AUTOMATIC ROBUST LOCALIZATION OF EYE RELATED LANDMARKS

Alessandra BANDRABUR[1], Laura Maria FLOREA[2],
Raluca BOIA[3], Corneliu FLOREA[4]

In this paper we present an accurate and robust framework for automatic localization of landmarks in the eye region. The system is based on compressed projections used for describing the regions of interest on the face. We start from an initial anthropometric extracted seed point and we search the exact landmark location around it. The investigated neighboring points are described by concatenation of integral and edge projections, which are later reduced through Principal Component Analysis technique. Landmark localization is performed by classifying these various candidates through a Multi-Layer Perceptron. The method is evaluated on Cohn-Kanade and BioID databases.

Keywords: eye localization; compressed image projections; feature descriptors

1. Introduction

The eyes are, probably, the most salient features of the human face. They are crucial in non-verbal communication or for recognizing and understanding the emotional states of humans. As shown in [1], [2], [3] there is an increasing interest on the eye location as they are a significant component in human computer interaction. Eye localization is used in applications as face alignment, face recognition, human computer interaction, gaze estimation and control devices for disabled people. Specific and distinct applications focused on the discrimination of spontaneous versus posed facial expressions, specifically in detection of fake smiles [4], [5]. An example of a practical application is in online (remote) interviewing via web-cam video transmission where the interviewer runs an application hinting at interviewed emotion states while answering to various queries.

[1] PhD student, The Image Processing and Analysis Laboratory, LAPI, University POLITEHNICA of Bucharest, Romania, e-mail: abandrabur@imag.pub.ro

[2] Lecturer, The Image Processing and Analysis Laboratory, LAPI, University POLITEHNICA of Bucharest, Romania, e-mail: laura.florea@upb.ro

[3] PhD student, The Image Processing and Analysis Laboratory, LAPI, University POLITEHNICA of Bucharest, Romania, e-mail: rboia@imag.pub.ro

[4] Lecturer, The Image Processing and Analysis Laboratory, LAPI, University POLITEHNICA of Bucharest, Romania, e-mail: corneliu.florea@upb.ro

While initially only the center of the iris was of interest, more recently, additional fiducial points (such as corners, bottom and upper limit of the eye socket) are searched for as they can be used for more extended applications.

In this paper we aim to solve a special case of face fiducial points (i.e. landmarks) localization. More exactly, the purpose is to localize 7 points for each eye, to a total of 14 points for a given face (illustrated in Fig. 2). These points are represented by two points on the eyebrows: the inner and outer corner of the eyebrow and by five points of the eye: the left, right, top, bottom corners and center of the eye. These points hold the most important information needed by specific applications and a further test showed [6] that are sufficient re separate the eye related AU.

The face landmarks localization methods can be divided into two distinct categories: intrusive techniques, which imply physical equipment installed on the user's head and image processing techniques, which imply regular cameras for capturing images of the face. The first category provides highly accurate information, but they require intrusive and expensive sensors [7]. On the other hand, the non-intrusive techniques should function real-time, with minimum calibration and under natural head movement.

This paper follows the second scenario assuming near frontal face with remote passive illumination and camera acquisition. There are two methodologies for the appearance based eye locators: model and feature based methods category. A review of the most relevant methods may be followed in [8]. In general, the model based methods employ the holistic appearances of the eye or of the face. Using the global appearance, these methods are not very accurate for the eye localization. Feature based methods make use of the eye properties such as symmetry and employ local image features, like corners, edges or gradients without requiring any model fitting. Therefore, these methods can be very accurate, if the eye areas are not affected by great levels of noise.

Regarding the state of the art, the primary facial landmarking techniques are developed from Active Appearance Models (AAM) [9] and Elastic Graph Matching [10]. The globally optimization from AAM was improved by turning into a local one by using the independent models from Constrained Local Models [11]. The iterative use of approximate matching algorithms is inquired for the facial connected spatial model by Valstar *et al.* [6] through Support Vector Machine regressed feature point location with the aid of conditional Markov Random Fields in the so-called BoRMaN algorithm.

The proposed method is developed from the iris localization procedure detailed in [12] and is a direct expansion of the method in [13]. In our previous work, by means of Principal Component Analysis we reduced the redundancy of image patch descriptors, namely the concatenation of integral and edge projections. This is followed by a classification of the descriptors using a multi-layer perceptron (MLP) network. We proved that starting from the positions provided by BoRMaN, we can increase the localization accuracy by 5-10%.

Within this proposal we differ from the method from [1] by turning the process into a fully automatic one, without initialization using BoRMaN results, but using instead locations extracted from the face detected square. Also by further elaborating the pre and post-processing methods, a higher accuracy is achieved.

The rest of paper structure is as follows: section 2 reviews various types of image projections as image patch descriptors, section 3 describes the overall proposed method, section 4 presents the achieved results and finally section 5 concludes and specifies possible continuation paths.

2. Image projections

An image region can be described by the integral projections. As one can see in Fig. 1, the region is depicted both by vertical and horizontal projections.

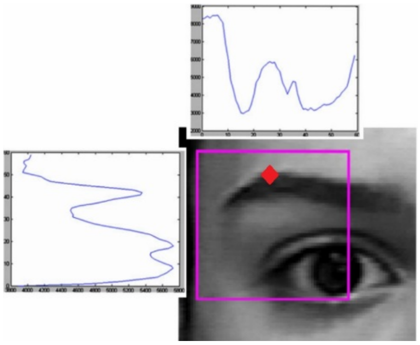

Fig. 1. Eye landmark projections

A. Integral Image projections

Integral projection functions are very powerful descriptors [14]. A gray-level sub-image *I(i,j)* with $i=i_1 \ldots i_2$ and $j=j_1 \ldots j_2$ has the projection on horizontal axis $P_H(j)$, which is the average gray–level along the columns, and the vertical axis projection $P_V(i)$ equal to the average gray-level along the rows:

$$P_H(j) = \frac{1}{i_2 - i_1} \sum_{i=i_1}^{i_2} I(i,j), \forall j = \overline{j_1, j_2}\ , \qquad (1)$$

$$P_V(i) = \frac{1}{j_2 - j_1} \sum_{j=j_1}^{j_2} I(i,j), \forall i = \overline{i_1, i_2}\ , \qquad (2)$$

B. Edge projections

The same technique as in the case of integral image projections can be used to compute the edge projections. The difference is that instead of using the original image, we will use an image containing the contours of the original image. In order to get the edge image, the original image is filtered using the well-known Sobel edge detector, resulting an image of edge magnitude *S(i, j)*. The final horizontal and vertical projections S_H and S_V are computed:

$$S_H(j) = \frac{1}{i_2 - i_1} \sum_{i=i_1}^{i_2} S(i,j), \forall j = \overline{j_1, j_2}\,, \tag{3}$$

$$S_V(i) = \frac{1}{j_2 - j_1} \sum_{j=j_1}^{j_2} S(i,j), \forall i = \overline{i_1, i_2}\,, \tag{4}$$

$$S(i,j) = S_H^2(i,j) + S_V^2(i,j)\,, \tag{5}$$

3. Algorithm

We aim to locate the eye related landmarks which are exemplified in neutral and respectively surprised faces as shown in Fig. 2. The first step of our algorithm is face detection. Faces are detected using the classical face detector Viola-Jones [15] and are further scaled to 300×300 pixels (which is typical size for a face framed in HD video transmission).

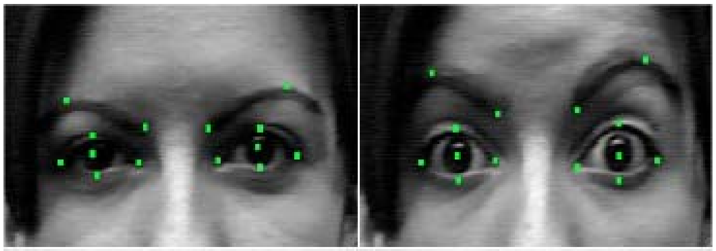

Fig. 2. Eye landmarks modelling emotional expressions

A. Region feature descriptor

For every eye landmark, we will extract a 60×60 pixels region around it, as can be seen in Fig. 1. The initial region position is extracted from the face square based on geometrical/anthropometrical criteria.

For the extracted region we will compute the integral projections and the edge projections, both horizontal and vertical, resulting four vectors P_H, P_V, S_H, S_V. These vectors are normalized to [-128; 127] range and then concatenated, resulting in a full vector with 240 elements. Its dimensionality is reduced by PCA to 50 elements, obtaining the proposed feature descriptor which will be further used. This workflow is visually depicted in Fig. 3.

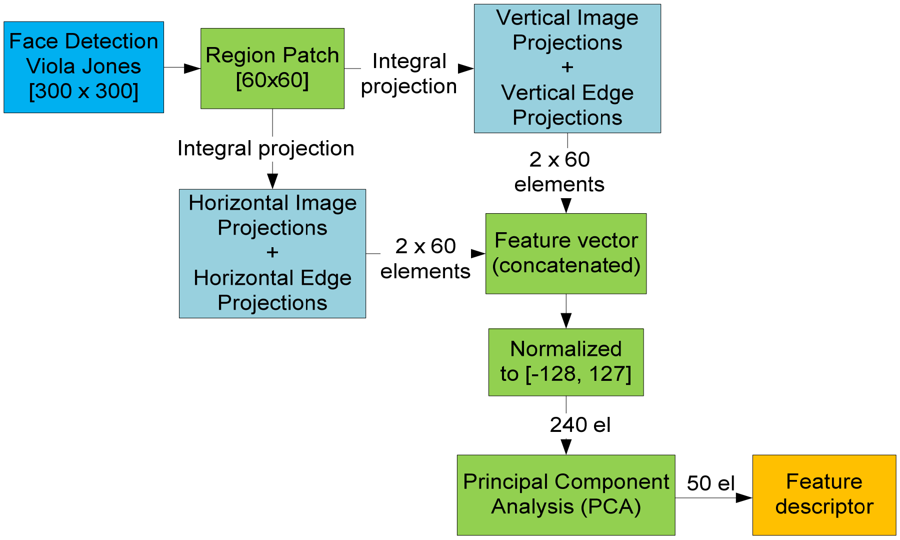

Fig. 3. Proposed scheme for computing the feature descriptor

B. Training

The classifying stage is ensured by a multi-layer perceptron (MLP) network with one hidden layer of 30 neurons (Matlab “patternnet” trained with “trainrp”). For training, we use images from public databases as will be discussed in the future section. The training data is formed by randomly selecting 30% of the total number of images in the databases.

For each point (manually annotated – ground truth) from the training images we compute the proposed feature descriptors. In order to enlarge the training set we consider the points that are at a 2 pixel distance from the real eye landmark point to be the positive ones, while the negative ones are positioned at a 40 pixels distance from the real point as can be seen in Fig 4. The distances are chosen with respect to 300x300 face size and to provide a balanced training dataset. These descriptors will be the training set for the MLP classifier.

C. Testing and scanning

We will use for testing the remaining images of the databases. The starting points are computed using anthropometric information using the procedure described in [16] for the iris center and extended by us to each landmark type.

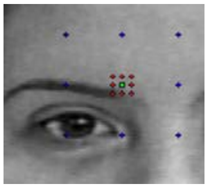

Fig. 4. Positive data set is marked with red, while the negative data set with blue and the ground truth with green

Supposing that these starting points are accurately positioned only to a certain degree, we consider a 20 pixel neighborhood area around each of them and we take 24 neighbors in this area (±10 pixels from the marked point, with a step of 5 pixels, since we observed on the training database that the maximum approximation error is ±10 pixels). These will be the candidates for the final desired landmark point, as can be seen in Fig. 5. These 25 points will be processed and one of them will be the winner for this stage of the algorithm.

In addition to the method in [1] we decided to refine the obtained location so we used the winner as the new starting point and we considered 10 pixels vicinity around it, we took 36 neighbors in this area distanced with a smaller step of 2 pixels. The same *process*, which we are going to describe below, will be used to estimate the final desired eye landmark.

i. Preprocessing

Given the starting point and its vicinity, we will remove those candidates which are clearly not landmark points. To do this we note that landmarks are consistently darker than most points in the vicinity as they are placed on hairy parts; neighboring skin pixels are expected to have higher intensities. Thus any initially presumed landmark position that has the intensity higher than a certain threshold, computed as a percentage of the average value of the region of interest, is removed as it does not have potential to be a valid candidate. This step also considerably speeds up the entire method.

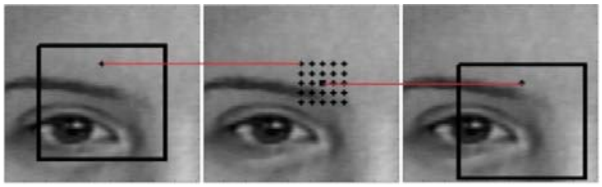

Fig. 5. Starting point vicinity at a distance of 10pixels with a 5pixels step and its selected areas

ii. Process

After the data is *preprocessed*, for each resulted point we will compute the proposed feature descriptor and we will feed it into the trained MLP. The MLP responses will be stored in a classification map, which will be further *post-*

processed. The final detected point (the winner) will be the weighted center of mass of the classification map. The entire algorithm can be seen in Fig. 6.

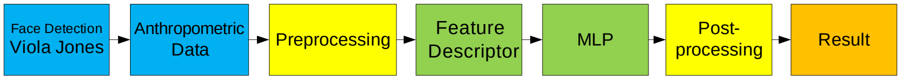

Fig. 6. Algorithm scheme

iii. Post-processing

The classification map contains the MLP responses of the candidates to the desired eye landmark. These responses are values between [0;1], where numbers closer to 1 are more likely to be the desired point, and those closer to 0 are less likely to be the winner, as exemplified in Fig. 7. In order to remove some of these responses, to prevent their contribution to the final decision, we filter the map, removing candidates with lower responses.

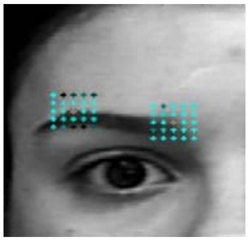

Fig. 7. Classification matrix. Black represents the smallest value (0) and cyan represents the maximum value (1) of the classification mapping.

4. Implementation and results

A. Databases

The algorithm is tested on two databases, namely the BioID database [17] and the Cohn-Kanade database [18]. The BioID database includes 1521 gray-level images, with the near frontal view face of 23 different subjects, taken in natural conditions with various backgrounds and different illuminations. The pictures are 384×286 gray scale images. The images have 20 face landmarks manually annotated, as can be seen in Fig. 8.

The Cohn Kanade dataset consists of 486 sequences from 97 posers. All sequences start with a neutral expression and proceed to a peak expression (apex). We consider only the neutral pose and the apex, resulting, thus, in 972 gray-level images of 640×490 or 640×480 pixels. The images contain frontal view faces with uniform illumination. The images have manual annotations of facial fiducial points [19]. These annotations contain 59 coordinates, as can be seen in Fig. 8.

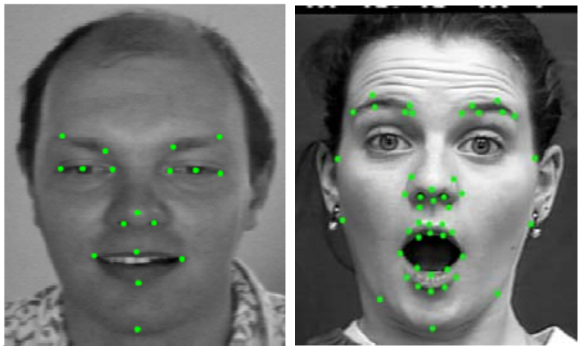

Fig. 8.Database annotated face images. The left image is from BioID database and the right image is from Cohn Kanade database.

The used expressions consist of the six basic universal emotions: happiness, anger, fear, surprise, sadness and disgust. These emotions are simulated. The apex is coded using Facial Action Coding System [20].The system is based on the fact that facial muscles position gives a great description of the basic emotions, which are thus described by facial Action Units.

B. Results

The system's performance is evaluated according to the stringent criterion for eye centers [21], ε, and the proximity measure for multiple landmarks [22], me. The point is correctly determined if the specific error is smaller than a threshold. The error is computed as:

(6)

where $\varepsilon_{L/R}$ is the Euclidean distance between the ground truth left/right landmark and the determined left/right landmark and D_{eye} is the distance between the ground truth eye centers, as one can see in Fig. 9.

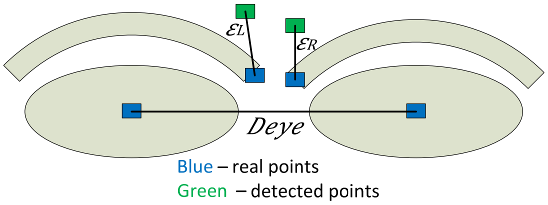

Fig. 9. Accuracy measure is the error of the Euclidian distances between the located eyes and the ones in the ground truth normalized by the distance between the real eye centers

The measure for proximity for ***t*** landmarks has the following formula:

$$m_e = \frac{1}{t \cdot D_{eye}} \sum_{i=1}^{t} \varepsilon_i \qquad (7)$$

where $\boldsymbol{\varepsilon_i}$ are the point to point errors for each landmark location, ***t*** is the number of searched points; in this case *t = 14* (7 for each eye).

The results for every landmark point are displayed in Table 1 for various thresholds: 0.05, 0.1 and 0.25, which are the usual choices as they have some meaningful interpretations: 0.05D_{eye} is the width of the eye pupil, 0.1D_{eye} is the width of the iris and 0.25 is the width of eye (sclera).

Table 1

The accuracy (Acc.) of the proposed algorithm for the eye landmark points

Eye landmarks	Acc. ε< 0.05	Acc. ε< 0.1	Acc. ε< 0.25
Eyebrow outer corner	17.71	65.42	99.14
Eyebrow inner corner	19.42	70.57	99.14
Eye outer corner	23.42	81.71	100
Eye inner corner	33.42	97.14	100
Eye upper limit	38.85	92.28	99.71
Eye bottom limit	32	91.14	100
Iris center	42	93.71	100

To establish the optimal parameters of the system we performed extended experiments. Concerning the parameters of the filters in the steps before and after the classification process, the best threshold to be used for the preprocessing removal of points is 10% of the average intensity level of the crop, as can be observed in the tests summarized in Table 2, while the best threshold for the classification map is achieved for a value 0.2, as shown in Table 3.

Table 2

The accuracy (Acc.) influenced by the variation of preprocessing gray-level threshold, represented by a specific percentage of landmark area average value

Gray Level Threshold	10%	50%	80%
Acc. ε< 0.05	32.28	20.98	12.11
Acc. ε< 0.1	95.41	70.57	59.55
Acc. ε< 0.25	99.85	83.98	71.54

Regarding the state of the art comparison, we report the results achieved by the method from [7] and our older method from [1]. Compared to the BoRMaN [7] method, we report an increase in accuracy of 19%. The comparative results can be seen in Table 4 and in Fig. 10.

Table 3

The accuracy (Acc.) influenced by the variation of post-processing classification threshold

Classification Threshold	**0.2**	**0.5**	**0.8**
Acc. ε< 0.05	***34.12***	30.01	17.14
Acc. ε< 0.1	***98***	86.25	69.63
Acc. ε< 0.25	***100***	99.98	80.12

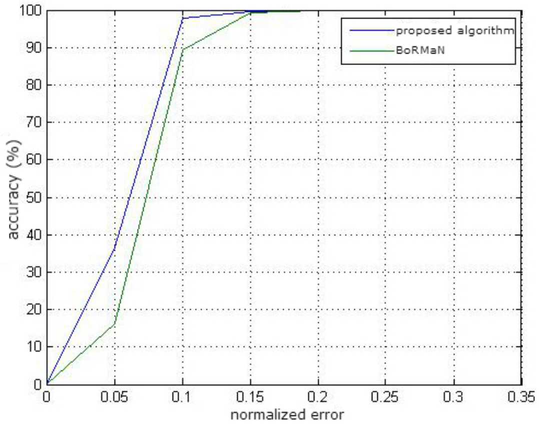

Fig.10. Results achieved: blue line – proposed method, green line – BoRMaN method

Table 4

The accuracy (Acc.) of the proposed algorithm, our previous method proposed in [1] and BoRMaN[12] prior art solutions. We emphasize the best results for each accuracy criterion.

Method	BoRMaN [8]	Bandrabur et al[1]	*Proposed*
Acc. ε< 0.05	16.28	23.42	***35.71***
Acc. ε< 0.1	89.42	91.42	***98.86***
Acc. ε< 0.25	100	100	***100***

The correct localization on Cohn Kanade database with precision of $\boldsymbol{\varepsilon}$< 0.1 was 98.86% and 35.71% with $\boldsymbol{\varepsilon}$< 0.05. Examples of results are presented in Fig. 11. In addition, we tested on the BioID database, where we obtained a precision of $\boldsymbol{\varepsilon}$< 0.1 was 92.64% and 33.82% with $\boldsymbol{\varepsilon}$< 0.05.

The described solution is completely autonomous and it is implemented in Matlab, where it takes 1s/frame on an Intel i7 processor to localize 14 landmark points, so we assume that in optimal C code implementation will take only 10 milliseconds / frame.

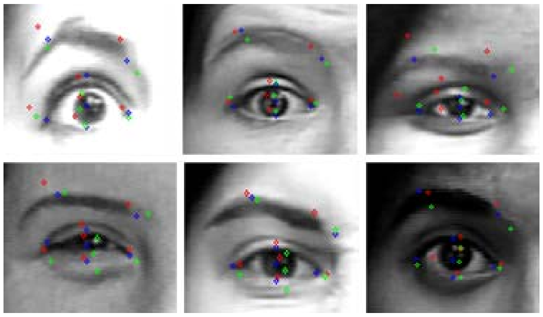

Fig. 11. Cropped eyes. The ground truth are marked with blue, while the detected point with green and The BoRMaN [8] points with red. Top row shows failure cases, while the bottom shows accurate localization.

5. Conclusion

This paper described a fast and efficient eye landmarks detector. The starting points are chosen through anthropometric reasons, unlike [1] where starting points where the results proposed by time consuming method from [7] (the BoRMaN algorithm). The proposed feature descriptor is computed using integral and edge projections. For each starting point, we consider its vicinity and we apply a refinement, in addition to [1]. The classification is done using the multi-layer perceptron. We tested the proposed algorithm on public and widely used databases, the Cohn Kanade and the BioID database showing considerable improvements over the state of the art algorithm [7] (a 19% increase in accuracy) and our previous method [1] (a 12% increase in accuracy).

While overall the method performs reasonably well, poorer results are encountered on images with less contrast on the landmarks (i.e. having the hair color closer to blonde) and in extremely shadowed images where the projections variation is perturbed by illumination variation.

Acknowledgment

This work was supported by the Romanian Sectorial Operational Program Human Resources Development 2007-2013 through the European Social Fund Financial Agreements POSDRU /89/1.5/S/132395 (InnoREASEACH) and POSDRU /89/1.5/S/132398 (Knowledge).

REFERENCES

[1] *P. Majaranta, A. Bulling* “Eye Tracking and Eye-Based Human–Computer Interaction” Advances in Physiological Computing Human–Computer Interaction, Series 2014, pp 39-65.

[2] *G. Ghinea, C. Djeraba, S. R. Gulliver, K. P. Coyne.* "Introduction to special issue on eye-tracking applications in multimedia systems".ACM Transactions on Multimedia Computing, Communications, and Applications,**vol. 3**, no.4, 2007

[3] *M.W. Schurgin, J.Nelson, S. Iida, H. Ohira,J.Y. Chiao, S.L. Franconeri*"Eye movements during emotion recognition in faces",Journal Vision, **vol.14**, no. 13-14. 2014.

[4] *J. Cohn and K. Schmidt.* "The timing of facial motion in posed and spontaneous smiles". Int. Journal of Wavelets, Multiresolution and Information Processing , 2:1–12, March 2004.

[5] *M. F. Valstar and M. Pantic.* "How to distinguish posed from spontaneous smiles using geometric features". In Proc. ICMI ,pp. 38–45, 2007.

[6] *M. Valstar, B. Martinez, X. Binefa, M. Pantic.* "Facial point detection using boosted regression and graph models". In Proc. of CVPR, pp. 2729- 2736, 2010.

[7] *H.Zhoua,H. Hub,*"Human motion tracking for rehabilitation - A survey", Biomedical Signal Processing and Control,**Vol. 3**, no. 1, 2008, pp. 1–18.

[8] *O. Celiktutan, S. Ulukaya, B. Sankur,* "A Comparative Study of Face Land marking Techniques", in *EURASIP Journal on Image and Video Processing*, **vol. 13**, 2013, doi:10.1186/1687-5281-2013-13.

[9] *T. F. Cootes, G. J. Edwards, C. J. Taylor*,"Active appearance models". IEEE Trans. Patt. Anal.Mach. Intel., **vol. 23,** no. 6, pp. 681 – 685, 2001.

[10] *T. Leung, M. Burl, P. Perona,"* Finding faces in cluttered scenes using random labeled graph matching".In Proc. of ICCV, pp. 637 – 644, 1995.

[11] *D. Cristinacce T. Cootes*. "Feature detection and tracking with constrained local models". In Proc. of BMVC, pp. 929 – 938, 2006.

[12] *L. Florea, C. Florea, R.Vranceanu, C.Vertan* "Zero-crossing Based Image Projections Encoding for Eye Localization", In Proc. of EUSIPCO 2012, pp. 150-154, 27-31 august 2012, Bucuresti, Romania.

[13] *A. Bandrabur, L. Florea, R. Boia, C. Florea* "Compressed Projections for Localization of Landmarks in the Eye Region", In Proc. of COMM2014, pp.1-4, Bucuresti, Romania, 2014

[14] *T. Kanade,* "Picture processing by computer complex and recognition of human faces", Technical Report, Kyoto University, Department of Information Science, 1973

[15] *P. Viola, M. Jones*. "Robust real-time face detection".International journal of Computer Vision, **vol. 57,** no. 2, pp. 137–154, 2004.

[16] *R. Valenti,* "What are you looking at? Automatic estimation and inference of gaze".PhD Thesis, University of Amsterdam, chap. 6, pp. 118, 2011.

[17] *BioID database* https://www.bioid.com/About/BioID-Face-Database

[18] *T. Kanade, J. F. Cohn, and Y. Tian,* "Comprehensivedatabase for facial expression analysis*,"* InProc. of the Fourth IEEE International Conference on AutomaticFace and Gesture Recognition (FG'00), 2000, pp.46–53.

[19] *G. Lipori*, Manual annotations of facial fiducial points on the Cohn Kanade database, LAIV laboratory, University of Milan, web url: http://lipori.dsi.unimi.it/download.html.

[20] *P. Ekman, W. Friesen,* „Facial Action Coding System: A technique for the measurement of facial movement. Palo Alto", U.S.: CA: Consulting Psychologists Press, 1978.

[21] *O. Jesorsky, K. Kirchberg, R. Frischolz*, "Robust face detection using the Hausdorff distance," in Audio and Video Based Person Authentication, Eds. J. Bigun and F. Smeraldi, 2000, pp. 90–95,

[22] *G. C. Feng, P. C. Yuen,* "Variance projection function and its application to eye detection for human face recognition", Pattern Recognition Letters, **vol. 19(**9), pp. 899 – 906, 1998.