AUTOMATIC FACIAL FEATURE DETECTION FOR FACIAL EXPRESSION RECOGNITION

Taner Danisman, Marius Bilasco, Nacim Ihaddadene and Chabane Djeraba
LIFL - UMR CNRS 8022, University of Science and Technology of Lille, Villeneuve d'Ascq, France

Keywords: Facial Feature Detection, Emotion Recognition, Eye Detection, Mouth Corner Detection.

Abstract: This paper presents a real-time automatic facial feature point detection method for facial expression recognition. The system is capable of detecting seven facial feature points (eyebrows, pupils, nose, and corners of mouth) in grayscale images extracted from a given video. Extracted feature points then used for facial expression recognition. Neutral, happiness and surprise emotions have been studied on the Bosphorus dataset and tested on FG-NET video dataset using OpenCV. We compared our results with previous studies on this dataset. Our experiments showed that proposed method has the advantage of locating facial feature points automatically and accurately in real-time.

1 INTRODUCTION

Over the last quarter century, there is an increased body of research on detection of facial feature points (e.g. locations of eye, eyebrow, and mouth) on different domains including face recognition, facial expression recognition, facial animation, head pose estimation etc. All of these application domains need efficient, accurate and fast detection of facial features. On the other hand, there are many factors that effects desired functionality including different lighting and illumination conditions, background changes in outer boundaries, high variation in static and dynamic parts in face and shadows.

In literature, research on facial feature detection can be grouped into color intensity based, projection based, segmentation based, classifier based (Neural Networks, Support Vector Machines, Bayesian Classifiers) and deformable template based approaches. Intensity and projection based approaches use a specific color model to get benefit from the intensity changes in facial area. Usually vertical and horizontal integral projection applied to get peak and/or valley points which presents the availability of facial features. These methods assume that skin color in facial area is homogeneous. However, lighting, illumination and pose changes affect the outputs of this approach in a negative way.

Segmentation based approaches also use the intensity values either in color or grayscale formats. Majority of the studies focus on skin color based segmentation of facial features especially for the mouth. In these approaches domain knowledge about the skin color is used. According to the mouth and lip segmentation studies, lip area includes more reddish colors than bluish colors. For this reason, researchers use nonlinear transformations to get benefit from this property. Unfortunately, this feature is not available all the time and previously mentioned problems still exists for this approach.

Classifier based approaches needs large training sets to learn facial features from extracted feature vectors. Huge amount of annotated data is need for better classification. Popular classifiers are neural networks, support vector machines and Bayesian classifiers.

Deformable template based approaches like Snakes (Kass, Witkin and Terzopoulos, 1987) and Active Appearance Models (Cootes, Edwards and Taylor, 1998) use an initial template object and deform it until reaching to the desired shape. Although these methods provide robust results, their operations are computationally complex to use in real time applications. In general, all of these approaches have advantages and disadvantages. This observation motivated us to use a hybrid approach to detect facial feature points.

In this study we used combination of image processing techniques (contrast enhancement, adaptive thresholding and quantization) and artificial neural networks to detect facial feature points in

Danisman T., Bilasco M., Ihaddadene N. and Djeraba C. (2010).
AUTOMATIC FACIAL FEATURE DETECTION FOR FACIAL EXPRESSION RECOGNITION.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 407-412
DOI: 10.5220/0002838404070412

grayscale images extracted from a given video. Our proposed system automatically detects the faces in a given video using tuned Viola-Jones face detector (Viola and Jones, 2001), then it detects the pupil positions using Neural Network based eye detector (Rowley, Baluja and Kanade, 1998) and estimates the orientation of the face according to the vertical position of left and right pupil. After correcting the orientation, it detects position of eyebrows and nose using adaptive-thresholding and finally left and right corners of mouth is detected using a trained Artificial Neural Network. Finally, we used the ratio of geometrical distance measures between corners of mouth and inter-pupillary distance (IPD) to detect the availability of an emotion. For our experimental study we considered neutral, happy and surprise emotions.

- For training, we used Bosphorus image dataset (Savran et al., 2008)
- For testing, we extracted images from FG-NET Wallhoff (2006) video dataset and compared our results with previous studies achieved on this dataset.
- Our approach uses both neural networks and adaptive thresholding for the facial feature detection.
- We considered eyebrows, center of pupils, nose and corners of mouth as facial feature points.

The rest of the paper is organized as follows. In section two, we explained our methodology used in this study. Section three presents the experimental results performed on FG-NET dataset. Finally section four concludes the study.

2 METHODOLOGY

Our hybrid approach consists of a set of independent but hierarchically dependent detection components shown in Figure 1 that works on different region of interests (ROI) of the image.

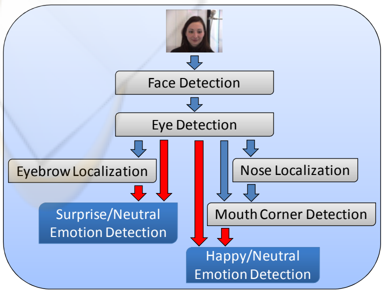

Figure 1: Dependency figure for our hybrid approach.

We start with the detection of the face using Viola-Jones face detector available in OpenCV library. We selected the scale factor 1.2, minimum neighbors 3 and set the flag CV_HAAR_FIND_BIGGEST_OBJECT to achieve the best speed and performance.

After that we used the neural network based eye detector Rowley (1998) available in STASM (Milborrow and Nicolls, 2008) library to locate the positions of pupils. STASM is a variation of Active Shape Model of Coote's implementation. STASM works on frontal views of upright faces with neutral expressions. Although it is highly robust on neutral expressions, it is not suitable for facial feature detection on faces with emotions. In addition, because of the high computational needs it is not suitable to use with real time applications. Therefore we derived only Rowley's eye detection code from the library which is a group of neural networks that provides eye positions. We did not use the Rowley face detection because of the speed constraints. Figure 2 shows original and superimposed images after the detection process.

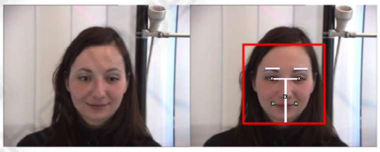

Figure 2: Face and facial feature detection (Sample from FG-NET database).

After the eye detection, we estimated the orientation of the face using the vertical positions of the two eyes. If they are not in the same position then we compute the angle between these two pupil points and simply corrected the orientation by setting the face center as origin point and we rotated the whole frame in opposite direction. In this way we guaranteed the frontal upright position of the face up to ±30 degree in both sideways.

2.1 Eyebrow and Nose Localization

For eyebrow and nose localization we performed adaptive thresholding technique. First we determined three ROI's that shows approximate positions of the two eyebrows and nose. After that, we performed contrast stretching. To do that first we find the low and high limit pixel values to contrast stretch image. In order to calculate these values first we compute the cumulative histogram of the detected face. After that we searched within the cumulative histogram to find index values that is closest to the 1% of all image pixels. More clearly, if the face size is 200×200 then total pixel values is 40,000 and 1% of it is 400. Thus, in the cumulative histogram we search for the number 400 from the top to find the low limit ($\mathrm{low_{in}}$) and we search the number 40,000−400=39,600 from the bottom to find the high limit (high_{in}). Let this numbers be the 20th and 230th index elements in the cumulative histogram. Then simply normalize these values by dividing them to 255 to get approximate intensity values in 8-bit scale which is ~0.07 and ~0.90 respectively. Finally, these numbers are then used to map desired scale which is 0 to 1. The following function is applied to the all pixels within the face. The γ value specifies the shape of the curve during the intensity mapping process where if it is greater than 1 then it produces dark values and if it is less than 1 it produces brighter values. During our experiments we used the γ value as 1 which means that our mapping is linear. Figure 3 shows the intensity transformation.

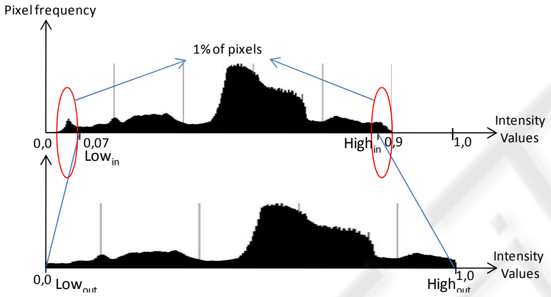

Figure 3: Intensity transformation.

$$\mathrm{I(x,y)} = \mathrm{pow}\left(\frac{\mathrm{I(x,y)} - \mathrm{low_{in}}}{\mathrm{err_{in}}}, \gamma\right) * \mathrm{err_{out}} + \mathrm{low_{out}} \qquad (1)$$

Where

$$\mathrm{err_{in}} = \mathrm{high}_{in} - \mathrm{low_{in}} \qquad (2)$$

$$\mathrm{err_{out}} = \mathrm{high_{out}} - \mathrm{low_{out}} \qquad (3)$$

Example output of this method is compared with the standard histogram equalization method in Figure 4. From the left to the right, original low contrast image, intensity transformed image and histogram equalized image is shown.

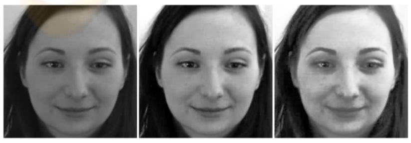

Figure 4: Original, intensity transformed and histogram equalized images (left to right order).

After the intensity transform, we got samples from the mid of forehead region and we compute the mean and variance values. Width and height of the forehead region is set to the quarter of the IPD. Center of this region (x1, y1) is represented by the center of the line between two eyes and quarter of inter-pupillary distance (IPD) above the vertical eye position. In order to detect the eyebrow and nose locations we compute the horizontal integral projections in previously selected ROI's as seen in Figure 5. We start searching position from top to bottom thus avoiding possible false positives. We use threshold value as $\overline{\mathrm{x}} - \sigma$ where $\overline{\mathrm{x}}$ represents the mean and σ represents the variance for each search region. ROI window sizes for eyebrows are set to the quarter of the IPD and a tenth of the IPD for nose ROI.

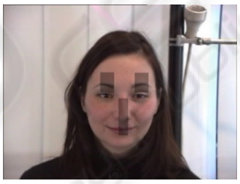

Figure 5: ROI areas for eyebrow and nose localization.

2.2 Mouth Corner Detection

Mouth is the most dynamic part of the face as its shape and size dynamically changes when we speak or show a facial expression. Research works (Ekman and Friesen, 1982) showed that there is a direct relation between happy emotion and the shape of the mouth. Therefore, detection of corners of mouth has an important role in emotion recognition. We used mouth corners for the detection of the "happy" emotion only. In order to detect corners we used artificial neural network based system having two hidden layers as shown in Figure 6.

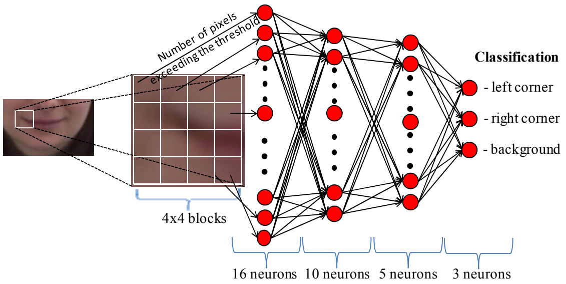

Figure 6: The structure of the neural network used for the detection of mouth.

Detection of mouth in neutral facial expression is easy. The simplest way is to use the integral projections. On the other hand, for real world situations where there is an emotion or person is speaking then this method does not provide robust results. In our study, we considered emotion invariant mouth corner detection which is much more difficult problem than traditional detection methods which considers only neutral faces with closed mouth. Therefore we used 2 hidden layers with more neurons than traditional approaches use in order to handle the variety of mouth shapes. The input layer has 16 neurons, first hidden layer has 10, second hidden layer has 5 and output layer has 3 neurons. We used sigmoid activation function.

We used 2D landmark files to extract corners of mouth from Bosphorus dataset (Savran et al., 2008) as our training set which consists of 2D and 3D facial coordinates of 105 subjects (61 male, 44 female). As the number of subjects is small, we ignored 16 beard men which have negative effect on training process. In addition, for each subject we ignored YR (Yaw rotation), O_EYE (Eye occlusion), O_MOUTH (Mouth occlusion), O_HAIR (Hair occlusion), CR_RD (Right Downwards), CR_RU (Right Upwards).

As our NN classifier is based on intensity values, each subject image in Bosphorus set have been converted to grayscale and then reduced to 250×250 pixels. Finally we crop 60×60 samples from these images before the feature extraction.

As our output layer has 3 neurons, one for the left mouth corner, one for the right mouth corner and third neuron is for the background. We used 2,000 positive samples for mouth left corner, 2,000 samples for mouth right corner and 40,000 samples for the background. Background samples selected from the near mouth corner area to eliminate possible false positives. Figure 7 shows samples for each class. First row includes samples from the mouth left corner, second row shows samples from the mouth right corner and last row shows background samples.

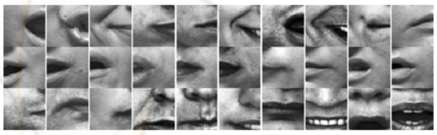

Figure 7: Samples from our training set (Left corner, right corner and background samples in each row).

Each sample is preprocessed to enhance the contrast as described in section 2.1. After that we compute the average image from the 2,000 positive samples as shown in Figure 8. According to the gray level properties of the average image, we manually set the gray-level threshold value $T = 50$. Then, each image is converted to binary image using this threshold value T.

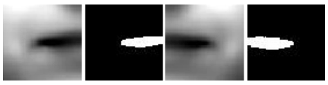

Figure 8: Average and segmented images for mouth corners from Bosphorus dataset.

Finally we divided these binary images into 4×4 sub blocks and for each block we compute the cumulative value. As a result we get a feature vector from each sample having 16 bins which is then used as an input for the input layer. After that we trained our NN model using sigmoid activation function.

Prior to testing, we need to preprocess captured frames so that they are processed in the same way that we used in training. As the frame size in our test set is not equal in Bosphorus dataset, we have to scale both detected face and sample size in the test frame. In Bosphorus dataset we have faces of size 250×250 and sample size of 60×60. On the other hand Viola-Jones face detector gives different face sizes. For this reason we resized the detected faces to 100×100 pixels and we used 24×24 pixels for the sample size by holding the aspect ratio.

One of the problems with neural network based classification approach is that searching each pixel increases the computation time. Therefore we set two ROI for the left and right mouth corner. Position of these regions depends on the vertical location of detected nose and IPD distance. Let the position of this ROI is represented by x, y, w, h. Vertical position y is set to the nose position plus 6% of face height. Horizontal position x is set to 10% of the face width for the left ROI and 45% of the face width for the right ROI. Width of the ROI w is set to 23% of face width and height h is set to 50% of the IPD distance. Figure 9 shows coarse coordinates of candidate mouth corners.

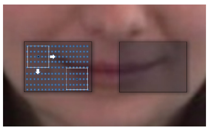

Figure 9: Example mouth search ROI image from FG-NET video dataset.

In order to classify mouth corners, we need to get samples from these ROI's. If there is more than one candidate exists for the corner points then we select the point with the maximum network output .

2.3 Facial Expression Recognition

Facial expression recognition is simply a classification of a human face according to predefined emotional classes. We created static rules that consider facial geometry for classifying different emotions. In this research, we used happiness, surprise and neutral classes for the classification. We examined the following measurements for emotion classification of a face;

$$Frame(I) = \begin{Bmatrix} Happy & \dfrac{Mouth_{width}}{IPD} \geq 0.8 \\ Surprise & Eye_y - EBrow_y \geq \dfrac{IPD}{1.9} \\ Neutral & otherwise \end{Bmatrix} \quad (4)$$

where

$$Mouth_{width} = MouthRightCorner_x - MouthLeftCorner_x \quad (5)$$

In this approach it is possible to assign more than one emotion class per frame. In order to classify a given face, we measured the ratio of facial features to the IPD value. If it exceeds the threshold values then we label the face with the corresponding emotion class. Threshold value for the happy emotion is computed by manually examining the samples from the Bosphorus dataset and for surprise emotion, by manually examining our faces.

3 EXPERIMENTAL RESULTS

In our research study, we used FG-NET (IST-2000-26434) dataset for testing purposes. It consists of 7 different emotions from 18 subjects where each subject repeats each emotional expression three times. As we selected to study on three classes then we used only 54×3=162 videos. Figure 10 shows sample frame sequence for happiness emotion from FG-NET dataset. In FG-NET dataset, minimum annotated unit is the video itself. Thus, our evaluation on this dataset is based on number of videos that correctly classified.

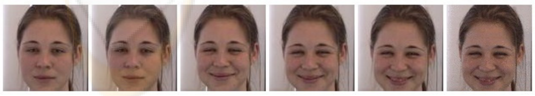

Figure 10: Sample cropped happiness frame sequence from FG-NET dataset.

These videos usually start with neutral face and in general the emotion appears in the middle of the video. As a result, there are many neutral faces exist in each video. To solve this labeling problem, we assigned a single emotion label for each frame and we compute the emotion histogram per video. Then we selected the emotion from the histogram with the second biggest histogram value. Because the top value in histogram shows the neutral emotion. In addition, we used a threshold value for the minimum number of frames that must be reached to classify a video which is 10% of total frames in the video. It means that if the video has 120 frames then, the second biggest value in the emotion histogram must be greater than or equal to the threshold value which is 12 frames. If there is no such emotion exists then it is classified as neutral emotion. More formally;

$$Video(I) = \begin{Bmatrix} emoClass(h2[1]) & h2 \\ neutral & otherwise \end{Bmatrix} \quad (6)$$

where $h2 = sort(hist(I)) | h2[1] \geq 0.1 * TF$ (7)

and TF is the total frames in $Video(I)$. Table 1 shows the confusion matrix and accuracy obtained from our experimentations using FG-NET dataset.

Table 1: Confusion matrix for FG-NET dataset.

Predictions / Actual	Accuracy in %		
	Happy	Surprise	Neutral
Happy	74,1	11,1	14,8
Surprise	68,5	22,2	9,3
Neutral	3,7	7,4	88,9

Experiments showed that there is a clear confusion between surprise and happy emotions which also supports previous studies. This is because of the fact that subjects show happy emotion while expressing surprise emotion when they see a surprising event. This is a general human behavior where a surprise event is turned out to be happy emotion as expressed in Izard, (1991). Moreover, some of the happy videos are classified as surprise videos. When we look at to these misclassified videos we see that there is a hair occlusion on forehead region which creates false positive results for eyebrow localization. In a similar way some of the neutral videos misclassified as surprise emotions because of the hair occlusion as seen in Figure 11.

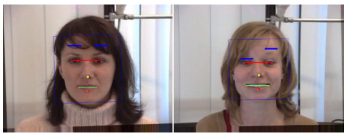

Figure 11: Example false positive detections for eyebrows due to hair occlusions.

We compared our results with Cerezo and Hupont, (2006). They used 10 characteristic MPEG4 feature points to extract emotional information to detect six basic emotions. They worked on static images and manually selected facial points. They tested Hammal's method (Hammal, Couvreur, Caplier and Rombaut, 2005) on FG-NET dataset. Although the evaluation criteria and feature extraction method are not the same (manual vs. automatic), Table 2 shows general information about results achieved on FG-NET dataset.

Table 2: Previous studies on FG-NET dataset.

Method	Feature	Accuracy in %		
	Extract	Happy	Surprise	Neutral
Hammal's method	Man.	87,2	84,4	88,0
Cerezo and Hupont,2006	Man.	36,8	57,8	100,0
Our method	Auto	74,1	22,2	88,9

In terms of speed, our hybrid approach runs at 28fps on Intel Core2Duo 2.8 GHz laptop for 26fps mpeg sized videos. In addition current prototype delivers 24.5fps for webcam video frames of size 640×480. Therefore the suggested approach is suitable for real-time processing.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a hybrid approach for facial feature detection for emotion recognition in video. Our system is detects seven facial feature points (eyebrows, pupils, nose, and corners of mouth) from grayscale images.

Experimental results showed that our system works well on faces with no occlusions thus we get acceptable emotion recognition results. On the other hand, different occlusions on facial area slightly affect the performance of the system.

As future work, we are planning to detect finer locations for eyebrows and radius of the pupillary area in terms of feature extraction and planning to work on hard cases (hair occlusion, etc.). In case of eyebrows, the shape of the eyebrow will give useful information about different emotion.

ACKNOWLEDGEMENTS

This study is supported by the Multimodal Interfaces for Disabled and Ageing Society (MIDAS) ITEA 2 – 07008 project.

REFERENCES

Cerezo, E. and Hupont, I., (2006). Emotional Facial Expression Classification for Multimodal User Interfaces. *LNCS,* (Vol. 4069, pp. 405-413).

Cootes, T. F., Edwards, G. J. and Taylor, C. J. (1998). Active appearance models. In *H.Burkhardt and B. Neumann, editors, 5th European Conference on Computer Vision*, (Vol. 2, 484–498), Springer, Berlin.

Ekman, P. and Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6, 238–252.

Hammal, Z., Couvreur, L., Caplier, A. and Rombaut, M. (2005). Facial Expressions Recognition Based on the Belief Theory: Comparison with Different Classifiers. *In Proceedings of 13th International Conference on Image Analysis and Processing*, Italy.

Izard, C. E. (1991). *The psychology of emotions*. New York: Plenum Press.

Kass, M., Witkin, A. and Terzopoulos, D. (1987). Snake: Active Contour Model, *International Journal of Computer Vision*. (Vol. 1, pp. 321-331).

Milborrow, S. and Nicolls, F. (2008). Locating facial features with an extended active shape model. In: *Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS,* (Vol. 5305, pp.504–513). Springer, Heidelberg.

Rowley, H. A., Baluja, S. and Kanade T. (1998). Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (Vol. 20, p. 23-38), http://vasc.ri.cmu.edu/NNFaceDetector

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B. and Akarun L. (2008). Bosphorus Database for 3D Face Analysis, The First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008), Roskilde University, Denmark.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *In Proceedings of Computer Vision and Pattern Recognition*, (Vol. 1, pp. 511–518).

Wallhoff, F. (2006). FG-NET Facial Expressions and Emotion Database. Technische Universität München. Retrieved from: http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html.