Face Detection Algorithm Based on a Cascade of Ensembles of Decision Trees

Anton Lebedev, Vladimir Pavlov, Vladimir Khryashchev, Olga Stepanova
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
lebedevdes@gmail.com, i@yajon.ru, {vhr, olga1stepanova}@yandex.ru

***Abstract*—Face detection algorithm based on a cascade of ensembles of decision trees (CEDT) is presented. The new approach allows detecting faces other than the front position through the use of multiple classifiers. Each classifier is trained for a specific range of angles of the rotation head. The results showed a high rate of productivity for CEDT on images with standard size. The algorithm increases the area under the ROC-curve of 13% compared to a standard Viola-Jones face detection algorithm. To test the applicability of the algorithm in the real world have been conducted research on a robustness. Robustness research shown that the algorithm based on the CEDT show that Gaussian noise, impulsive "salt-and-pepper" noise exert a strong influence on the algorithm (in the worst case decrease in the area under the ROC-curve of 21.2% with a decrease in PSNR metric to 17.99 dB). At the same time blurring, JPEG-compression and JPEG2000 algorithms distortion have little effect on the proposed face detection algorithm (reduction of the area under the ROC-curve by 3.5% while reducing PSNR metric to 21.58 dB).**

I. Introduction

Face detection is an attractive field for computer vision research [1-3]. The face detection task is global because it is used in commercial and law enforcement applications [4-7]. The task of face detection on real images was created in real conditions, the so-called "faces in-the-wild", is relevant at the moment, despite significant progress in the development of such algorithms [11-17].

The Viola-Jones algorithm is the classical face detection approach [3, 8-10]. Viola and Jones proposed to use the signs based on Haar wavelets. They has introduced the two kinds of two rectangular, two kinds of three rectangular view and one four rectangular signs. The value of two rectangular features is the difference between the sum of the intensities of the pixels in a dark box and the sum of the intensities of pixels in a light box. The three rectangular sign sum of the intensities of pixels considered for two bright rectangles. Even for a small 3x3 pixel image, the number of features is essential (12 double rectangular features three 6-square and 4 four-square, for a total of 22 sign). For an image size of the 4x4 number of attributes increases to 136. If we consider the standard size of an image in 24x24 pixel, which is used for the training of face detector in most implementations of the algorithm Viola-Jones, the feature set will consist of 162,336 values. This detector is capable of processing images extremely [3].

Viola and Jones have proposed for their detection cascade structure consisting of units of layers in the form of strong classifiers [3]. This structure allows quick cast a "not face" at the first stage, and the second stage they are calculating a few pairs of rectangular signs. For each stage chose the threshold level so that the relatively high to provide some minimum level of detection at relatively low requirements to the level of a false alarm. Thus, a cascade of rejects at each stage of increasingly sophisticated "not face" passing on all or nearly of the "face".

In this paper, the novel face detection algorithm is based on a cascade of ensembles of decision trees (CEDT). Our approach is a modification of the standard Viola-Jones algorithm with an image-scanning cascade of binary classifiers. If the image's area passes through all the stages of the cascade, it will be classified as an object of interest. Each binary classifier comprises an ensemble of decision trees, which compare the intensity of the pixels in a binary test of their internal nodes. The learning process consists of a procedure for constructing regression tree was based on the greedy algorithm. Most modern algorithms construct regression trees are greedy. The greedy algorithm creates trees from top to bottom by a recursive division of the training data and may be briefly described as follows:

- selection the best separation (providing an extremum of a criterion);
- separation of raw data into subsets;
- recursive application of this procedure for each of the selected subsets.

Greedy algorithms have low complexity, good scalability, but have several disadvantages: a) regression tree is created slowly without returning to previous decisions; b) each step of the algorithm is locally optimal solution. It solution gives the maximum effect on the current step, without regard to impact on the overall solution. Greedy algorithms conduct an optimal separation of data.

To solve the problem that based on regression, we will use the optimized binary decision trees. This approach uses a comparison of pixel intensity as a binary test in its internal nodes. This strategy was proposed by Amit and Geman [18], and later successfully used by researchers and engineers.

A pixel intensity comparison binary test on image I is defined as:

$$bintest(I;l_1,l_2)=\begin{cases}0, & I(l_1)\le I(l_2),\\ 1, & otherwise,\end{cases}$$

where $I(l_i)$ is the pixel intensity at location l_i. l_1 and l_1 are normalized coordinates from the set $[-1;+1]\times[-1;+1]$. It allows resizing binary tests, if necessary. Each terminal node of the tree contains the scalar which models the output value.

Viola and Jones have made object detection feasible in real applications. This is related to the fact that the system based on their algorithm can process the image faster than other approaches with similar results. Mobile devices have limited processing power. Mobile developers are interested in the development of faster detection. Developers are ready to sacrifice precision for the best detection processing speeds for the system to work with limited resources. CEDT algorithm is used to process images and video at high speed. This algorithm maintains the accuracy of the comparison. It allows the re-training algorithm to a new set of data. Also, it is able to classify individuals rotated at different angles relative to the vertical axis. The algorithm is invariant to rotation of the image plane of the screen by using at training multiple copies of the original image rotated by angles uniformly selected from the interval [0;2π) and for small shifts.

II. The face detection algorithm

The face detection algorithm based on CEDT is trained on the following dataset: $\{(I_s, v_s, w_s): s = 1,2,\ldots,S\}$, where v_s is the ground truth for image I_s, w_s is a factor of importance (weight). For example, in the case of binary classification, ground truths have two class labels: positive and negative samples are annotated with +1, -1, respectively. Weights w_s allow ranking these samples according to their importance. The binary test in each node of the tree is chosen in a way to minimize the weighted mean squared error obtained after splitting the input data by the test. The minimization is made according to the following equation:

$$WMSE = \sum_{(I,v,w)\in C_0} w\cdot(v-\overline{v}_0)^2 + \sum_{(I,v,w)\in C_1} w\cdot(v-\overline{v}_1)^2,$$

where C_0 and C_1 are groups of training samples for which the results of the binary test are equal to 0 and 1, respectively. Scalars $\overline{v}_0$ and $\overline{v}_1$ are weighted mean values for ground truths in C_0 and C_1, respectively.

Since the number of comparisons pixel intensity is very large, while optimizing each internal node is created only a small portion of the sample by repeated two coordinates from a uniform distribution on the square $[-1;+1]\times[-1;+1]$ The training data are recursively grouped together so long until the terminating condition is satisfied. The depth of the trees is restricted to minimize the training time, to increase the processing speed and according to memory requirements. The output value for every terminal node is equal to the weighted mean value for ground truth that is obtained in a training process.

If you limit the depth of the tree through D and considered B binary tests in each internal node, as a result the training time will be $O(D\cdot B\cdot C)$ for the training set with S samples. Each training sample is tested with B comparing the intensity of pixels for each internal node, which it passes on the path length D of the root node to the terminal. Construction of a tree requires $O(2^D)$ byte of storage and speed of their work is proportional to $O(D)$.

The single decision tree usually provides the medium accuracy. On the other hand, the ensemble of trees can achieve impressive results. The Gentle-Boost algorithm (the modification of widely used AdaBoost) is used to create the discriminative ensemble fitting the decision tree to an appropriate least squares problem [14].

The following steps are required to generate an ensemble of K trees using training dataset $\{(I_s, c_s): s = 1,2,\ldots,S\}$:

1) Choosing the start weights w_s for each image I_s and its class label $c_s \in \{-1;+1\}$:

$$w_s = \begin{cases} 1/p, & c_s = +1, \\ 1/N, & c_s = -1, \end{cases}$$

where P and N are the total numbers of positive and negative samples, respectively.

2) For $k = 1,2,\ldots,K$:

a) Fit a decision tree T_k by weighted least squares c_s for image I_s with weight w_s

b) Update weights:

$$w_s = w_s \exp\left(-c_s T_k(I_s)\right),$$

where $T_k(I_s)$ is the real-valued output T_k for image I_s.

c) Renormalize weights so that their sum is equal to 1.

3) Output ensemble $\{T_k : k = 1,2,\ldots,K\}$.

During runtime, outputs of all trees in the ensemble are summed and the resulting value is thresholded to obtain the class label. The detection rate is adjusted by varying the ensemble output threshold for every stage of detectors. Each stage uses the soft output ("confidence") of the previous stage as additional information to improve its discriminability. This is achieved by progressively accumulating the outputs of all classification stages in the cascade.

The detector is resistant to small changes in the position and scale around each region of interest may be a few frames. These overlapping detections are combined as a result of post-processing. Two detection combined if the overlap there between is more than 30%:

$$\frac{D_1 \cap D_2}{D_1 \cup D_2} > 0.3.$$

Two datasets are required for the detector training: a dataset with positive samples that contain faces and a dataset with negative samples that do not contain faces. Database GENKI that consists of 3500 annotated faces is used for frontal detector training. In order to improve the algorithm performance, the original images from the database are transformed in different ways [19]. 15 positive training samples with variations in pose

and scale of a face are obtained from every original image after transformation. This makes the detector more robust to noises. 300 000 negative samples are also used for training. The training parameters are set previously. The depth of each tree is fixed at 6 and use 20 classification degrees. Each stage has a predetermined amount of classification trees and the level of detection. Optimization for each internal node of the tree included 256 binary tests. The optimization process significantly improves the performance stage.

New training dataset, which consists of images from color FERET, CMU Multi-PIE [20] databases and video frames from the test area, is collected to detect faces that are rotated left through angles from 300 to 600. This dataset contains 2966 images with annotated frame and four key points (the eye centers, the nose tip, the center of the mouth) marked manually. The negative samples are similar to samples were chosen for training the frontal detector. The following transformations are applied to this dataset: in-plane rotation through angles $\pm 5°, \pm 10°$, shifting the image on $\pm 2.5°, \pm 5°$, scaling $\pm 5°$.

The final detector (CEDT Multi) consists of five trained modules: CEDT frontal, CEDT left 30-60°, CEDT left 60-90°, CEDT right 30-60°, CEDT right 60-90° as shown in Fig. 1.

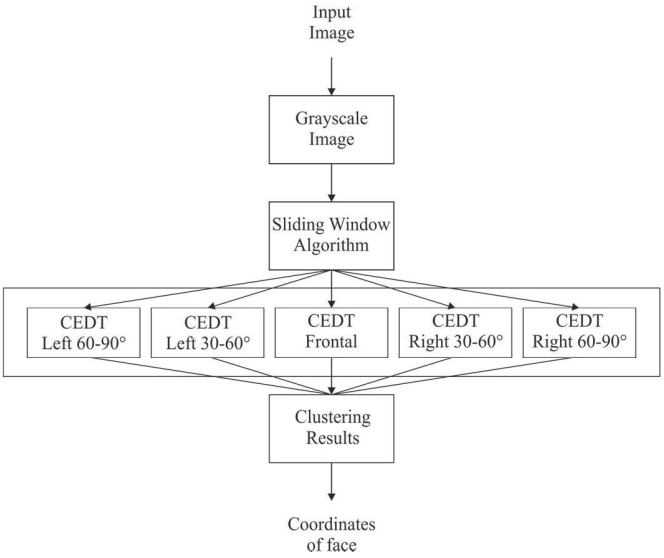

Fig. 1. The final scheme of face detection algorithm using CEDT approach

III. SIMILATION RESULTS

Database Robotics is chosen for testing and analyzing the detector characteristics. This database contains 6623 images of 90 subjects [21]. Each subject has 74 images, where 37 images were taken every 5 degrees from a right profile (defined as +90°) to left profile (defined as -90°) in the pan rotation. Examples of images from the training set are shown in Fig. 2.

ROC-curves for different modules of CEDT detector are presented in Fig. 3a. The areas under ROC-curves are equal to 0.932 (CEDT frontal), 0.856 (CEDT left 30-60°), 0.852 (CEDT right 30-60°), 0.830 (CEDT left 60-90°), 0.852 (CEDT right 60-90°) . In Fig. 3b the areas under ROC-curves are equal to 0.830 (Viola-Jones), 0.932 (CEDT frontal), 0.951 (CEDT multi). Thus, the proposed CEDT algorithm increases the area under ROC-curve by 13% in comparison to Viola-Jones algorithm.

In order to evaluate the robustness of face detection algorithms, the following types of distortion are applied to images: blur, additive white Gaussian noise (AWGN), impulse and multiplicative noises, JPEG and JPEG2000 compression. The results of experiments are presented in Fig. 4.

The linear low-pass averaging filter with mask sizes 20, 30 and 40 pixels simulates blur (Fig. 4a). PSNR=24.33dB for the mask of 20 pixels, PSNR=22.68dB for the mask of 30 pixels, PSNR=21.58dB for the mask of 40 pixels. The areas under ROC-curves are equal to 0.942 (without distortion), 0.903 (mask size is 20 pixels), 0.872 (mask size is 30 pixels), 0.877 (mask size is 40 pixels). Thus, we conclude that blur in test images decreases the area under ROC-curve by 7% for proposed method when PSNR decreases to 21.58dB.

The results of experiments on images distorted by AWGN with different standard deviations (sigma=15, 25, 35) can be seen in Fig. 4b. The medium values of PSNR measure are: 24.94dB, 20.71dB and 17.99dB. The areas under ROC-curves are equal to: 0.942 (without noise), 0.894 (sigma=15), 0.794 (sigma=25), 0.752 (sigma=35). Thus, we conclude that distortion of test images by AWGN has the most significant impact on the CEDT detector, and reduces the area under the ROC-curves by 15% when PSNR is reduced to 20.71dB and by 19% when PSNR is reduced to 17.99dB. The results of similar experiments for impulse noise and multiplicative noise are shown in Fig. 4c and Fig. 4d, respectively.

The results of experiments with Baseline JPEG compression are presented in Fig. 4e. The areas under ROC-curves are equal to: 0.942 (without compression), 0.941 (quality=20, PSNR=35.21dB), 0.930 (quality=10, PSNR=32.17), 0.872 (quality=5, PSNR=24.73). Thus, the low level of compression has a weak impact on detector performance. The increase of compression ratio to PSNR=24.73 dB reduces the area under the ROC-curve by 7%. The results of similar experiments for JPEG2000 compression are shown in Fig. 4f.

Fig. 2. Sample face images from the training dataset

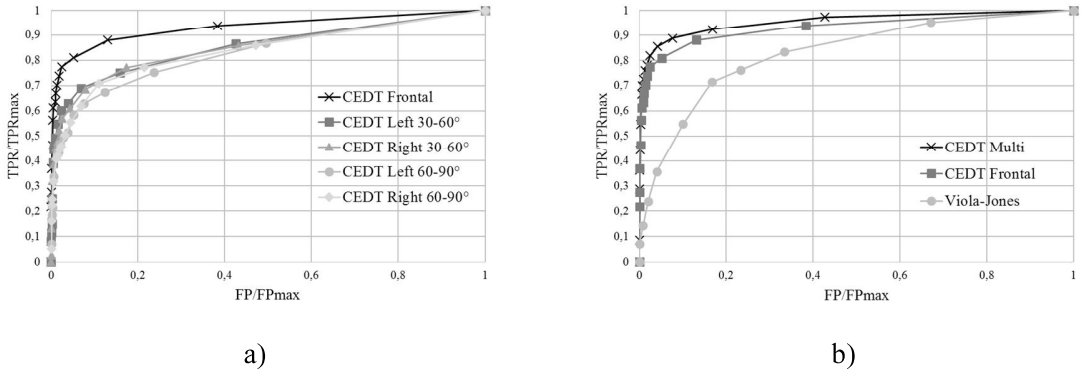

Fig. 3 ROC-curves comparison: a) different modules of CEDT face detector; b) CEDT face detector vs Viola-Jones algorithms

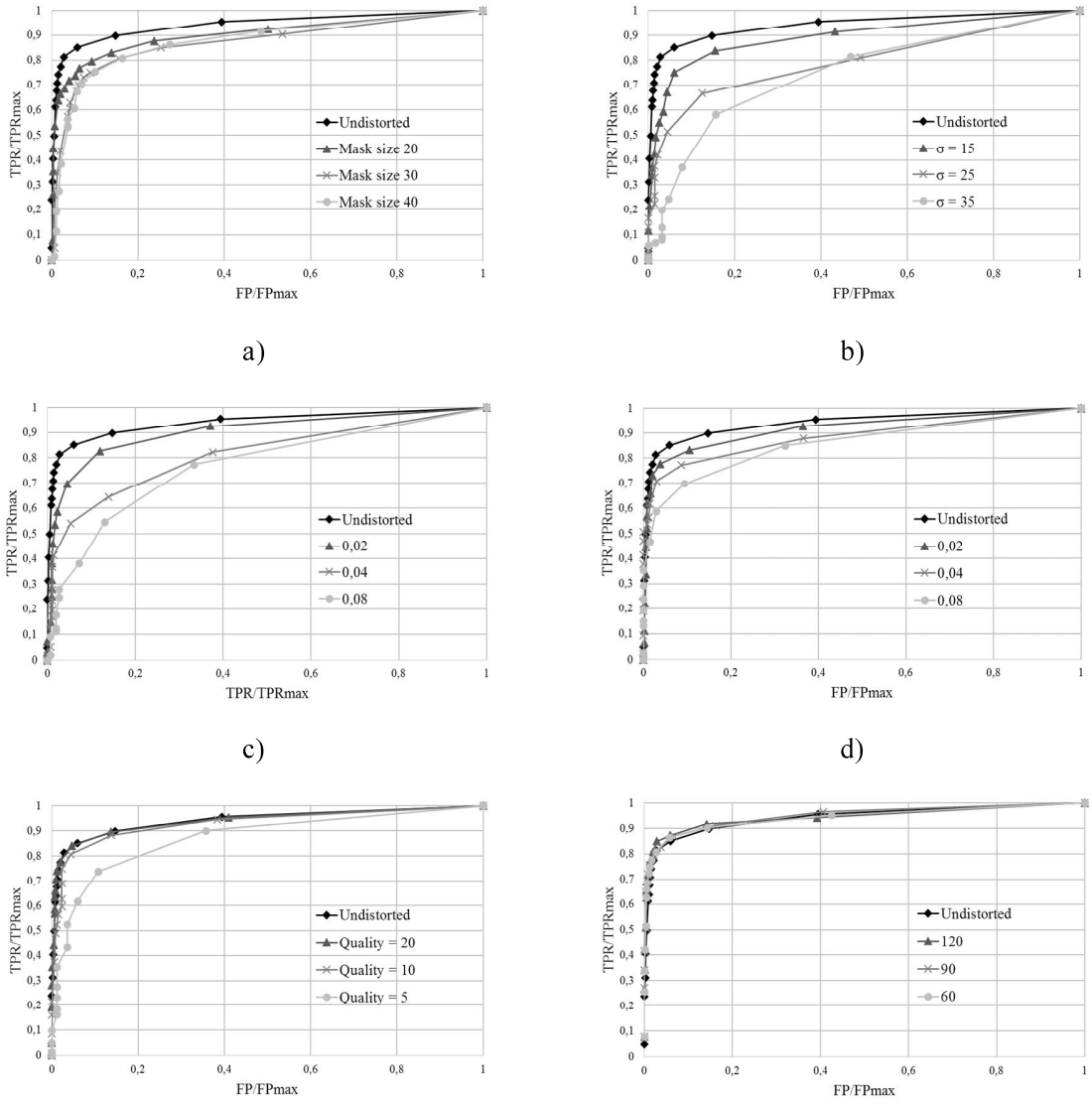

Fig. 4 ROC-curves for images with different types of distortion: a) blur; b) AWGN; c) impulse noise; d) multiplicative noise; e) JPEG compression; f) JPEG2000 compression

The experiment was performed on Python programming language and PC platform with the Intel Core i7-4770 3,40 GHz processor. The average time of CEDT face detection on the 1024×768 pixels image resolution and at the minimum window size of 40×40 pixels is 0.19 seconds. At each iteration the frame size increased produced by 20% of the previous size. We have compared the proposed approach with the Viola-Jones detector from OpenCV library. The average time of the detector is 0.26 seconds under the same settings. The final time of the algorithm is not significantly increased in parallel operation of CEDT detectors. This allows the detection system to use 3-5 detectors for the detection of faces with different orientations relative to the camera.

Visual comparison of face detection quality between Viola-Jones algorithm and CEDT approach is shown in Fig.5. This picture shows the practical improvement of face detector quality which can achieve without increasing the computational complexity.

IV. Conclusions

The proposed algorithm based on CEDT increases the area under ROC-curve by 13% in comparison to standard Viola-Jones detection method.

The experiments on algorithm robustness show that AWGN, multiplicative and impulse noises have a significant impact on algorithm performance (reduction of the area under the ROC-curve: by 21.2% when PSNR decreases to 17.99 dB for AWGN; by 8.8% when PSNR decreases to 15.71 dB for salt-and-pepper impulse noise; by 18.4% when PSNR decreases to 18.75 dB for multiplicative noise).

The experiments also show that blur and JPEG/JPEG2000 compression has a weak impact on the CEDT algorithm: reduction of the area under the ROC-curve: by 3.5% when PSNR decreases to 21.58 dB for blur; by 7.5% when PSNR decreases to 24.73 dB for JPEG; by 0.3% when PSNR decreases to 31.79 dB for JPEG2000).

One more thing which we take from the simulation results is a low computational complexity of CEDT algorithm in comparison with standard Viola-Jones approach. This could prove important in the embedded system and mobile device industries because it can reduce the cost of hardware and make battery life longer.

Acknowledgment

This work was supported by Russian Foundation for Basic Research grants (№ 15-07-08674 and № 15-08-99639).

References

[1] R. Szeliski, *Computer Vision: Algorithms and Applications.* Springer, 2010.

[2] C. Cortes and V. Vapnik, "Support-vector networks", *Machine Learning*, vol. 20, Sept. 1995, pp. 273-297.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *in Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR),* 2001, pp. 511–518.

[4] T. Valentine and J. Davis, *Forensic Facial Identification: theory and practice of identification from eyewitnesses, composites and CCTV.* Chichester: Wiley-Blackwell, 2015.

[5] P.K. Suri and A. Verma, "Robust face detection using circular multi block local binary pattern and integral Haar features", *Int. Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, 2011, pp. 67–71.

[6] L.R. Cerna, G. Camara-Chave and D. Menotti, "Face detection: histogram of oriented gradients and bag of feature method", *in Proc. of the Int. Conf. on Image Processing, Computer Vision & Pattern Recognition (IPCV)*, July 2013, pp. 657-662.

[7] Y. Taigman, Ming Yang, M. Ranzato and L. Wolf, "DeepFace: closing the gap to human-level performance in face verification", *in Computer Vision and Pattern Recognition IEEE Conf. (CVPR)*, June 2014, pp. 1701-1708.

[8] X. Tan and B. Triggs, "Fusing Gabor and LBP feature sets for kernel-based face recognition", *Analysis and Modeling of Faces and Gestures, Lecture Notes in Computer Science*, vol. 4778, 2007, pp. 235-249.

[9] A. Krizhevsky, I. Sutskever and G.E. Hinton, "ImageNet classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems 25*, 2012, pp. 1106-114.

[10] D. Chen, X. Cao, F. Wen and J. Sun, "Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification", *in Computer Vision and Pattern Recognition IEEE Conf. (CVPR)*, 2013, pp. 3025-3035.

[11] E. Zhou, H. Fan, Z. Cao, Y. Jiang and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade", *IEEE Int. Conf. on Computer Vision (ICCV) Workshops*, 2013, pp. 386-391.

[12] J. Li and Y. Zhang, "Learning SURF cascade for fast and accurate object detection", *in Computer Vision and Pattern Recognition IEEE Conf. (CVPR)*, 2013, pp. 3468-3475.

[13] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection", *in Computer Vision and Pattern Recognition IEEE Conf. (CVPR)*, vol. 1, June 2005, pp. 886-893.

[14] T. Riopka and T. Boult, "The eyes have it", *in Proc. of the ACM SIGMM Multimedia Biometrics Methods and Applications Workshop*, 2003, pp. 9-16.

[15] J. Yan, Z. Lei, L. Wen and S. Z. Li, "The Fastest deformable part model for object detection", *in Computer Vision and Pattern Recognition IEEE Conf. (CVPR)*, June 2014, pp. 2497-2504.

[16] P. Felzenszwalb, D. McAllester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model", *in Computer Vision and Pattern Recognition IEEE Conf. (CVPR)*, June 2008, pp. 1-8.

[17] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild", *in Computer Vision and Pattern Recognition IEEE Conf. (CVPR)*, June 2012, pp. 2879-2886.

[18] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees", *Neural Computation*, vol. 9, July 1997, pp.1545–1588.

[19] GENKI Database, MPLab, United Kingdom, Web: http://mplab.ucsd.edu/wordpress/?page_id=398.

[20] Annotated facial landmarks in the wild, Graz University of Technology, Austria, Web: https://lrs.icg.tugraz.at/research /aflw/.

[21] Robotics Database, National Cheng Kung University, Taiwan, Web: http://robotics.csie.ncku.edu.tw/database.htm.

Fig. 5. Visual examples of face detection algorithm quality: a), c), e), g) Viola-Jones algorithm; b), d), f), h) CEDT approach