ROBUST FACIAL EXPRESSION RECOGNITION USING SPATIALLY LOCALIZED GEOMETRIC MODEL

Ashutosh Saxena
Department of Electrical Engineering
IIT Kanpur
Kanpur 208016, India
ashutosh.saxena@ieee.org

Ankit Anand
Dept. of Computer Sc. and Engg.
IIT Kanpur
Kanpur 208016, India
ankanand@cse.iitk.ac.in

Prof. Amitabha Mukerjee
Dept. of Computer Sc. and Engg.
IIT Kanpur
Kanpur 208016, India
amit@cse.iitk.ac.in

ABSTRACT

An efficient, local image-based approach for extraction of intransient facial features and recognition of four facial expressions from 2D image sequences is presented. The algorithm uses edge projection analysis for feature extraction and creates a dynamic spatio-temporal representation of the face, followed by classification through a feed-forward net with one hidden layer. A novel transform for extracting lip region for color face images based on Gaussian modeling of skin and lip color is proposed. The proposed lip transform for colored images results in better extraction of lip region in the feature extraction stage. The algorithm achieves an accuracy of 90.0% for facial expression recognition from grayscale image sequences.

Categories and Subject Descriptors

[**Cybernetics**]: Pattern Recognition and Analysis, Artificial Intelligence & Applications – *facial expression recognition, color images.*

General Terms

Algorithms, Design.

Keywords

Color images, face, facial expression recognition, lip region extraction.

1. INTRODUCTION

Identifying human facial expressions has become an important field of study in recent years because of its inherent intuitive appeal and also due to possible applications such as human-computer interaction, face image compression, synthetic face animation and video facial image queries.

While approaches based on 3D deformable facial model have achieved expression recognition rates of as high as 98% [2], they are computationally inefficient and require considerable apriori training based on **3D information**, which is often unavailable. Recognition from 2D images remains a difficult yet important problem for areas such as image database querying and classification. The accuracy rates achieved for 2D images are around 90% [3,4,5,11]. In a recent review of expression recognition, Fasel [1] considers the problem along several dimensions: whether features such as lips or eyebrows are first identified in the face (local [4] vs holistic [11]), or whether the image model used is 2D or 3D. Methods proposed for expression recognition from 2D images include the Gabor-Wavelet [5] or Holistic Optical flow [11] approach.

This paper describes a more robust system for facial expression recognition from image sequences using 2D appearance-based local approach for the extraction of intransient facial features, i.e. features such as eyebrows, lips, or mouth, which are always present in the image, but may be deformed [1] (in contrast, transient features are wrinkles or bulges that disappear at other times). The main advantages of such an approach is low computational requirements, ability to work with both colored and grayscale images and robustness in handling partial occlusions [3].

Edge projection analysis which is used here for feature extraction (eyebrows and lips) is well known [6]. Unlike [6] which describes a template based matching as an essential starting point, we use contours analysis. Our system computes a feature vector based on geometrical model of the face and then classifies it into four expression classes using a feed-forward basis function net. The system detects open and closed state of the mouth as well. The algorithm presented here works on both color and grayscale image sequences. An important aspect of our work is the use of color information for robust and more accurate segmentation of lip region in case of color images. The novel lip-enhancement transform is based on Gaussian modeling of skin and lip color.

To place the work in a larger context of face analysis and recognition, the overall task requires that the part of the image involving the face be detected and segmented. We assume that a near-frontal view of the face is available. Tests on a grayscale and two color face image databases ([8] and [9,10]) demonstrate a superior recognition rate for four facial expressions (smile, surprise, disgust and sad against neutral).

2. IDENTIFICATION OF FEATURE REGIONS

We have used integral projections [6] of the edge map of the face image for extraction of facial features. Let I (x, y) be the input image. Vertical and horizontal projection vectors (fig 1) in the rectangle $[x_1, x_2]$ x $[y_1, y_2]$ are defined as

$$V(x) = \sum_{y=y_1}^{y=y_2} I(x,y) \qquad (1a)$$

$$H(y) = \sum_{x=x_1}^{x=x_2} I(x,y) \qquad (1b)$$

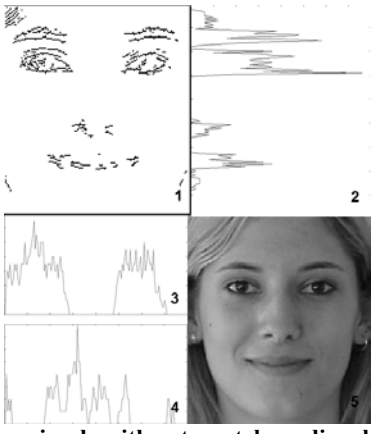


Figure 1. Generic algorithm to get bounding box. (1) Edge map, (2) Vector H (y), (3) Vector V (x) in upper half portion, (4) Vector V (x) in lower half portion, (5) Original image.

A typical human face follows a set of anthropometric standards, which have been utilized to narrow the search of a particular facial feature to smaller regions of the face. We use the following generic algorithm for the facial feature extraction from the localized face image—

1. An approximate bounding box for the feature is obtained using the anthropometric standards.
2. Sobel edge map (fig 1.1) is computed to obtain edges along the boundary of the feature.
3. The integral projections V(x) and H(y) are calculated on the edge map (fig 1.2, 1.3, and 1.4).
4. Median filtering followed by Gaussian smoothing smooths the projection vectors so obtained. Higher value of projection vector at a particular point indicates higher probability of occurrence of the feature. The relative probability E(i) of the i^{th} region containing the feature is calculated as—

$$E(i) = \sum_{y=y_i}^{y=y_{i+1}} H(y)\, w(y) \qquad (2)$$

Where, 'y_i' is the position of the i^{th} region; and w(y) is the gaussian weighing factor, calculated on the basis of anthropometric standards.

The region with maximum E(i) gives the vertical extent of the region containing the feature. Similar approach is used for getting the vertical extent from the vertical projection V(x).

5. The bounding box so obtained is processed further to get an exact binary mask of the feature.

The following sections describe in detail the application of this generic algorithm for each particular case of eyebrow, lip and nose.

2.1. Eyebrow

The approximate bounding box is the top half of the face. The generic algorithm uses horizontal sobel edges to compute bounding box containing *eye* and *eyebrow*. The segmentation algorithm cannot give bounding box for the eyebrow exclusively because the edges due to eye also appear in the chosen bounding box. Brunelli [6] suggests use of template matching for extracting the eye, but we use another approach as described below.

Eyebrow is segmented from eye using the fact that the eye occurs below eyebrow and its edges form closed contours (fig 2), obtained by applying *Laplacian of Gaussian* operator at zero threshold. These contours are filled and the resulting image containing masks of eyebrow and eye is morphologically filtered by horizontally stretched elliptic structuring elements. From the two largest filled regions, the region with higher centroid is chosen to be the mask of eyebrow.

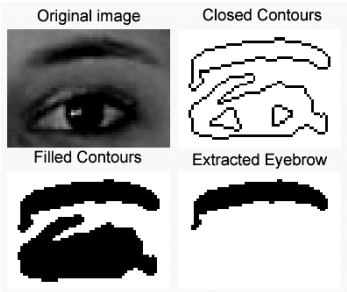


Figure 2. Post processing of approximate bounding box to get exact binary mask of eyebrow.

2.2. Lip

The approximate bounding box is the lower half of the face. In case of colored images, lip pixels significantly differ from those of skin in *YCbCr* color space. Therefore the color image is preprocessed to produce pronounced demarcation between lip and other skin regions (see section 3 for the details of the proposed lip enhancement transform). For the case of grayscale images, no such preprocessing is applied.
The generic algorithm calculates edge maps on the transformed image. Edges for lips occur both in horizontal and vertical direction. In the bounding box computed by the generic algorithm, closed contours are obtained by applying *Laplacian of Gaussian* operator at zero thresholds. These contours are filled and morphologically filtered using elliptic structuring elements to get binary mask for the lips.

2.3. Nose

The approximate bounding box for the nose lies between the eyes and the mouth. The generic algorithm uses vertical sobel edges to compute the vertical position, which is required as a reference point on face.

3. LIP ENHANCEMENT TRANSFORM

Sadeghi [12] proposed a lip segmentation method based on Gaussian mixture modeling of mouth area images, followed by Bayesian decision-making system, which labels the pixels as lip or non-lip. This approach gives a binary image, on which usual feature extraction algorithms cannot be applied. Approach by Lievin [13,14] using Bayesian segmentation also results in a binarized image.

To follow the usual feature extraction algorithms (for example as in [15]), the method would be to first convert the vector (color) image to a scalar (grayscale) image by

$$S_1 = a_1 R + a_2 G + a_3 B \qquad (3)$$

Where (*R,G,B*) are the components in the *RGB* color space. This is followed by other operations like calculating edge maps to find the lip region in a color image. This approach would use intensity information only. This compromises the accuracy of lip segmentation, because intensity difference need not be significant between the lip and skin pixels, and lightening variations may also be present.

We propose a transform, based on Gaussian modeling of mouth area images, to convert the vector (color) image to a novel scalar image, on which further feature extraction algorithms can be applied. This method uses the chromaticity components in *YCbCr* color space.

3.1. Gaussian modeling of skin and lip color

Color of skin and lip is a very important property that can be used for identification of the skin and lip regions. To model skin color, one has to look for color spaces in which the distribution of color components is concentrated in a small area. Researchers have looked at various color spaces—normalized RGB, YCbCr, HIS, etc. Chai [7] has studied various color spaces for modeling skin pixels. Skin and lip pixels are localized in a small portion in a two-dimensional *Cb-Cr* space. Note that the color space chosen should be such that it is independent of intensity component.

The sample distributions were calculated experimentally from over randomly selected 93 images (1000 pixels for lip), and (40,000 pixels for skin). The skin and lip pixels in *Cb-Cr* space are confined in a small region (figure 3).

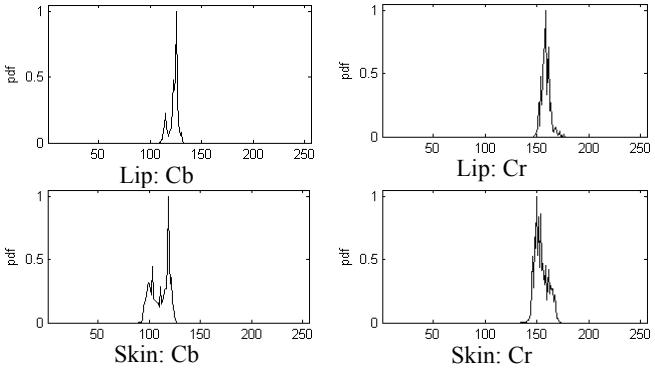


Figure 3. Sample distribution of lip and skin pixels along Cb and Cr axis.

On one extreme, distribution of skin and lip each can be modeled by a simple vector $\boldsymbol{\mu} = (\mu_{Cb}, \mu_{Cr})$, which is the mean in *Cb-Cr* space. On the other extreme, one can model the distribution by summation of *G* gaussians. In former case, the distribution could not be properly modeled. In latter case, the number of parameters (= 6 *G* for each) becomes very large, rendering the transform computationally intensive. We used a compromising solution, using one two-dimensional Gaussian to model each of skin and lip pixels as in [7]. Therefore, we get statistically estimated mean of skin pixels and lip pixels as

$$\hat{\boldsymbol{\mu}}_{skin} = (\mu_{Cb}, \mu_{Cr}) \qquad (4a)$$

$$\hat{\boldsymbol{\mu}}_{lip} = (\mu_{Cb}, \mu_{Cr}) \qquad (4b)$$

We get one covariance matrix each for skin and lip distributions. The *Cb* and *Cr* components have very high cross-covariance C_{CbCr}. The covariance matrices are given by

$$\mathbf{C}_{Skin} = \begin{bmatrix} C_{Cb^2} & C_{CbCr} \\ C_{CbCr} & C_{Cr^2} \end{bmatrix} \qquad (5a)$$

$$\mathbf{C}_{Lip} = \begin{bmatrix} C_{Cb^2} & C_{CbCr} \\ C_{CbCr} & C_{Cr^2} \end{bmatrix} \qquad (5b)$$

3.2. The proposed Transform

Let p_{skin} and p_{lip} be apriori probabilities of skin and color, i.e., what is the expected fraction of skin and lip pixels in the chosen window. The probability of an image pixel to be lip independently is P_{lip} and the probability of an image pixel to be skin independently is P_{skin}.

$$P_{Lip} = \frac{1}{\sqrt{2\pi|C_{Lip}|}} \exp\left(-\frac{1}{2} (\mathbf{I} - \hat{\boldsymbol{\mu}}_{\mathbf{Lip}}) \mathbf{C}_{\mathbf{Lip}} \left(\mathbf{I} - \hat{\boldsymbol{\mu}}_{\mathbf{Lip}} \right)^T \right) \qquad (6)$$

$$P_{Skin} = \frac{1}{\sqrt{2\pi|C_{Skin}|}} \exp\left(-\frac{1}{2} (\mathbf{I} - \hat{\boldsymbol{\mu}}_{\mathbf{Skin}}) \mathbf{C}_{\mathbf{Skin}} \left(\mathbf{I} - \hat{\boldsymbol{\mu}}_{\mathbf{Skin}} \right)^T \right) \qquad (7)$$

Where, $\mathbf{C_{Lip}}$ and $\mathbf{C_{Skin}}$ are estimated the covariance matrices in *CbCr* space. $\mathbf{I} = (Cb, Cr)$ is the image pixel under consideration.

The lip region is surrounded by skin area. Hence, probability of an image pixel to be a lip-pixel in presence of skin pixels is given by—

$$P_{Lip/Skin} = p_{Lip} P_{Lip} - p_{Skin} P_{Skin} \qquad (8)$$

The above value is normalized to give $\mathrm{S}_2 \in [0,1]$. The plot of $\mathbf{S_2}$ is shown in figure 4.

$$S_2 = \left\{ P_{lip/skin} - \min\left(P_{lip/skin}\right) \right\} / \max\left(P_{lip/skin}\right) \qquad (9)$$

$$S_3 = Y \log(1 + K S_2) / \log(1 + K) \qquad (10)$$

Where, *Y* is the *Y*-component of *YCbCr* representing intensity (similar to grayscale), and K is a parameter that decides the relative importance to be given to color information. A large value of K implies less importance to color and more importance to intensity information. S_3 is the scalar value of the pixel in the transformed image. Thus, calculating this transform for each pixel, we get a new image $\mathbf{S_3}$ that is lip-enhanced.

To localize lip region in the face image, K should be small (e.g. K = 1). However, to study features inside lip or find exact curves after localization of lip, K should be made large (e.g. K = 100). Thus, transform has the flexibility given variable level of features within lip.

While calculating edges either for segmenting lips or for approximating lips with curves, it is desirable that the edges occur only at the boundary of lip and skin but not in other regions of the face. The equations (8), (9) and (10) tilt the projection plane in such a way that the gradient while moving from a skin pixel to a lip pixel is maximized in this plane. Therefore, the lip and skin regions are demarcated more accurately. The gradient while moving from skin pixel to a lip pixel is shown by an arrow in figure 4. Not only the gradient is maximized, but also the absolute value of S_2 (eq. 9) is maximum for the lip pixels. This makes the transform suitable for various applications like edge projection, template based methods, or binarizing by thresholding. Previously [13-16], Mahalnobis distance (i.e. the distance from the Gaussian) served as a criterion to identify a lip pixel, without involving any concept of maximizing the gradient. The improvement in lip region segmentation on applying the lip enhancement transform is shown in figure 5.

4. FEATURE VECTOR AND CLASSIFICATION

A spatio-temporal representation of the face is created, based on geometrical relationships between features using Euclidean distance (figure 6). Such a representation allows robust handling of partial occlusion. Seven parameters form the feature vector **F**—

$$\mathbf{F} = \{H_e, W_e, H_m, W_m, R_{ul}, R_{ll}, N_L\} \qquad (11)$$

All components of the vector are normalized against the first frame to achieve scale independence. Radii of curvature of the upper and lower lips R_{ul} and R_{ll} are computed by approximating the binary mask of the lips with two parabolas. N_L is the number of distinct peaks detected for upper and lower lips during edge projection analysis, indicating whether mouth was open or closed. The change in feature vector **F** when the face undergoes change from neutral state to some expressional state—

$$\Delta\mathbf{F} = \{\Delta H_e, \Delta W_e, \Delta H_m, \Delta W_m, \Delta R_{ul}, \Delta R_{ll}, \Delta N_L\} \quad (12)$$

Such dynamic characteristic of the feature vector provides shape-independence. $\Delta\mathbf{F}$ serves as an input to the classifier. The classifier is a feed-forward basis function net with one hidden layer (figure 7). The activation function (figure 8) implemented in the hidden layer is—

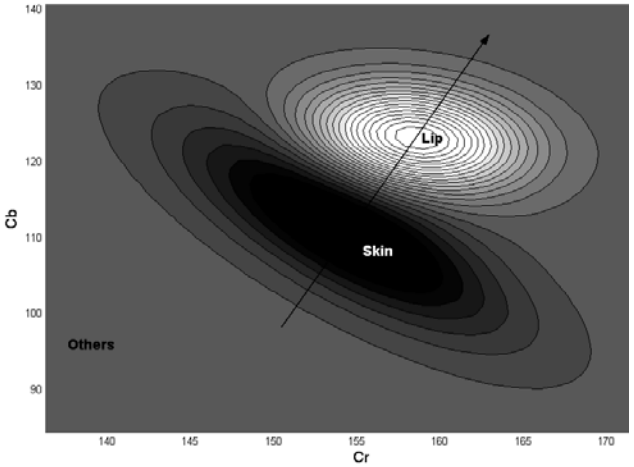


Figure 4. Contours for Normalized S_2 in Cb-Cr space. Legend: White = 1, Black = 0. Note the arrow shows the movement from an ideal skin pixel to an ideal lip pixel. Gradient is maximized in the direction indicated by the arrow.

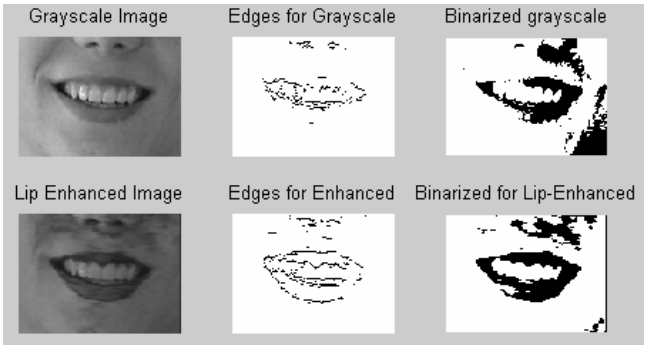
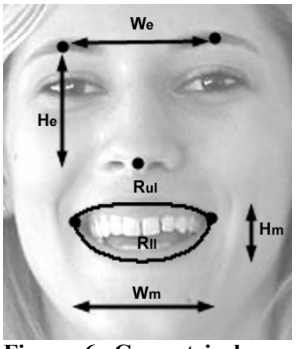


Figure 5. ***First Row:*** **(A) Color image converted to grayscale, (B) Edges for A, (C) Binary image obtained by thresholding A.** ***Second Row:*** **(A) Transformed color image, (B) Edges for A, (C) Binary image obtained by thresholding A.**
Notice the improvement in the second row, in terms of edges for lips, and binarized image.



H_e: Height of eyebrow
W_e: brows distance
H_m: mouth height
W_m: mouth width
R_{ul}: upper lip curvature
R_{ll}: lower lip curvature

Figure 6. Geometrical parameters of the face, forming the feature vector.

Where σ_i is the variance of i^{th} component of **F.** As compared to the standard sigmoid function, this activation function has a small value in a larger interval near zero. Thus, it makes the net tolerant towards small errors present in $\Delta\mathbf{F}$. For output layer,

$$O_i = sign(x_i)\left[1 - \exp\left(-\frac{x_i^2}{2\sigma_i^2}\right)\right] \quad (13)$$

$$Y_j = \sum_{i=1}^{7} w_{ij} O_i \quad (14)$$

Where, weights w_{ij} are correlation coefficients between O_i and desired output Y_j. The output at each node Y_j gives the confidence level of the corresponding expression.

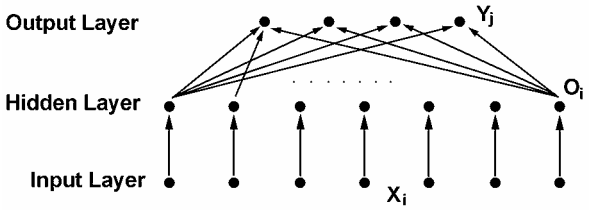


Figure 7. The basis function net for classification.

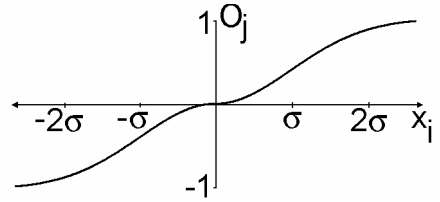


Figure 8. The activation function for the hidden layer.

5. RESULTS

Experiments were performed on colored as well as grayscale image databases. Cohn Kanade [8] database consists of grayscale image sequences depicting the four facial expressions— smile, surprise, sad and disgust. The subjects varied in ethnicity, age and skin color. A set of 50 sequences (825 images) was randomly selected as test samples. This database was used to test the accuracy of the facial expression recognition algorithm.

To test the effectiveness of lip enhancement transform in improving lip region extraction in case of colored images, 93 images from AR [9] and CVL [10] database (containing considerable lightening variations) were used. In the following sections, we describe the accuracy obtained in the conducted tests.

5.1. Feature Extraction

The accuracy of facial feature extraction for colored and grayscale images is shown in figure 9. In case of colored images, the lip enhancement transform was applied to the images.

An average accuracy of 92.2% was obtained for grayscale image database (data size=825 images). In case of color image database, a slightly better accuracy of 95.4% was obtained (data size = 93).

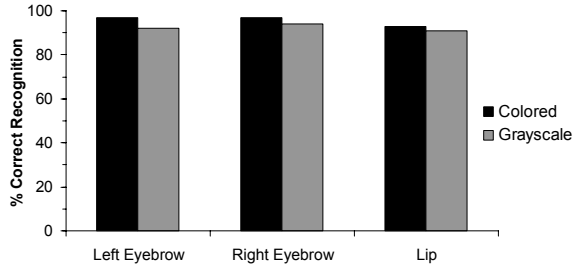


Figure 9. Accuracy for facial feature extraction

5.2. Facial expression recognition

Each grayscale image sequence in the database depicted one of the expression classes (smile, surprise, sad and disgust against neutral). The first image in the sequence was a neutral image. Confidence level of each expression was calculated for each of the subsequent images against the neutral image. The calculated vector of confidence levels was added to give total confidence for each of the expressions. Expression having the highest total confidence level was declared as the expression of the sequence.

On a test set of 50 sequences (825 images), an accuracy of 90.0% was achieved for grayscale images (figure 10). Bourel [3] has reported an accuracy of 89% over the same database. Their system has used feature point tracking followed by k-nearest neighbor algorithm. The holistic approach of Yacoob [11] has reported an accuracy of 89%.

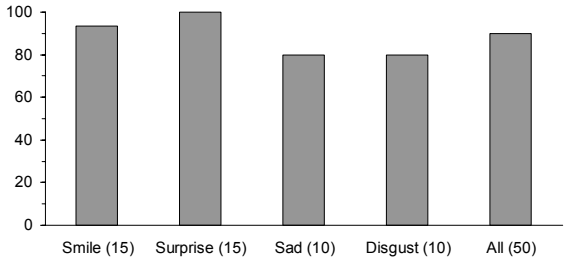


Figure 10. Accuracy for grayscale image sequences.

6. CONCLUSION

An efficient, local image-based approach for extraction of intransient facial features and recognition of four facial expressions was presented. A lip-enhancement transform for better segmentation of lip region in color images was proposed.

Our system shows superior performance in comparison to the other facial expression recognition systems. The system requires no manual intervention (like initial manual assignment of feature points, as in system described by Bourel [3]). The system, based on a local approach, is able to detect partial occlusions also.

7. ACKNOWLEDGEMENT

We thank Dr. Sumana Gupta for helping us in formalizing the statistical approach for lip enhancement transform.

8. REFERENCES

[1] B. Fasel, et al. Automatic Facial Expression Analysis: A Survey. Pattern Recognition, 36, 259-275, 2003

[2] I. Essa, A. Pentland. Coding, analysis, interpretation and recognition of facial expression. IEEE Trans. Pattern Anal. Mach. Intel., 19(7), 757-763, 1997.

[3] F. Bourel, et al. Robust Facial Expression Recognition Using a State-Based Model of Spatially-Localised Facial Dynamics. Proc Fifth IEEE Int'l Conf on Automatic Face and Gesture Recognition (FGR-02), Washington D.C., 113-118, 2002.

[4] M. Black, et al. Recognizing facial expressions in Image Sequences using local parameterized models of image motion. Int'l J. Comp Vision, 25(1), 23-48, 1997.

[5] Z Zhang, et al. Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron. Proc Third IEEE Int'l Conf on Automatic Face and Gesture Recognition, Nara, Japan, 454-459, 1998.

[6] R. Brunelli, et al. Face Recognition: Feature versus Templates. IEEE Trans Pattern Anal. Mach. Intel., 15(10), 1042-1052, Oct 1993.

[7] D. Chai, et al. Locating Facial Region in Head and Shoulder colored image. Proc. Third IEEE Int'l Conf of Automatic Face and Gesture Recognition, Nara, Japan, 124-129, 1998.

[8] Kanade T., et al. Comprehensive Database for Facial Expression Analysis. Proc. Fourth IEEE Int'l Conf on Automatic Face and Gesture Recognition (FG'00). Grenoble, France. March 2000.

[9] Martinez AM, et al. The AR Face Database. CVC Tech Report #24, 1998.

[10] CVL Face Database. Available: [Online] http://lrv.fri.uni-lj.si/facedb.html, ŠCV, PTERŠ, Velenje.

[11] Y Yacoob, et al. Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow. IEEE Trans on Pattern Anal. Mach. Int., vol 18, no. 6, 1996.

[12] Sadeghi M., et al. Segmentation of Lip pixels for lip tracker initialization. IEE Proc on Vision, Image and Signal Proc., vol 149 (3), 179-184, June 2002.

[13] Lievin, M.; Luthon, F. Lip features automatic extraction. Proceedings International Conference on Image Processing, ICIP 98, vol. 3, 168-172, 4-7 Oct. 1998.

[14] Lievin, M.; Delmas, P.; Coulon, P.Y.; Luthon, F.; Fristol, V. Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme. IEEE International Conference on Multimedia Computing and Systems, vol. 1, 691 -696, 1999.

[15] Kaucic, R.; Reynard, D.; Blake, A. Real-time lip trackers for use in audio-visual speech recognition. IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication (Digest No: 1996/213), 3/1 - 3/6, 28 Nov. 1996.

[16] Ramos Sanchez, M.U.; Matas, J.; Kittler, J. Statistical chromaticity-based lip tracking with B-splines. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-97, vol. 4, 2973 -2976, 21-24 April 1997.