Deformable Part-based Robust Face Detection under Occlusion by Using Face Decomposition into Face Components

Darijan Marčetić, Slobodan Ribarić
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
{darijan.marcetic, slobodan.ribaric}@fer.hr

Abstract - In this paper, we propose modifications of deformable part-based models in order to increase the robustness of face detection under occlusion. The modifications are: i) the tree, representing the deformable part-based model of the frontal face, which is partitioned into 11 subtrees representing face components; ii) the weight of each face component which is obtained based on the results of psychological experiments; iii) the introduction of new scoring functions and thresholds; and iv) a new procedure for robust face detection based on the valuation of scoring functions and thresholds. The experiment was performed only for frontal face images, and thus this work is used only as a proof of concept. Based on the encouraging experimental results, we conclude that the proposed method is suitable for extension to detect faces with different poses under occlusions.

Keywords: Robust Face Detection, Occlusion, Deformable Part-based Model, Face Components

I. INTRODUCTION

Face detection is the process of finding and localizing faces in an image. Finding and analysing faces in unconstrained real-world images is a very difficult computer vision task. Constraining factors are the ambiguities of the face, such as poses, occlusions, expressions, illumination, scale, colours and textures. Diverse methods have been developed based on scanning window classifiers [1], deep neural networks [2], part-based models [3-6] and regression [7]. Several papers have been published related to face detection under occlusion [8-11], but they are constrained to detect only frontal faces under occlusion. In [8], a support vector machine (SVM) with local kernels was used for robust frontal face detection under partial occlusion. A test was performed on HOIP and CMU face databases. The proposed method was demonstrated for the detection of faces wearing sunglasses or a scarf. It was confirmed that the proposed method was better than a method based on conventional SVM with a global kernel. In [9], a generalization of the work of Viola and Jones for the efficient detection of faces with occlusions was described. The main idea is based on boosting algorithms and cascade structures to efficiently detect faces with occlusions. Several improvements were presented: (i) a robust boosting scheme; (ii) reinforcement training; and (iii) cascading with evidence to extend the system to handle occlusions. Experimental results for detecting frontal faces with artificial occlusions were obtained on CMU and MIT face databases. The authors in [10] divided a face region into multiple patches, and mapped those weak classifiers to the patches. The weak classifiers, inspired by the Viola and Jones approach, belonging to each patch, were combined to create a new classifier to determine if this was a valid face patch without occlusion. All of the valid face patches, with different weights, were used to make a final decision whether the input subwindow was a face. The experimental results showed that the proposed method is promising for the detection of occluded faces. In [11], a novel approach for automatic and robust face detection was presented. It used a component-based approach that combined techniques from both the statistical and structural pattern recognition domain. The method relied on Haar-like features and an AdaBoost trained classifier cascade, and the topology verification was based on graph matching techniques. The system was applied to frontal face detection and the experiments showed improved performance compared to conventional face detection in the presence of partial occlusions, uneven illumination, and out-of-plane rotations, thus yielding higher robustness.

Initially, face detection, pose estimation, and landmark localization were handled separately. These three tasks can be simultaneously handled with a unified part-based approach that uses mixtures of trees with a shared pool of part templates for face landmarks, and utilizing an encoding scheme for modes of elastic deformation, and a view-based topology for face detection and pose estimation [5]. The parameters of part-based models are discriminatively trained in a max-margin framework. Facial landmarks are visible in multiple views and thus the corresponding parts can be shared among different views, which reduces the complexity of the learning procedure for the model. Most part-based models are sensitive to the presence of occlusions because they are learned on non-occluded faces. Facial landmark scores are processed one by one in a cascade, which also decreases robustness.

In this paper, we propose robust frontal face detection obtained by modifying the method presented in [5]. In our approach, a face is decomposed into distinct face components. These face components are selected by psychological experiments related to face recognition by humans [12]. Robustness is achieved by fusion with a

weighted voting scheme where the relative relations among weights are selected based on recommendations given in [12]. In this paper, we present a proof of concept for the detection of frontal faces under occlusions which is suitable for extension to multi-viewpoint occluded face detection.

II. Deformable Part-based Models

Deformable part-based models (DPMs) are one of the most popular face detection, pose estimation and landmark localization methods [5].

The DPM was used to find a visual object in an actual image in [4]. The authors describe the approach of breaking down visual objects into a number of "primitive parts" and specify a permitted range of spatial relations which these primitive parts must satisfy for the object to be present in an image. The object model is represented by primitive parts held together by "springs". The springs joining the primitive parts serve both to constrain relative movement and measure the "cost" of the movement by how much they are "stretched" [3].

The maximal matching between an object model and a visual object in the scene is expressed by the total cost of embedding parts at their locations. For total cost optimization, the authors proposed a method based on dynamic programming.

This initial idea was extended to a mixture of multi-scale deformable part models, represented by HOG features, and methods for discriminative training with partially labelled data [5].

In [6], the authors proposed a method that solved the speed bottleneck of DPMs without impairing the accuracy of detection.

The method proposed in [5] uses a model based on mixtures of trees with a shared pool of parts represented by HOG descriptors. Facial landmarks are modelled as parts. A global mixture of parts is used to capture topological changes due to face viewpoints. A set of 13 trees was used to represent different face viewpoints (Fig. 1). Each tree has only one root node. From 13 trees which represent different face viewpoints, we selected only the tree associated with the frontal face, thus simplifying the model description. This simplification was used for a proof of concept, related to the robust detection of faces under occlusion.

The tree associated with the frontal face $T = (V, E)$ has $|V| = 68$ vertices (or parts) and $|E| = 67$ edges (Fig. 1). Let us write I for an image and $l_i = (x_i, y_i)$ for the pixel location of part V_i. A configuration of parts $L = \{l_i : i = 1,..., 68\}$ is scored as follows [5]:

$$Score(I,L) = App(I,L) + Shape(L) + \alpha \quad (1)$$

$$App(I,L) = \sum_{i=1}^{68} w_i \cdot \phi(I,l_i) \quad (2)$$

$$Shape(L) = \sum_{ij \in E} a_{ij} dx^2 + b_{ij} dx + c_{ij} dy^2 + d_{ij} dy. \quad (3)$$

Equation 2 sums the appearance evidence for placing a HOG-based template w_i for part V_i at location l_i. A HOG-based feature vector $\phi(I,l_i)$ is extracted from pixel location l_i in image I. Equation 3 scores the mixture-specific spatial arrangement of parts L, where $dx = x_i - x_j$ and $dy = y_i - y_j$ are the displacement of the V_i part relative to the V_j part. Each term in the sum can be interpreted as a spring that introduces spatial constraints between a pair of parts, where a, b, c and d represent spring parameters (rest location and rigidity) and were learned on a set of training images [3]. The term α is a scalar bias associated with the front face view.

Inference corresponds to maximizing $Score(I,L)$ in Eq.1 over L:

$$Score^*(I,L) = \max_L [Score(I,L)]. \quad (4)$$

Since $T = (V, E)$ is a tree, the inner maximization can be done efficiently with dynamic programming [3].

Let us denote L for which the $Score$ (I, L) is maximal with L^*, i.e. $Score^*(I,L) = Score(I,L^*)$.

If $Score$ $(I,L^*) \geq \theta_1$, then it is assumed that a face is detected in the image (i.e. locations of facial landmarks are given by L^*). θ_1 is a threshold value applied in [5].

III. Robust Face Detection

In the context of this paper, robustness is related to the ability to detect frontal faces in an image I with different types of occlusions in cases where the original approach [5] fails due to false negative detection (i.e. a face has $Score(I,L^*) < \theta_1$, Fig. 2). In order to increase the robustness of frontal face detection, we modified the original approach described in Section 2 by introducing a new algorithm based on modifications as follows:

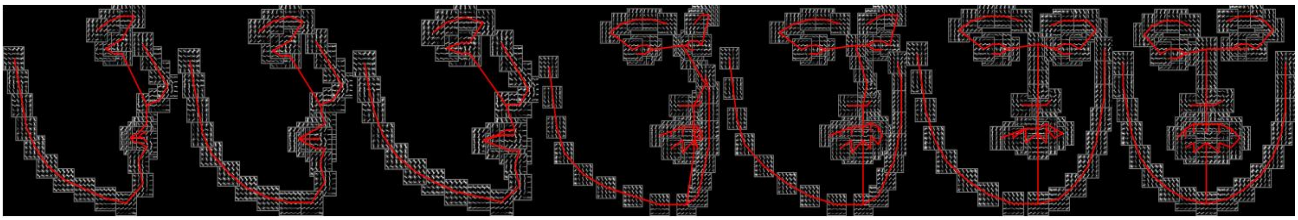

Figure 1. Mixture of trees that encode topological change due to viewpoint [5]. The red lines denote springs between pairs of parts.

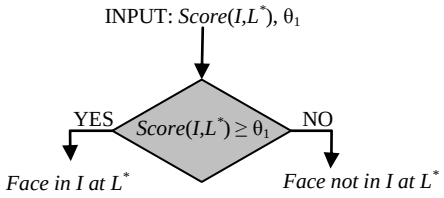

Figure 2. Flowchart of the decision step for the original face detection algorithm [3].

i) The first modification partitions the tree $T = (V, E)$, which corresponds to the frontal face (seventh tree in Fig. 1), into 11 subtrees $T^k = (V^k, E^k)$; $k = 1, 2, ...11$, by removing the edges that connect different face components (Fig. 3). Each subtree is associated with one face component: *left upper edge, left lower edge, right upper edge, right lower edge, left side of the lips, right side of the lips, nose, left eye, left eyebrow, right eye, right eyebrow*.

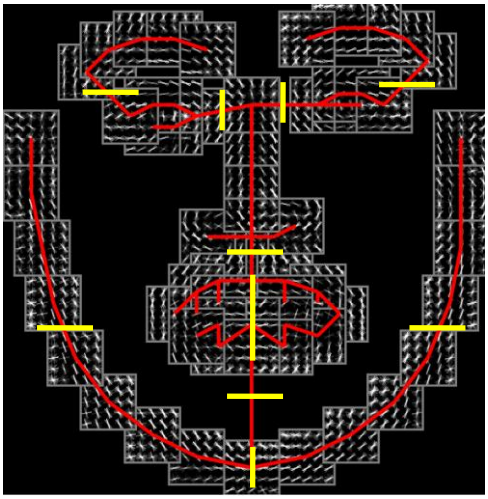

Figure 3. The face region is partitioned into 11 face components containing parts corresponding to the face components. Removed edges are marked with short yellow lines.

ii) The second modification introduces three new thresholds (θ_2, θ_3 and θ_4) and three new scoring functions ($Appearance_k(I,L^*)$, $ScoreAppearance(I,L^*)$ and $ScoreThreshold(I,L^*)$).

The original threshold θ_1 proposed in [5] is -0.65, and the threshold θ_2 is determined based on experiments (described in Section 4). These two thresholds define three intervals for $Score(I,L^*)$. Interpretation of these three intervals is clarified in Fig. 4 b).

The thresholds θ_3 and θ_4 divide the 2-dimensional decision space into face and non-face space (Fig. 4 b)).

Fig. 2 illustrates the original decision step [5]. Fig. 4 a) depicts a flowchart for the final decision step of our frontal face detection algorithm.

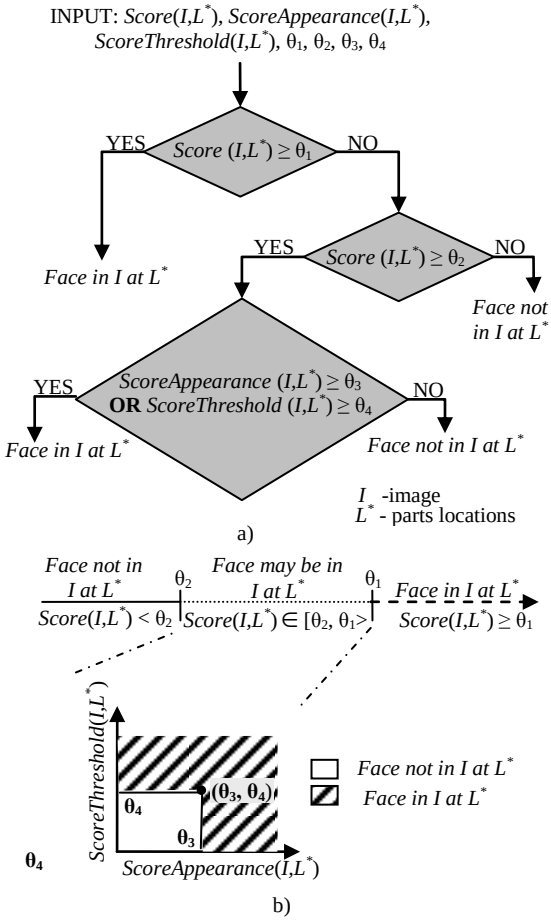

Figure 4. a) Flowchart of the decision step for the proposed face detection algorithm; b) Intervals

A. *Final decision step for the face detection algorithm*

The final decision step for the face detection algorithm is given as follows:

i) If $Score(I,L^*) \geq \theta_1$, then a face is detected in image I at location L^*.

ii) If $Score(I,L^*) < \theta_2$, then a face is not detected in image I at location L^*.

iii) If $Score(I,L^*)$ falls into the interval $[\theta_2, \theta_1>$, then $ScoreAppearance(I,L^*)$ and $ScoreThreshold(I,L^*)$ are used in conjunction with the newly introduced thresholds θ_3 and θ_4, respectively (Fig. 4).

If $ScoreAppearance(I,L^*) \geq \theta_3$ or $ScoreThreshold(I,L^*) \geq \theta_4$, then there is a face in an image I at location L^* (Fig. 4).

The $ScoreAppearance(I,L^*)$ and $ScoreThreshold(I,L^*)$ are described as follows.

A face appearance score is defined as:

$$ScoreAppearance\left(I,L^*\right)=\sum_{k=1}^{11}\mu_k\cdot Appearance_k\left(I,L^*\right),\quad(5)$$

where μ_k is the weight of a k-th face component selected based on the results of psychological experiments [12] which estimate the importance of the face components in the face recognition process performed by humans. The experimental results [12] indicate the following order of importance of facial components (in decreasing order): eyes, eyebrows, mouth, nose and contour of face (Table I).

TABLE I. THE EXPERIMENTALLY DETERMINED IMPORTANCE OF FACE COMPONENTS IN FACE RECOGNITION BY HUMANS

Face component	Weight μ_k
Left/right upper contour - μ_1/μ_2	1
Left/right lower contour - μ_3/μ_4	1
Left/right side of the lips - μ_5/μ_6	2
Nose - μ_7	1.5
Left/right eye - μ_8/μ_{10}	3
Left/right eyebrow - μ_9/μ_{11}	2.5

For each face component k = 1, 2, ...11, the appearance score is defined as:

$$Appearance_k\left(I,L^*\right)=\sum_i w_i\cdot\phi\left(I,l_i\right),\quad(6)$$

where index i determines the parts V_i from the set of parts V^k associated with a k-th face component, w_i is a HOG-based template for part V_i, and $\phi\left(I,l_i\right)$ represents the HOG-based feature vector extracted from pixel location l_i in an image I. For example, k = 8 corresponds to a left eye (Table I), and there are six parts $V^8 = \{V_{21}, V_{22}, ..., V_{26}\}$.

A face appearance threshold score is defined as:

$$ScoreThreshold\left(I,L^*\right)=\sum_{k=1}^{11}c_k\quad(7)$$

where

$$c_k=\begin{cases}1 & \text{if } Appearance_k(I,L^*)\geq Threshold_k\\ 0 & \text{if } Appearance_k(I,L^*)<Threshold_k\end{cases}\quad(8)$$

and $Threshold_k$ defines the minimal value of the appearance score of a k-th face component for which it is supposed that this face component is correctly detected. Thresholds $Threshold_k$; k = 1, 2, …, 11 are selected based on the experiments conducted in the XM2VTS [13] and INRIAP [14] databases described in Section 4. The interpretation of $ScoreThreshold\,(I,L^*)$ is as follows. For example, if $ScoreThreshold\,(I,L^*)$ = 5, this means that 5 face components are visible and properly detected, while the remaining six (11 − 5) face components are not detected due to poor quality of an image or occlusion.

The main reason for introducing $ScoreAppearance\,(I,L^*)$ and $ScoreThreshold\,(I,L^*)$ is the fact that in the interval $[\theta_2, \theta_1>$ there are many false positive and false negative faces, so it is used to minimize the total error rate (TER). The thresholds θ_2, θ_3 and θ_4 define the operating point and they are selected to minimize the TER.

IV. EXPERIMENTAL RESULTS

Two databases were used to perform the experiments. From the XM2VTS database [13], 201 images of frontal faces without occlusions were selected (first column of Fig. 5). For each of these 201 images, six new images were created for simulating six typical configurations of face occlusions by covering different face regions with black rectangles (from the second to seventh column of Fig. 5).

Additionally, 73 images without faces were selected

Figure 5. Results of face detection for frontal face images from the XM2VTS database with three score values S1 for $Score(I,L^*)$, S2 for $ScoreAppearance(I,L^*)$ and S3 for $ScoreThreshold(I,L^*)$. The first column is faces without occlusions and the other columns represent faces with different modes of occlusion. For example, for the 5th face in the top row S1 = − 0.69887, which means that the original method fails to detect the face, while S3 = 3, which means that the modified method detects the face.

from the INRIA person database [14] (Fig. 6). In total, 1,480 (i.e. 7 × 201 + 73) images were used for testing. Let us now justify somewhat the unusual configuration of the testing. Typically, some "in the wild" datasets, such as AFW [15], which contain complex scenes with face occlusions, are used for testing face detection performance. All such databases violate our assumption about the presence of only frontal faces. In order to circumvent this shortcoming of the popular datasets used for face detection testing, we were forced to create an ad hoc face dataset just for the purpose of the proof of concept. Because the XM2VTS database contains only face images, the INRIA person database was used to enable an evaluation of false positive face detections.

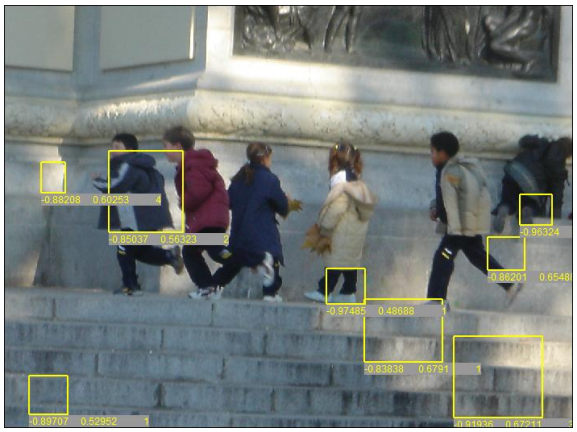

Figure 6. False detections of faces for an image from the INRIA person database with $Score(I,L^*) > -1$. For each detection frame, the values of $Score(I,L^*)$, $ScoreApperance(I,L^*)$ and $ScoreThreshold(I,L^*)$ are given in the same line, respectively. For example, in the image at the lower left frame, the values are $Score(I,L^*) = -0.89707$, $ScoreApperance\ (I,L^*) = 0.52952$ and $ScoreThreshold(I,L^*) = 1$.

First, we ran the originally proposed algorithm with the predefined threshold $\theta_1 = -0.65$ proposed in [5]. The following results were obtained: 160 frontal faces were not detected (false negatives) among 1,407 faces (Fig 9.a)), and 1 image region was detected as a face false positive among 1,480 images (one image from INRIA had one image region with a $Score\ (I,L^*) > -0.65$, Fig. 7.), so we concluded that the robustness was relatively high compared with the state-of-the-art-method [7]. Among 160 false negatives, there were 145 frontal face images having the type of occlusion depicted in the 5th column in Fig. 5, and the remaining 15 having the type of occlusion depicted in the 2, 3, 6, or 7th column in Fig. 5. Note that there is no false negative detection for 201 images without occlusions, and a single false positive was in one of the 73 images from the INRIA person database (Fig 7.).

For our method, we searched for the values for thresholds θ_2, θ_3, and θ_4 to find an operating point that minimizes the total error rate (TER). The minimum TER = 0.087 was achieved for the following threshold values: $\theta_2 = -0.83$, $\theta_3 = 0.69$, and $\theta_4 = 3$. The threshold θ_1 is assumed to be -0.65, as in the original paper [5]. The values of the thresholds $Threshold_k$; $k = 1, 2, \ldots, 11$ are determined as the value of the first quartile of $Apperance_k(I, L^*)$ obtained on 201 frontal images without occlusions (Fig. 8).

Figure 7. The single false detections of faces for the image from the INRIA person database with $Score(I,L^*) \geq -0.65$.

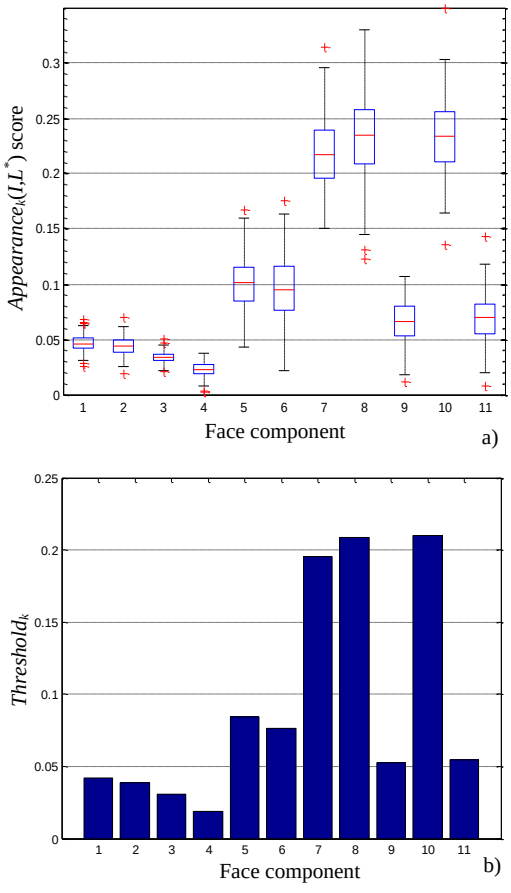

Figure 8. Appearance scores of 11 face components for 201 images of frontal faces without occlusions from the XM2VTS database: a) $Appearance_k(I,L^*)$ distribution, the red line represents the mean value, the box edges the first and the third quintile, the data range is represented with the dashed line, and the red crosses represent the data outliers; b) the values of $Threshold_k$ are obtained as the first quartile from the distribution in a).

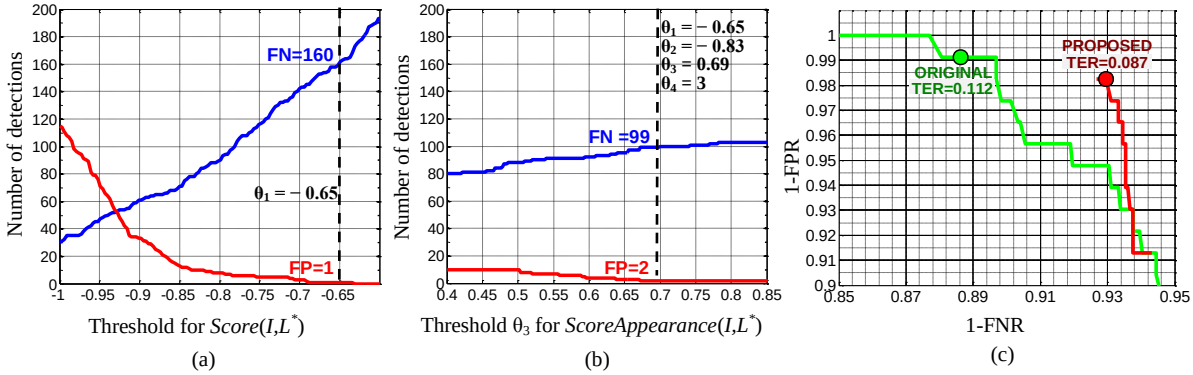

Figure 9. Comparison of the original and the proposed method: a) FN and FP for the original method; b) FP and FN for the proposed method where θ_1 = -0.65, θ_2 =-0.83, and θ_4 = 3; c) 1-FPR and 1-FNR for the original and the proposed method, the green circle represents the working point of the original algorithm for threshold $\theta_1 = -0.65$, and the red circle represents the working point for the proposed algorithm for $\theta_1 = -0.65$, $\theta_2 = -0.83$, $\theta_3 = 0.69$, and $\theta_4 = 3$ for which the TER is minimal.

By using the proposed method and optimal thresholds we achieved a reduction of false negative face detections from 160 to 99, while false positive face detections increased from 1 to 2. The TER was reduced from 0.112 for the original approach to 0.087, which corresponds to an improvement of 22.32% (Fig. 9 c).

V. CONCLUSION

In this paper, we have analysed the influence of face occlusion on the face detection method based on deformable part-based models as originally proposed by X. Zhu and D. Ramanan [5]. We proposed modifications of the method in order to increase the robustness of face detection and to minimize the total error rate. The main modifications are: i) the tree, representing the part-based model of the frontal face, which is partitioned into 11 subtrees representing the face components (left upper edge, left lower edge, right upper edge, right lower edge, left side of the lips, right side of the lips, nose, left eye, left eyebrow, right eye, right eyebrow); ii) the weight of each face component which is obtained based on the results of psychological experiments [12]; iii) the introduction of new scoring functions and thresholds; iv) a new procedure for the robust detection of faces under occlusion based on an evaluation of scoring functions and thresholds.

By using our method on a database consisting of 1,480 images (201 non-occluded face images, 1,206 face images with six types of occlusion and 73 non-face images), the TER was improved by 22.32 % compared to the original method [5]. Note that the experiment was performed only for frontal face images, and thus this work is used only as a proof of concept. Based on the encouraging results, we plan to extend the proposed method by applying the original mixture trees for 13 different face poses and in a total of 146 parts [5].

ACKNOWLEDGMENTS

This work has been supported by the Croatian Science Foundation under project 6733 De-identification for Privacy Protection in Surveillance Systems (DePPSS). It is also the result of activities in COST Action IC1206 "De-identification for Privacy Protection in Multimedia Content".

LITERATURE

[1] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” inProc. IEEE Comput. Soc. Conf. Comput.Vis. Pattern Recog., 2001, pp. I-511–I-518.[

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. “Deepface: Closing the gap to human-level performance in face verification,” In CVPR, 2014, pp. 1701–1708.

[3] M. Fischler and R. Elschlager, “The representation and matching of pictorial structures,” IEEE. Trans. Computers, C-22(1), 1973, pp.67–92.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” IEEE TPAMI, 32(9), 2010, pp. 1627–1645.

[5] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 2879–2886.

[6] J. Yan, Z. Lei, L. Wen, and S. Z. Li, “The fastest deformable part model for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 2497–2504.

[7] S. Liao, A. K. Jain, S. Z. Li, “A Fast and Accurate Unconstrained Face Detector,” IEEE TPAMI, vol. 38, Issue: 2, 2016, pp. 211–223.

[8] K. Hotta, “A robust face detector under partial occlusion,” in Proc. Int. Conf. Image Process., 2004, pp. 597–600.

[9] Y. Lin, T. Liu, and C. Fuh, “Fast object detection with occlusions,” in Proc. Eur. Conf. Comput. Vis., 2004, pp. 402–413.

[10] J. Chen, S. Shan, S. Yang, X. Chen, and W. Gao, “Modification of the adaboost-based detector for partially occluded faces,” in Proc. 18th Int. Conf. Pattern Recog., 2006, pp. 516–519.

[11] L. Goldmann, U. Monich, and T. Sikora, “Components and their topology for robust face detection in the presence of partial occlusions,” IEEE Trans. Inf. Forensics Security, vol. 2, no. 3, 2007, pp. 559–569.

[12] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russel, “Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About,” Proc. IEEE, vol. 94, no. 11, Nov 2006, pp. 1948–1962.

[13] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre; “XM2VTSbd: The Extended M2VTS Database, Proceedings 2nd Conference on Audio and Video-base Biometric Personal Verification (AVBPA99)” Springer Verlag, New York, 1999.

[14] INRIA Person Dataset, http://pascal.inrialpes.fr/data/human/

[15] The Annotated Faces in the Wild (AFW) testset, https://www.ics.uci.edu/~xzhu/face/