been studied by other researchers. The outcome of surveys proved the eyes to be the most dependable and easily located of all facial features, and as such they were used. The techniques that were developed and tried separately, utilize a combination of template matching and Gabor filtering [6].

After having isolated the restricted regions of interest from the frontal image, the localization of the predetermined points ensues. The approximate center of the eye's pupil is searched as the darkest point both at the integrated horizontal and vertical direction of the eye's block. The exact position of the nostrils is sought from the sides to the center of the nose block. The mouth tips are met in a similar manner. Finally, the right and left head edges at the altitude of the eyes are retrieved from the horizontally integrated vector describing the area where the temple hair starts. It is obvious that the whole search procedure was attempted to be as close to human perception as possible.

3.3 The Hybrid Method

The basic quest of the desired feature blocks is performed by a simple template matching procedure. Each feature prototype is selected from one of the frontal images of the face base. The practiced comparison criterion is the maximum correlation coefficient between the prototype and the repeatedly audited blocks of a smartly restricted area of the face.

In order for the search area to be incisively and functionally limited, the knowledge of the human face physiology has been applied, without hindering the satisfactory performance of the algorithm in cases of small violations of the initial limitations.

However, the final block selection by the mere use of this method has not always been crowned with success. Therefore, the need of a measure of reliability came forth. For that reason, the use of Gabor filtering was deemed to be one suitable tool. As it can be mathematically deduced from the filter's form, it ensures simultaneous optimum localization in the natural space as well as in frequency space.

The filter is applied both on the localized area and the template in four different spatial frequencies. Its response is regarded as valid, only in the case that its amplitude exceeds a saliency threshold. The area with minimum phase distance from its template is considered to be the most reliably traced block.

3.4 Expression recognition within MPEG-4 streams

Utilization of the rationale and data used in an MPEG-4 encoded synthetic or natural objects can enhance our capability to recognize emotions and expressions. The geometry of a synthetic object is encoded as a tree of vertices and planar faces, while the deformations that take place, for example, during the apex of an expression are described with tokens termed FAPs (Facial Action

Parameters). These parameters can be combined or even substitute with the automatic feature extraction mechanism, as they incorporate the results of muscle flexing and the skin deformation that ensues [4], [5].

4. Feature point motion estimation

Block matching methods are being broadly used in various motion estimation problems for video sequences, mainly due to their ease in implementation and their relevant accuracy, as far as the calculated motion vectors are concerned. These are actually the reasons that such a kind of method has been used in our approach, in order to estimate how the feature points have progressively moved within a specific video sequence.

The executed block matching method aims at the computation of the specified points' transposition from one frame to its successive. Let the current frame be I1 and its successive be I2. For each pixel of the current frame I1(i,j) that is known to be a feature point ((i,j) Œ FP), we wish to find a displacement vector

$$d(i,j)=[d1(i,j),d2(i,j)],$$

such that I1(i,j)=I2(i+d1(i,j),j+d2(i,j)). For each pixel position (i,j) Œ FP of the current frame, we consider an n¥n block, the center of which is the specific pixel. A search procedure follows, which tracks the defined block of the current frame into its consecutive frame. This procedure will determine the motion vector of the feature point with respect to the block's displacement. Searching in frame I2 is performed within a limited N¥N search window, the center of which is this frame's (i,j) position.

Concerning the block matching criteria and the search methods, several variations have been proposed in the bibliography. The current implementation utilizes the Mean Absolute Difference (MAD) criterion and the so-called three-step exhaustive search method respectively [10].

5. Expression estimation

We reduce the problem of expression estimation to the encoding of motion curve sets that correspond to the feature points, followed by feature based classification using a multi-layer perceptron neural network. Fig. 1 presents the estimation system. As can be seen, preprocessing is necessary to encode the synthetic expressions database and train the network.

From each video sequence or synthetic expression, q motion curves observing the benchmark points in a three dimensional [x y t] space are extracted. Let S = {Ci: i Œ [1,q]} be this set of curves and _ be the definition vector of x, y coordinates that describes the curves over time. We project each curve to the space axis, obtaining two-dimensional curves, for the x- and y-axis, described by the definition vectors _x and _y respectively. The matching algorithm processes the projected curves independently and ultimately combines the results. In this way a trade off between space correlation and improved efficiency is obtained. The more complex three-dimensional problem can be tackled using techniques, but is out of the scope of this paper. In the following we will consider _ ∫ _x assuming that the results apply to the _y as well.

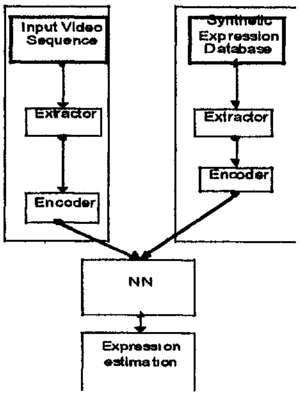

Fig. 1 The expression estimator

For the purpose of this paper we assume that the input video sequences start from and conclude to the neutral expression. The active expression period _ = [t_, t_] lies in between, where t_ is the last frame where all benchmark points are stabilized and following the muscle activation, t_ is the frame when the benchmark points are re-stabilized. Information outside this period is rejected by appropriately cropping _. Using _ coordinates over time as knots, we obtain a continuous approximation of the curve, which is subsequently re-sampled to m samples yielding a normalized description vector _.

While such a vector describes C^i, it is not really appropriate for the matching process. For robust classification we demand a fine invariance properties from the representation space. Our approach is to transform _ to a feature vector _ using a carefully designed encoding scheme, employing central moment descriptors. The composite feature vector _ invariantly describes the two-dimensional curve set by concatenating the _ vectors describing each curve in S.

Using the _ vector we employ a multi-layer neural network to match the real world input curve set against our synthetic expression database. The network consists of Ni = dim (_) = 60 input neurons that correspond to the moments' parameters, n output neurons that correspond to the expression classes, and 2 layers of hidden neurons.

The supervised learning process we use for adapting the neural network consists of the following steps:

- Motion curve sets are extracted for each synthetic expression sequence.
- The encoder splits the curves in two-dimensional components to be transformed from definition space _ to the feature space _ using the encoding scheme previously described.
- The calculated feature vectors _i and their corresponding output vectors provided by the supervisor are fed to the multi-layer perceptrons that compose our classification neural network, thus adapting the neuron weights to our problem domain.

During the allocation stage, the NN is fed with the feature vector of the input video sequence. The output vector is transformed by a sigmoid transfer function to normalize and threshold the results producing the expression mix vector _, the coordinates of which represent the matching of the input against the respective expression class.

6. Results

The following results indicate the discrimination between the time-related paths that are drawn from the natural images. This difference is in all tested cases enough to distinguish muscle activation, despite the presence of noise and error.

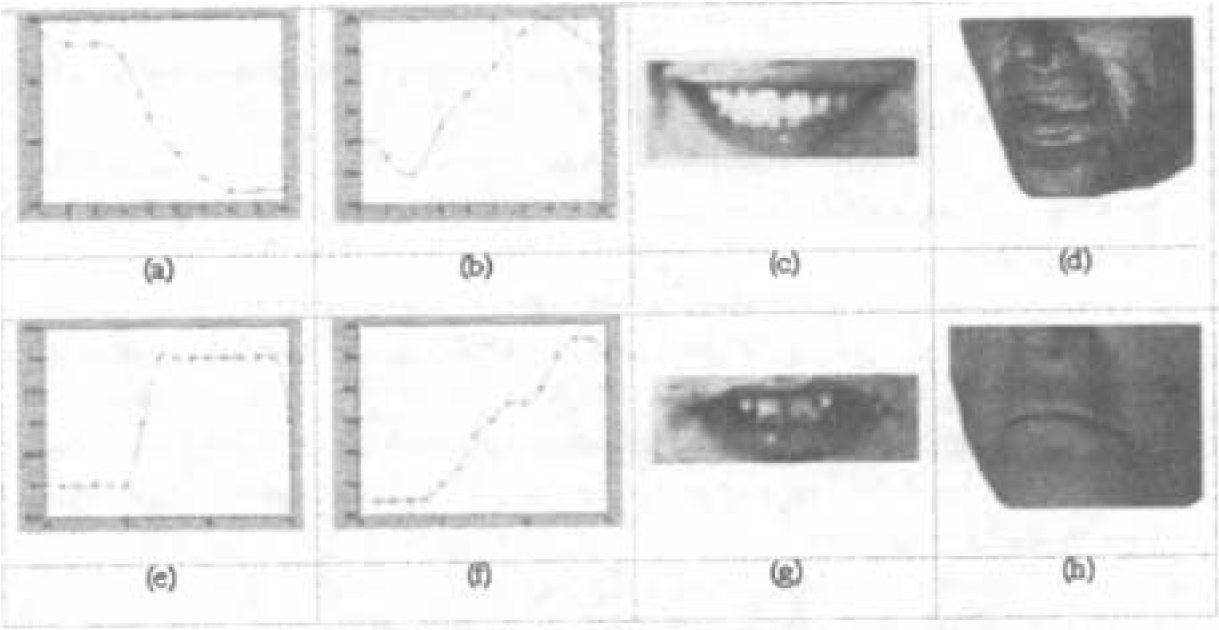

Fig. 2 (a) and (b) show the horizontal movement of the right and left mouth corner for 'smile', (c) shows the detected feature points from the video sequence and (d) shows the synthetic expression, (e) and (f) show the horizontal movement of the right and left mouth corner for 'anger', (g) shows the detected feature points from the video sequence and (h) shows the synthetic expression

7. Conclusion

The proposed system utilizes automatic feature extraction and motion estimation techniques, along with 3-d face models to compare motion data using predefined templates. The muscle activation information that is calculated can be mapped to groups of emotions, either through the Mimic Language or using neural networks. The above notion includes MPEG-4 encoded streams or observations of rotated heads, instead of standard frontal views. The outcome of this procedure is used to provide the system with additional information with respect to the users' reaction to what is presented to them. This knowledge can then be used in a feedback manner to retrieve relevant information for the user or offer them the possibility to change their preferences.

References:

[1] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W. and Taylor J., Emotion Recognition in Human-Computer Interaction, submitted for publication

[2] Ekman P., Friesen W., Manual for the FACS, Consulting Psychologists Press, 1978