FACE POSE RECOGNITION BASED ON MONOCULAR DIGITAL IMAGERY AND STEREO-BASED ESTIMATION OF ITS PRECISION

V. Gorbatsevich [a, *], Yu. Vizilter [a], V. Knyaz [a] and S. Zheltov [a]

[a] St. Res. Institute of Aviation Systems (GosNIIAS), 125319, 7, Victorenko str., Moscow, Russia - gvs@gosniias.ru

Commission V, WG V/5

KEY WORDS: Biometrics, face 3D reconstruction, facial imagery modelling, facial imagery analysis

ABSTRACT:

A technique for automated face detection and its pose estimation using single image is developed. The algorithm includes: face detection, facial features localization, face/background segmentation, face pose estimation, image transformation to frontal view. Automatic face/background segmentation is performed by original graph-cut technique based on detected feature points. The precision of face orientation estimation based on monocular digital imagery is addressed. The approach for precision estimation is developed based on comparison of synthesized facial 2D images and scanned face 3D model. The software for modelling and measurement is developed. The special system for non-contact measurements is created. Required set of 3D real face models and colour facial textures is obtained using this system. The precision estimation results demonstrate the precision of face pose estimation enough for further successful face recognition.

INTRODUCTION

A problem of object orientation determination based on single image often arises in image analysis field. An estimation of human face orientation is important for different applications, especially, for biometrical applications. Estimation of face pose is very important for automated person identification systems because it allows increasing the probability of correct identification.

There are two main approaches to face pose account in face recognition algorithms:

1) Account of face pose at the stage of biometrical template (feature vector) calculation (Jie et al, 2010, Zhang et al, 2013).
2) 3D-model-based face pose estimation and artificial frontal view generation before the biometrical template calculation (Choi et al, 2010, Kemelmacher-Shlizerman and Basri, 2010, Kemelmacher-Shlizerman et al, 2013).

The second approach allows using the existing 2D frontal face recognition algorithms for processing of non-frontal face images. So, it is very popular and useful approach. And the problem of evaluation of face pose is addressed in some special researches, for example, (Vatahska et al, 2007).

In this paper a technique for automated face detection and pose estimation using single image is developed and reported. The developed algorithm contains following steps:

- Face detection
- Facial points and features localization
- Face/background image segmentation
- Face pose estimation
- Face pose correction (image transformation to a frontal position).

The photogrammetric approach for precision of face pose estimation is developed too based on comparison of synthesized facial 2D images and scanned face 3D model. This approach allows separating the influence of pose correction errors from the influence of other factors to the final results of biometric face recognition.

1. FACE DETECTION AND SEGMENTATION

1.1 Face detection and tracking

Recent years the Viola-Jones (Viola and Jones, 2001) face detection algorithm and its modifications are considered as state-of-the-art. This approach presumes the design of multi-stage strong classifier based on the weak classifiers at each stage. The selection of weak classifiers is performed using the boosting algorithm. The initial set of weak classifiers is usually formed as a some set of Haar-like features: differences of mean intensity values of some local rectangular regions. The main advantage of these features is a high speed of their computation, which allows to provide the real-time calculation. The further increasing of computational speed is obtained using the multi-stage scheme. Each hypothesis about the face position in image is tested by a sequence of strong classifiers (stages). The negative response of any stage directly forces the final negative answer. The first stages use the lower number of features than the last stages. So, typical numbers of features by stages are: 1, 5, 10, 10, 20. However, this approach meats some problems in practice due to high variability of face images depending on face pose and some other factors. Because of this, the tree of classifiers is usually designed as a set of staged classifiers learned for different face poses with common first stages. Finally, such classifiers are applied in a set of sliding windows of different scales for detecting faces with different size and pose in image.

* Corresponding author.

Unfortunately, the performance of such algorithm is not enough for the real-time search of multiple small faces in a high-resolution video. For this case the special detection-and-tracking algorithm was developed. This algorithm uses the full detection algorithm in some rare basic frames of video sequence, but for other frames it uses just the simplified "tracking-detection" algorithm for tracking of previously detected faces. Such tracking-detection utilizes the information about the size and position of detected faces that decreases the search area and number of pyramid levels (scale hypotheses).

The detection-and-tracking algorithm includes following steps:

1. Creation of tracking-detectors for a set of predefined scales.
2. Processing of basic frame:
    a. The full search of faces for some basic frame of video sequence.
    b. Initialization of tracking process: the each detected face is assigned to one tracking detector with close face scale.
3. Processing of next *n* frames. For each face detected on a basic frame:
    a. Determination of bounds of the search zone.
    b. Testing of face position hypotheses in a search zone using the corresponding tracking-detector assigned at step 2.b.
4. If video sequence contains the next frame, then go back to Step 2.

Testing of this detection-and-tracking algorithm over a set of real video sequences demonstrated that the performance of face detection essentially increases while the quality practically remains the same. In contrast to the correlation-based face tracking algorithms, this approach provides the stability relative to lighting conditions and allows to obtain the exact face size at each frame.

1.2 Detection of facial feature points

The next stage of 3D face model construction is a detection of facial feature points (eyes, nose, mouth corners, etc.). In this paper we apply the Viola-Jones (Viola and Jones, 2001) algorithm too trained for detection of most stable facial feature points (Figure 1). The feature selection and training were performed over a large base of facial images with manually marked ground truth features.

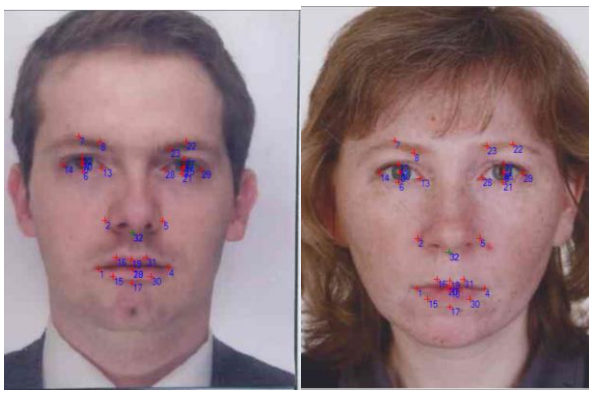

Figure 1. Detection of most stable facial feature points

1.3 Face region segmentation

The next step of face image processing is a segmentation of face region or simply face segmentation. It presumes the detection of exact face borderline. The correct face segmentation provides the possibility for determination of face type, face width, face borderline and coordinates of peripheral feature points required for facial 3D model adjustment.

Currently the graph-cut methods show the best results among automated techniques for image segmentation onto "object" and "background". But they could not be applied without some preliminary image marking (usually manual) (Boykov et al, 2001, Rother et al, 2004). So, such technique is not applicable to automatic face segmentation.

In this paper we propose to use the automatically detected facial features as a preliminary marking for graph-cut face segmentation. Figure 2 demonstrates the example of face image with automatic pre-marking of skin (face) part (green triangle) and background part (red border).

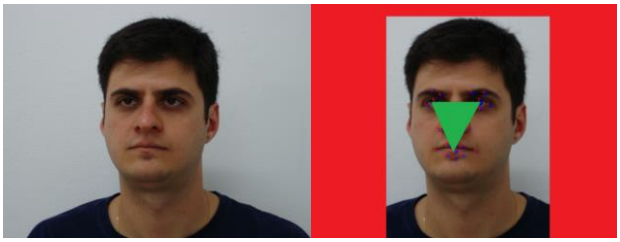

Figure 2. Automatic face pre-marking based on facial feature point detection

Such pre-marking allows applying of graph-cut technique (Boykov et al, 2001) for face segmentation. The neighborhood graph *G* is formed using the 8-neighbor aperture. The weights of graph edges are assigned corresponding to color difference formula of CIEDE2000 (Sharma et al, 2005):

$$D_{ij} = e^{\left(\frac{-\Delta E_{00}(\mathrm{Im}(x_i,y_i),\mathrm{Im}(x_j,y_j))}{\sigma}\frac{1}{D((x_i,y_i),(x_j,y_j))}\right)}, \qquad (1)$$

where

D_{ij} - weight of edge between *i*-th and *j*-th image pixels;

ΔE_{00} – color difference between pixels calculated according to the CIEDE2000;

Im(*x*,*y*) – color and intensity values of pixel with (*x*,*y*) coordinates;

(x_i, y_i) - coordinates of *i*-th image pixel;

σ – algorithm tuning parameter;

$D((x_i,y_i),(x_j,y_j))$ – Euclidean distance between pixel positions.

Then we can use following energy function for segmentation:

$$E(L) = \sum_i D_i(L_i) + \sum_{(i,j)\in G} D_{ij} \cdot \delta(L_i \neq L_j) \qquad (2)$$

where

L_i - label of *i*-th pixel: 1 for "face" and 0 for "background";

$$\delta(L_i \neq L_j) = \begin{matrix} 1: L_i \neq L_j \\ 0: L_i = L_j \end{matrix}$$

$D_i(L_i)$ - "data" term of *i*-th pixel: 0 if label L_i is equal to pre-marking label (or pixel is not pre-marked); some penalty constant elsewhere;

D_{ij} - weights of pairwise energy term described by (1).

This energy is sub-modular, so, the graph-cut minimization technique provides the global minimum of (2) (Kolmogorov and Zabih, 2004). Resultant optimal labelling produces the required face-to-background segmentation.

Thus, the following face segmentation algorithm is implemented:

1. Searching of basic facial feature points (for example, centers of eyes and mouth).
2. Preliminary object and background marking based on in-face and out-of-face points (like in Figure 2).
3. Graph-cut image segmentation to object (face) and background (Boykov et al, 2001) using the energy function (2).

This algorithm combines the good features of energy-based graph-cut segmentation and possibility for automatic image processing. The original expression for weights (1) allows to operate in different and variable lighting conditions. Figure 3 demonstrates the example of automatic face region segmentation: red region is set of pixels classified as a background; grayscale region is a facial image inside the detected facial borderline.

Figure 3. Example of automatic face region segmentation: red region is set of pixels classified as a background

2. FACE POSE ESTIMATION AND CORRECTION

The problem of head orientation estimation based on single image is an ill-posed problem. So for estimation of face rotation angles in the image some indirect methods are usually applied such as face proportion analysis or deformable head 3D model matching (Williamson, 2011).

In this paper we use the algorithm for 3D face pose estimation and correction applying the deformable head 3D model (Figure 4). This algorithm contains following steps:

1. Searching of basic facial feature points.
2. Preliminary positioning of 3D model based on feature points.
3. Deformation of 3D model based on precise feature points matching with 3D model points.
4. Repainting of invisible face regions.
5. Face pose correction via rotation of 3D model to the "frontal" pose and generation of face frontal view.

The positioning of 3D model by feature points is performed via solution of collinear equations for pairs of corresponding points in image and 3D model:

$$x = x_0 - f\frac{a_{11}(X - X_S) + a_{12}(Y - Y_S) + a_{13}(Z - Z_S)}{a_{31}(X - X_S) + a_{32}(Y - Y_S) + a_{33}(Z - Z_S)}$$

$$y = y_0 - f\frac{a_{21}(X - X_S) + a_{22}(Y - Y_S) + a_{23}(Z - Z_S)}{a_{31}(X - X_S) + a_{32}(Y - Y_S) + a_{33}(Z - Z_S)}$$

where

x, y – coordinates of point projection in a picture coordinate system;

x_0, y_0 – coordinates of the main point in a picture coordinate system; f – focal distance; $a_{i,j}$ – elements of transform matrix; X, Y, Z – coordinates of point in an object coordinate system; X_S, Y_S, Z_S – coordinates of the center of projection S in an object coordinate system.

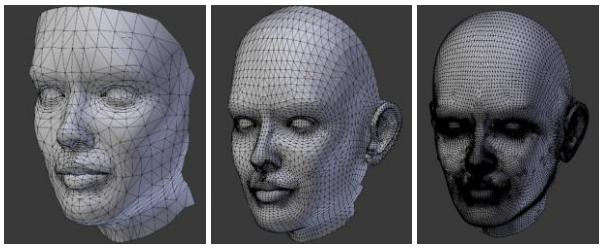

Figure 4. Deformable head 3D model with "low" (~900 triangles), "medium" (~5000 triangles) and "high" (~20000 triangles) spatial resolution.

The solution of this system of non-linear equations is performed by the modified Newton technique. After the preliminary head positioning, the deformation of the head model is performed using the precise matching of 2D image and 3D model points. This procedure takes in account the special properties of the human skull geometry. Figure 5 demonstrates the set of feature points applied for 3D head model deformation.

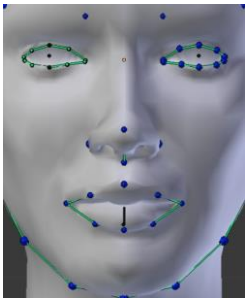

Figure 5. Set of feature points applied for 3D head model deformation

The next step of algorithm is a repainting (restoring) of invisible face regions. Such regions occur due to the occlusion of some face parts by other face parts (nose and so on) in different views far from the frontal one. The restoration of such invisible regions of facial texture is based on the hypothesis of face symmetry relative to vertical (nose) plane.

Let's consider a set of invisible triangle vertices ***INV*** and the corresponding border set of vertices Γ_{INV}. Let it be a set of symmetrical vertices ***SYM*** with a border set Γ_{SYM}.

The intensity of vertices from ***SYM*** can be expressed relative to intensities of vertices from Γ_{SYM} as follows:

$$K_{i,j} = \mathrm{SYM}_i / \Gamma_{\mathrm{SYM}_j},$$

where $K_{i,j}$ – ratio of intensity of i–th vertex of SYM to the intensity of j-th vertex of Γ_{SYM};
SYM_i – intensity of i–th vertex of SYM;
Γ_{SYM_j} – intensity of j-th vertex of Γ_{SYM}.

Then the intensities of invisible vertices are restored by the following formula:

$$\mathrm{INV}_i = K_{\mathrm{norm}i} \sum_j \Gamma_{\mathrm{INV}_j} K_{i,j}$$

where
INV_i – intensity of i–th vertex of ***INV***;

$K_{\mathrm{норм}\,i} = \dfrac{1}{\sum_j K_{i,j}}$ - normalization coefficient;

Γ_{INV_j} – intensity of i–th vertex of Γ_{INV}.

Figure 6 demonstrates the example of face pose correction using the algorithm described. The result of this procedure is a generation of face frontal view.

Figure 6. Example of face pose correction

3. TECHNIQUE FOR EVALUATION OF FACE POSE ESTIMATION

For estimation of precision of pose estimation the following technique was applied. Accurate face 3D model of a face was acquired by photogrammetric system, which then was used for generating synthetic photographs of this face with given orientation (yaw/roll/pitch angles, see Figure 7). Then these photographs were processed by 2D pose estimation procedure and results of estimation were compared with real angle values.

A face 3D model is acquired by calibrated photogrammetric system which provides high accuracy and high density of spatial coordinates and accurate colour texture mapping (Knyaz, 2010). The 3D model is generated in exterior coordinate system defined by calibration field. After 3D model generation it is transform to face coordinate system using given anthropometric points.

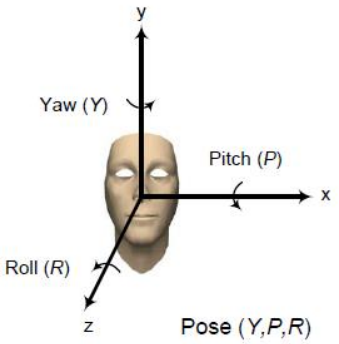

Figure 7. Face system of coordinates

Further this 3D model is used for generation of synthetic person photographs with given known face orientations. These photographs are processed by various indirect methods of head pose estimation for accuracy characteristics calculation. Figure 8 presents 3D models of a mannequin head and a real person used for 2D pose estimation procedure.

Figure 8. Textured 3D models of human and mannequin faces in frontal and non-frontal poses

For accurate face 3D model acquisition a photogrammetric system was used.

For using in biomedical application a photogrammetric system has to meet a set of requirements such as safety and convenience for a person, short time of 3D model acquisition caused by possibility of a person movement. Also the system has to produce accurate textured 3D model of a high resolution for providing a possibility for accurate anthropometric measurements.

To meet these requirements the photogrammetric system used for 3D reconstruction includes BASLER A601f IEEE-1394 camera and InFocus IN10 multimedia projector. These hardware configuration supports operating in synchronic mode.

The photograph of the developed photogrammetric system is presented in Figure 9.

Figure 9. Photogrammetric system for Face 3D model acquisition

BASLER A601f camera has 696x491 pixel resolution and supports external synchronization, program control for shutter, gain, and acquisition up to 60 frames per second.

InFocus IN10 multimedia projector has 1024x768 pixel resolution. It can work at frequencies 60-85 Hz. Special cable provides camera image capturing in synchronic mode with the projector.

The developed digital projector-single camera photogrammetric system uses personal computer as processing unit and original software for calibration and 3D reconstruction.

The additional parameters describing CCD camera (and a projector) model in co-linearity conditions are taken in form:

$$\Delta x = a\bar{y} + \bar{x}r^2K_1 + \bar{x}r^4K_2 + \bar{x}r^6K_3 + (r^2 + 2\bar{x}^2)P_1 + 2\bar{x}\bar{y}P_2$$

$$\Delta y = a\bar{x} + \bar{y}r^2K_1 + \bar{y}r^4K_2 + \bar{y}r^6K_3 + 2\bar{x}\bar{y}P_1 + (r^2 + 2\bar{y}^2)P_2$$

$$\bar{x} = m_x(x - x_p);\ \bar{y} = -m_y(y - y_p);\ r = \sqrt{\bar{x}^2 + \bar{y}^2}$$

where x_p, y_p,-the coordinates of principal point,
m_x, m_y - scales in x and y directions,
a – affinity factor,
K_1,K_2,K_3 – the coefficients of radial symmetric distortion
P_1,P_2 - the coefficients of decentring distortion

The common procedure for determining unknown parameters of camera model is bundle adjustment procedure using observations of test field reference points with known spatial coordinates.

Image interior orientation and image exterior orientation (X_i, Y_i, Z_i – location and α_i,ω_i,κ_i and angle position in given coordinate system) are determined as a result of calibration. The residuals of co-linearity conditions for the reference points after least mean square estimation σ_x, σ_y are concerned as precision criterion for calibration.

The results of system calibration provide the values of σ_x, σ_y at the level of 0.05 mm.

4. EXPERIMENTAL RESULTS

4.1 Results of photogrammetric experiments

Two 3D models were used for the accuracy of 2D pose estimation: a mannequin head and a real person head. For each 3D model a set of images with given yaw/roll/ pitch angle orientation was produced. The maximum/minimum angles were ±20° and an angle step was 1°.

The errors of 2D pose estimation for different angles (yaw/roll/ pitch) are presented in Figure 10.

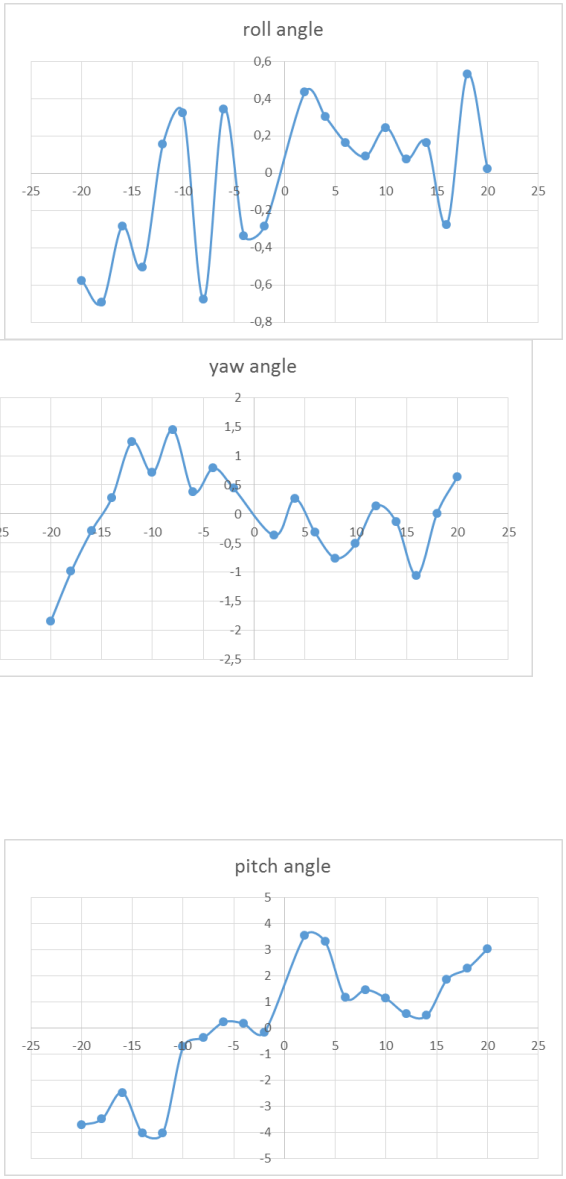

Figure 10. Experimental results: precision of pose estimation

The experimental estimates of head pose determination accuracy are following:

- roll angle – maximum error 0,69°; mean absolute error – 0.32°; mean squared error – 0.37°;

- yaw angle – maximum error 1.84°; mean absolute error – 0.63°; mean squared error – 0.78°;

- pitch angle – maximum error 4.05°; mean absolute error – 1.91°; mean squared error – 2.36°.

Relatively low accuracy in pitch angle estimation is caused by high variability of face vertical proportion.

Let's compare these results with competitive approach. In (Kolmogorov and Zabih, 2004) the analogous face orientation estimation algorithm was proposed, and the evaluation of its accuracy was performed too. The absolute mean errors of face orientation estimation were the following: Roll - 1.65°, Pitch - 2.74° and Yaw - 3.13° (Vatahska et al, 2007). So, the results of photogrammetric evaluation of proposed pose estimation technique are essentially better. This gives us the hope for better applicability of developed pose correction technique in biometrical applications.

4.2 Results of biometric experiments

The described approach is tested on the public FEI Face Database (FEI, 2012) containing 2800 images of 200 persons captured in different poses (views) relative to camera. Face recognition is performed using the face recognition software IdotFace RT SDK v.1.5.0.3 (Institute of Digital and Optical Technologies B. V.,2014).
Characteristics of biometric verification and identification were tested in our experiments. Biometric *verification* (1:1 testing) presumes that the decision about recognition of input face image *A* as a person with sample face image *B* is accepted, if the classifier score λ(A,B) exceeds some threshold *t*. In this case the verification quality is characterized by two basic distributions: $f_{gen}(x)$ in one person tests, and $f_{imp}(x)$ in impostor tests (comparisons with images of other people).

Two types of errors are estimated: False Recognition Rate (FRR) estimates the probability of false decision in person tests, False Acceptance Rate (FAR) estimates the probability of false decision in impostor tests. If *t* is fixed, then these error probabilities take a form:

$$FRR(t) = \int_{-\infty}^{t} f_{gen}(x)dx;$$

$$FAR(t) = \int_{t}^{+\infty} f_{imp}(x)dx$$

If *t* is variable, then 2D point $\langle FAR(t); FRR(t)\rangle$ draws the so called *ROC* or *DET* curve in a plane *FAR*−*FRR*. This curve completely describes the behavior of biometric classifier in a verification mode.

Biometric *identification* problem (1:*N* testing) presumes the comparison of input image *A* with all images (templates) from database $\mathbf{B} = \{B_i\}_{i = 1...N}$. The identification decision taking is performed using the nearest neighbor rule:

$$c_\lambda(A; \mathrm{B}) = \arg\max_{i=1...N} \lambda(A, B_i)$$

The other formulation of identification problem (1:*n*:*N* testing) presumes «finding of *n* most similar candidates» or «creation of best *n* candidates list». This problem is addressed as *n-identification*. Its solution means finding of nearest *n* neighbors for *A* in base B, $n \ll N$. Quality of identification procedures is estimated by probability of person matching in first *n* candidates in descending order of λ. This probability depends on the size of base *N*. So, matching in *n/N* is often referred as matching with «first *x*% of base». So, cumulative match characteristic (CMC) describes the behavior of biometric classifier in identification mode.

Figure 11 demonstrates the results of face recognition (verification and identification) for non-frontal views (~30 degrees from the frontal view). These graphs (FAR-FRR and CMC correspondingly) demonstrate the essential increasing of 2D face recognition characteristics due to the usage of proposed face pose estimation and correction algorithm.

The total computational performance of proposed pose estimation and correction procedure in our experiments was about 300 ms under one tread computations on Intel Core i7-860/8Gb with "medium" resolution of head 3D model (~5000 triangles) and 800x600 resolution of input face image.

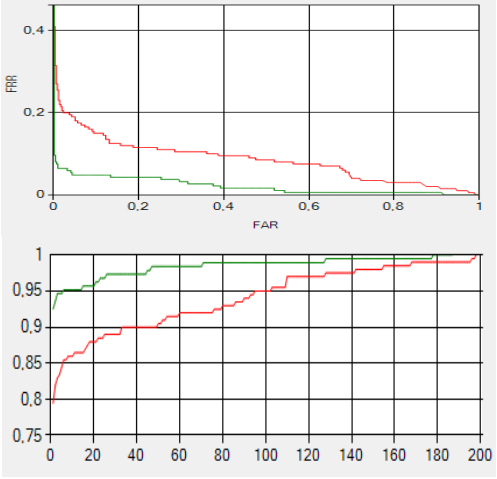

Figure 11. Face recognition results for test base FEI for non-frontal views: FAR/FRR (top) and CMC (bottom) graphs with (green) and without (red) face pose correction

5. CONCLUSIONS

A technique for automated face detection and its pose estimation using single image is presented. The algorithm included face detection, facial features localization, face/background segmentation, face pose estimation, image transformation to frontal view is developed and evaluated. The precision of face orientation estimation based on monocular digital imagery is estimated using photogrammetric technique. The precision estimation results demonstrate high precision of face pose estimation (mean squared error – 0.37°, 0.78°, 2.36° for roll/yaw/ pitch angles respectively) being quite enough for further successful face recognition.

The results of biometrical evaluation show that developed algorithm provides the essential increasing of 2D face recognition characteristics due to the usage of proposed face pose estimation and correction algorithm.

REFERENCES

Boykov, Y., Veksler, O. and Zabih, R., 2001. Fast Approximate Energy Minimization via Graph Cuts // IEEE Transactions On Pattern Analysis And Machine Intelligence, 2001

Choi, J., Medioni, G., Lin, Y., Silva, L., Bellon, O., Pamplona, M., Faltemier, T. C., 2010. 3D Face Reconstruction Using A

Single or Multiple Views. International Conference on Pattern Recognition, 2010

FEI Face Database, 2012. http://fei.edu.br/~cet/facedatabase.html

Jie, F., Zhihua, H., Hong-Jiang, Z., Tsuhan, C. Z., 2010. Pose Invariant Face Recognition. //Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2010

Kemelmacher-Shlizerman, I., Basri, R., 2010. 3D Face Reconstruction from a Single Image using a Single Reference Face Shape.//IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2010

Kemelmacher-Shlizerman, I., Basri, R., Nadler, B., 2013. 3D Face Reconstruction from Single Two-Tone and Color Images //In Shape Perception in Human and Computer Vision, Springer London, 2013

Knyaz, V., 2010. Multi-media Projector – Single Camera Photogrammetric System For Fast 3D Reconstruction. Proceedings of the ISPRS Commission V Mid-Term Symposium 'Close Range Image Measurement Techniques', ISSN 1682-1777, Vol XXXVIII, Part 5. pp. 343-348, 2010

Kolmogorov, V., Zabih, R., 2004. What energy functions can be minimized via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004

Rother, C., Kolmogorov, V., Blake, A., 2004. Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (SIGGRAPH'04), 2004

Sharma, G., Wu, W., Dalal, E. N., 2005. The CIEDE 2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations // Color Research and Application, vol. 30. No. 1, February 2005

Vatahska, T., Bennewitz, M., Behnke, S., 2007. Feature-based Head Pose Estimation from Images. // Humanoid Robots, 7th IEEE-RAS International Conference, 2007

Viola, P., Jones, M., 2001. Robust Real Time Object Detection. IEEE ICCV Workshop Statistical and Computational Theories of Vision, July 2001

Williamson, J., 2011. http://cgcookie.com/blender/cgc-courses/learning-mesh-topology-collection/ by

Zhang, Y., Shao, M., Wong, E. K., Fu, Y., 2013. Random Faces Guided Sparse Many-to-One Encoder for Pose-Invariant Face Recognition. //ICCV 2013 - Sydney, Australia, 2013

Institute of Digital and Optical Technologies B. V. IdotFace RT SDK v.1.5.0.3 // http://idotbio.com/?page_id=19