Emotion Recognition Using Feature Extraction and 3-D Models

KOSTAS KARPOUZIS, GEORGE VOTSIS, GEORGE MOSCHOVITIS AND STEFANOS KOLLIAS
Image Processing, Video and Multimedia Systems Group
Computer Science Division, Electrical and Computer Engineering Department
National Technical University of Athens
Heroon Polytechniou 9 GR - 157 73 Zographou
GREECE

Abstract:- This paper describes an integrated system for human emotion recognition. While other techniques extract explicit motion fields from the areas of interest and combine them with templates or training sets, the proposed system compares evidence of muscle activation from the human face to relevant data taken from a 3-d model of a head. 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 is then used in conjunction with neural networks that map groups of muscles to emotions. The notion of describing motion with specific points is also supported in MPEG-4 and the relevant encoded data may easily be used in the same context.

Key-Words :- Expression recognition, 3-d muscle mesh, feature extraction, motion estimation CSCC'99 Proc.pp.5371-5376

1. Introduction

The facial expression recognition problem has lately undergone various approaches that may be divided in two main categories: static and motion dependent. In static approaches, recognition of a facial expression is performed using a single image of a face. Motion dependent approaches extract temporal information by using at least two 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 containing a facial expression, which normally lasts 0.5 to 4 seconds. The latter case, which seems to be the most promising, has up to now involved data analysis in the form of optical flow energy estimation and energy templates, or region tracking and region deformation estimation, or even 3-d model alignment [1]. In our approach, automatic feature point extraction and motion estimation upon the extracted points is performed. This aims at observing the temporal movement of facial key-points, which reveals the type of emotion taking 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

The interaction of the facial muscles and the expression of emotional states has been in the focus of attention of many scientists. Researchers during the 19th century divided the facial muscles into groups, with respect to the emotions during which they are activated. Although this mapping does not conform with recent anatomical studies, it was used as a basis to minimize the continuous and perceptive nature of a human emotion into discrete and, in a way, countable features.

This mapping was well improved by Ekman's pioneering work in FACS [2]. Ekman conceived the notion of an action unit, which is in essence the recognizable result of the flexing of a single or a small group of facial muscles. The 66 action units (AUs) that can be identified can be combined to generate or infer 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 AU 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 scope of FACS imposes some limitations in the sense that action units can only describe visually distinguishable facial movement and not any *possible* changes. In fact, this is not exactly the case with the

lower part of the face, as the independent movement of the jaw and the flexibility of the lips allow a great number of visually perceivable, but virtually identical actions. Also, it does not tackle the problem of emotion or expression synthesis and recognition.

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, that is, the visual or intuitive fact that there is motion in a human face can be connected with muscle movement. Thus, if we can detect and recognize a change in a human face, we can safely deduce that at least a single muscle has flexed.

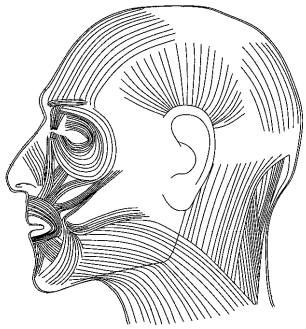


Fig 1. Structure and position of facial muscles

Not all facial muscles contribute to expressions ; some perform actions related with moving the jaw or forming the phonemes. The muscles that are related to expressions are superficial, i.e. near the surface of the head, and are responsible of the general shape of the head. Most of these muscles work collectively; as a result, it is difficult to separate the margins between the areas of influence of distinct muscles (see Fig. 1). 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. For example, the muscle that is accountable for the motion in the cheeks, *zygomatic major*, contracts in a linear fashion, while *orbicularis oculi*, the muscle around the eyes compresses its area of influence circularly. Thus, one can assume made 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 assume that we have detected motion in the inner eyebrow area; the anatomy of the face inform us that this location moves under the influence of three distinct muscles, the *frontalis*, which is the muscle that covers the forehead, the *depressor supercilii*, the muscle under the medial end of the eyebrow and the *depressor glabelle*, which runs parallel and 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 combination of this fact and the position of these muscles can help us conclude that if the median part of the eyebrow moves upward, this is a result of the flexing of the *frontalis* muscle, while the 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.

3. Utilization of the 3-d model

In order to classify the motion of the different areas of the face, we use a 3-d model of a complete human head. Although the employment of a simple mask might simplify numerical operations, it would not be suitable for re-synthesizing the expression. Despite the diversity of the shape of the head between humans, the assumption that any conceivable motion in this area occurs as a result of muscle flexing and the knowledge of facial anatomy help us extend these results to the vast majority of the human faces. Most people smile in different ways and with unlike visual results; in any occasion, though, this expression is a result of the contraction of the same muscles to the same track.

The model that is employed is a medium-resolution polygonal model of the surface of the head (about 8K vertices). Higher polygon counts would not assist our goal and would make operations much more complex. Areas of vertices 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; that is, the flexing of linear muscles results in the forming of an elevated or furrow shape in the surrounding area, while circular muscles produce radial motion in their areas of influence.

The outcome of the modeling process is a library of possible muscle actions. Some of these emotions are termed universal [3] as they can be recognized across different cultures. Humans can recognize and describe these fundamental emotions in a standard manner. For example, the eyebrows of an afraid 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. The system utilizes the fact that facial expressions are the result of dynamic and static factors, both being influenced by the mental and emotional condition of the subject. While the static aspects are determined by the shape and structure of the specific face and therefore cannot be generalized, the dynamic expressions are produced by the universal muscle influence.

4. Feature point extraction

4.1 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, but 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. This involves multi-resolution representations through the use of gaussian pyramids. 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.

4.2 Gabor Filtering

It is possible for Gabor filtering to be used in a facial recognition system. The neighboring region of a pixel may be described by the response of a group of Gabor filters in different frequencies and directions, which have a reference to the specific pixel. In that way, a feature vector may be formed, containing the responses of those filters.

One form of the Gabor filter is the following:

$$g(x,y,u,v) = \exp\left(-\left(\frac{x^2}{s_x^2} + \frac{y^2}{s_y^2}\right) + 2pj(ux + vy)\right)$$

where x, y is a point location, u, v govern the filter's central spatial frequency, which determines its orientation, and the parameters ó control the width of the Gaussian window relative to the wavelength corresponding to the central frequency [4].

As mentioned above, the feature vector consists of the responses of the filters in different central frequencies and for a specific position. During the comparison of two images, the phase difference of the response of each filter corresponds to local variation towards the filter's direction. These local position variations are combined to express the 'distance' of the two images.

4.3 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 that 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 [4].

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.

4.4 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.

4.5 Correspondence with the MPEG-4 synthetic model

The MPEG-4 standard uses the VRML as a starting point for its synthetic capabilities. However, this description proved insufficient for the needs of the standard. This is the reason that various capability extensions have been incorporated. Among others is the synthetic face and body (FAB) animation capability, which is a model-independent definition of artificial face and body animation parameters. Through the adopted coding, one has the potential of compactly representing facial expressions via the movement of a set of feature points. This set consists of a numerous gathering of points that are defined on a 2-d or a 3-d head mesh, as it may be seen in [5]. The feature point set supported in this work is a subset of the MPEG-4 coding system, merely because that is sufficient for discrimination purposes, which is the aim of the current work. However, progress is being done by our group upon the enhancement of the original, automatically extracted feature point set. Moreover, feature point motion estimation in combination with 3-d recovery techniques for human faces in video sequences, supports the MPEG-4 context, a fact that directly means the embodiment of facial expression coding in terms of the standard.

5. 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) ∈ 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) ∈ 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, plenty 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.

The MAD criterion produces the proposed displacement through the minimization of the sum:

$$MAD(d1,d2) = \frac{1}{n2} \sum_{(k,l)\in B} | I1(k,l) - I2(k+d1,l+d2)$$

|It may be seen at the above equation that this criterion is easily and quickly realizable, even in the case wheresome kind of an exhaustive search is being used.

The three-step exhaustive search procedure is described in the following:

1. For each pixel position (i,j) ∈ FP of the current frame, we select the corresponding pixel position (i,j) on the next frame, as well as its eight neighboring points that have a horizontal or/and vertical distance of four pixels from the specific

position. As a result, we get the initial set of nine points:

S1 = {(i,j), (i+4,j), (i-4,j), (i,j+4),(i,j-4), (i+4,j+4), (i+4,j-4), (i-4,j+4), (i-4,j-4)}

upon which the first step of the search procedure will be performed.

2. Having determined the initial point set S1, we perform a comparison between the n×n block defined in the current frame by the pixel position (i,j) and each of the n×n blocks defined in the following frame by the point set S1. The comparison is accomplished through the use of the MAD criterion. The element of the set S1 that minimally satisfies this criterion, is selected to be the center pixel position (ic1,jc1) of the next phase.
3. Except for the point (ic1,jc1), we consider its eight neighboring points that have a horizontal or/and vertical distance of two pixels from the specific position. As a result, we get the intermediate set of nine points S2, upon which the next step of the search procedure will be performed.
4. Having determined the intermediate point set S2, we repeat step 2, where S2 and (ic2,jc2) are involved instead of S1 and (ic1,jc1) respectively.
5. Except for the point (ic2,jc2), we consider its eight neighboring points that have a horizontal or/and vertical distance of one pixel from the specific position. As a result, we get the intermediate set of nine points S3, upon which the final step of the search procedure will be performed.
6. Having determined the intermediate point set S3, we repeat step 2, where S3 and (if,jf) are involved instead of S1 and (ic1,jc1) respectively. The pixel position (if,jf) on the frame I2, is the one to which the original position (i,j) on the current frame I1 has moved. The displacement vector for each feature point between frames I1 and I2 is then calculated as:

$$d(i,j) = [(if,jf) - (i,j)]$$

6. Expression estimation

We reduce the problem of expression estimation to the encoding of motion curve sets that observe the feature point paths, followed by a feature based classification using a multi-layer perceptron neural network (NN). Fig. 2 presents the estimation system. As it can be seen, a preprocessing step is necessary to encode the synthetic expressions database and train the network. To simplify the implementation, several assumptions are made which will be discussed in the following.

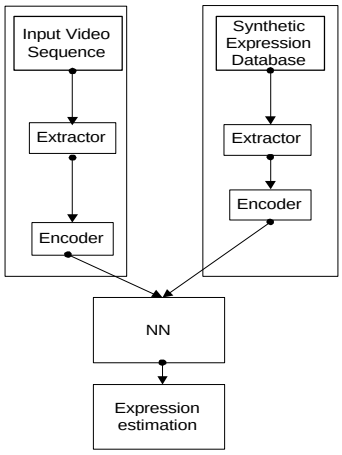


Fig. 2 The expression estimator

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 = $\{C^i: i \in [1,q]\}$ be this set of curves and ä 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 $ä_x$ and $ä_y$ 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 complex three-dimensional problem can be tackled using techniques discussed in [6], but is out of the scope of this paper. In the following we will consider ä ≡ $ä_x$ assuming that the results apply to $ä_y$ as well.

For the purpose of this paper we assume that the input video sequences start from and conclude to the neutral expression. In between lies the active expression period Ô = [$t_{á}$, $t_{â}$], where $t_{á}$ is the last frame where all feature points are stabilized and following the muscle activation, $t_{â}$ is the frame when the feature points are re-stabilized. Information outside this period is rejected by appropriately cropping ä. Using ä coordinates over time as knots, we obtain a Non-Uniform Rational B-Spline (NURBS) approximation of the curve, which is subsequently re-sampled to m samples yielding a normalized description vector ã.

While such a vector describes C^i, it is not really appropriate for the matching process. For robust classification we demand affine invariance properties from the representation space. Our approach is to transform ã to a unique feature vector ö using a

carefully designed encoding scheme, employing central moment descriptors. The composite feature vector ù invariantly describes the two-dimensional curve set by concatenating the ö vectors describing each curve in S.

Using the ù 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 N_i = dim (ù) = 60 input neurons that correspond to the moments parameters, n output neurons that correspond to the expression classes, and 2 layers of hidden neurons.

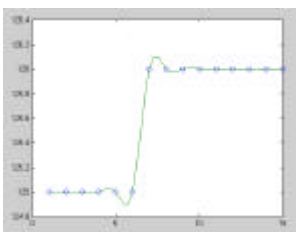
The supervised learning process we use for adapting the neural network consists of the following steps:

- 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 Ä to the feature space Ù using the encoding scheme previously described.
- The calculated feature vectors $ù^i$ 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.

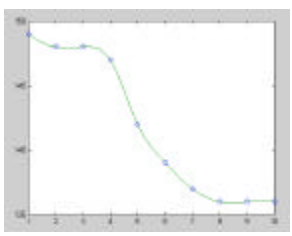
During the allocation stage, the NN is fed with the feature vector of the input video sequence. The output vector is transformed by a sigmoid transfer function to normalize and threshold the results producing the expression mix vector ì, the coordinates of which represent the matching of the input against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

7. Results

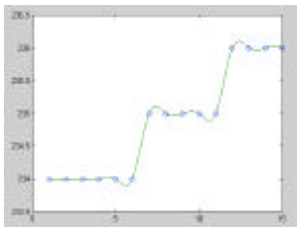
The following results indicate that there is some discrimination between the time-related paths that are drawn from the natural images. This difference is generally enough to distinguish muscle activation, despite the presence of noise and error.



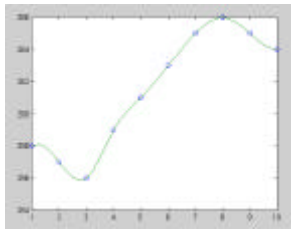
(a)



(b)



(c)



(d)

Fig. 3 (a) and (b) show the horizontal movement of the right mouth corner for 'anger' and 'smile' respectively. (c) and (d) show the horizontal movement of the left mouth corner for 'anger' and 'smile' respectively.

8. Conclusion

The proposed system utilizes automatic feature extraction and motion estimation techniques, along with 3-d face models to compare motion data to pre-defined prototypes. This results to muscle activation information which is mapped to groups of emotions, through the Mimic Language. The above notion can be extended to include MPEG-4 encoded streams or observation of rotated heads, instead of the standard frontal view.

References:

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

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

[3] F. Parke, *Computer Facial Animation*, A K Peters, 1996

[4] G. Votsis et al., *A Simplified Representation of 3D Human Faces Adapted from 2D Images*, NMBIA, 1998, Scotland

[5] B. Haskell et al., Image and Video Coding, *IEEE Trans. on CAS for VT*, Vol. 8, No. 7, 1998, pp. 814-837

[6] Y. Xirouhakis, Y. Avrithis and S. Kollias, *Image Retrieval and Classification using Affine Invariant B-Spline Representation and Neural Networks*, Proc. IEE, 1998.