Generic Face Detection and Pose Estimation Algorithm Suitable for the Face De-identification Problem

Aleksandar Milchevski[1], Dijana Petrovska-Delacrétaz[2] and Dejan Gjorgjevikj[3]

[1] Faculty of Electrical Engineering and Information Technologies, Skopje, R. of Macedonia
milchevski@gmail.com
[2] TELECOM SudParis, Évry, France
dijana.petrovska@telecom-sudparis.eu
[3] Faculty of Computer Science and Engineering, Skopje, R. of Macedonia
dejan.gjorgjevikj@finki.ukim.mk

Abstract. In this work we tackle the problem of face de-identification in an image. The first step towards a solution to this problem is the design of a successful generic face detection algorithm, which will detect all of the faces in the image or video, regardless of the pose. If the face detection algorithm fails to detect even one face, the effect of the de-identification algorithm could be neutralized. That is why a novel face detection algorithm is proposed for face detection and pose estimation. The algorithm uses an ensemble of three linear SVM classifiers. The first, second and the third SVM classifier estimate the pitch, yaw and roll angle of the face and a logistic regression is used to combine the results and output a final decision. Second, the results of the face detection and a simple space variant de-identification algorithm are used to show the benefits of simultaneous face detection and face de-identification.

Keywords: De-identification, Nonfrontal face detection, Pose estimation, Classifier fusion, SVM, Logistic regression

1 Introduction

The issue of privacy protection in video surveillance has drawn a lot of interest lately. There are different levels of privacy protection schemes that can be applied. Regarding the resolution of the video, people silhouettes or faces need to be protected. In order to be efficient, frontal as well as nonfrontal faces need to be protected. Both of them require a face detection algorithm in order to localize the region that needs to be hidden, encrypted, etc. There are already some research efforts of privacy protection solutions to hide distinguishing frontal facial information and to conceal identity. The available face detection algorithms work well, however the problem of nonfrontal face detection needs to be further studied. In order to be efficient, privacy protection schemes for nonfrontal faces have to be studied also. However, such research efforts are still lacking.

S. Loshkovska and S. Koceski (eds.), *ICT Innovations 2015*,
Advances in Intelligent Systems and Computing 399,
DOI: 10.1007/978-3-319-25733-4_23

2 Previous Work

2.1 Face Protection

The problem of face privacy protection can be in general defined as finding a way for the protection of the identity of the subject in the image or the video, while keeping the usability of the image or the video. The image is transformed in such a way that the subject cannot be identified by face recognition algorithm or a human observer. The definition of the usability of the video and the answer to the question: "Why not just delete the face region?" depend on the specific area of use of the de-identified video e.g. usually, it is important for the de-identification algorithm to retain the facial expressions.

In [1] the k-same algorithm is proposed. The algorithm determines the similarity between faces based on a distance metric and replaces the face with a new face which is an average of components of several faces. However, in order for the algorithm to be successful all of the used faces should be from different subjects. Several experiments are done using Eigen face recognition algorithm. The experiments made show that the naïve approaches, such as blurring, pixelization, adding noise, etc. although produce results from which a human observer cannot identify the subject, they do not provide good protection against face recognition algorithms.

The de-identification algorithm presented [2] is based on the k-same algorithm previously described; however, an AAM (Active Appearance Model) is first fitted for the face which is de-identified. The result of this improvement is that the output of the de-identification algorithm is without artifacts i.e. with better visual quality. The experiments made are also with an Eigenfaces recognition algorithm and they show successful de-identification. The experiments are limited on frontal faces and the AAM ground-truth is manually established.

In [3] a scrambling technique as a solution to the face de-identification problem. The sign of the H.264 transformed image is pseudo randomly flipped. The advantages of this approach are the low computational cost and the full reversibility of the applied modification of the image. The authors also provide an alternative of the algorithm by using permutation of the coefficients instead of sign change.

In [4] an algorithm for privacy protection in video surveillance is proposed which uses geometric warping of the face region. Several experiments are made using the OpenCV's Viola Jones implementation for the face detection and FLDA (Fisher Linear Discriminant Analysis) for the face recognition.

In [5] a system for automatic face replacement in images is proposed. First the pose of the face is estimated and a face with similar face is found from a large data set (yaw and pitch angles differ by no more than 3^{o} from the yaw and pitch of the original face). After that, the face is replaced while keeping some of the original features. The new face is than color and light adjusted. The algorithm is fully automated and produces highly plausible results.

2.2 Face Detection

Face detection is probably one of the most researched problems in the areas of computer vision and image processing. There are a vast number of published algorithms and different approaches, but the most revolutionary is the Viola – Jones algorithm.

Very good survey on face detection algorithms is done in [6]. In the following text the most important and recent algorithms are summarized.

In the work of [7] a multi-view face detector is proposed using a detector pyramid. They use coarse-to-fine approach to deal with the out of plane rotations of the head. The full range of possible rotations is partitioned into several partitions, and ranges are narrowed as the level of the pyramid increases. The detector at the top of the pyramid is very simple with a main task to reject as much of the non-face images. They deal with in-plane rotations by applying their detector on rotated test images with 30^o and -30^o rotations.

In [8] the authors present a simple solution for the multi-view face detection problem. They use the Viola – Jones framework, however they modify it by using LUT as weak classifiers instead of the stump weak classifier, used in the original work of Viola – Jones.

Human faces are divided into several categories and a cascade is trained for every category individually. For the yaw axis there are 5 categories with the following intervals: [-90°, -50°], [-50°, -20°], [-20°, +20°], [+20°,+50°], [+50°, +90°].

The authors in [9] build a low dimensional face manifold parameterized by the pose of the face. They train a convolution network and use Energy Minimization Framework to map the face images onto the face manifold and non-face image far away from the face manifold. The authors elaborate that the multi-view face detection and pose estimation are very closely related so they should not be done separately. The authors claim that the system is highly reliable, and runs in real time on standard hardware.

In [10] the authors present a method for simultaneous face detection, pose estimation and landmark localization. For the landmark localization they use mixtures of trees with a shared pool of parts, instead of densely-connected elastic graphs. They used HOG (Histogram of Oriented Gradients) as a feature descriptor. The authors claim that their method is better or comparable to the state-of-the-art algorithms in all three categories. The presented results clearly show that the algorithm works well, however, they have limited the test set to images with relatively big faces, so that the landmarks are clearly visible.

In [11] the authors motivated by the success of [10] propose a face detection algorithm which uses part models. However, they propose Cascade Deformable Part Models, arguing that the use of Tree Structure Model in [10] is suboptimal for face detection, because is too slow and limited to high resolutions. The presented results show that the algorithm works well even on the AFLW data set. The average detection time reported for the method is 0.52s, compared to 26.06s for the TSM algorithm published in [10].

3 Proposed Algorithm for Generic Face Detection

Usually the HOG is described as a feature descriptor with great descriptive power, but also as very computationally expensive. In FDDB [12] the best scoring algorithm uses HOG for the feature description. Many authors have suggested simplification or ways to compute the feature in a faster way. Because of the superior descriptive power the HOG is chosen as a feature descriptor.

Almost all of the reviewed work on unconstrained face detection have treated the face detection and pose estimation as a combined problem. However, the two problems are conflicting: the pose estimation tries to find differences between the several view groups and the face detection tries to find similarities in all view groups. Nevertheless, because of the difficulty of the multi-view face detection problem (large number of possible variations) the pose information should be used even if the detection of the face is of main concern.

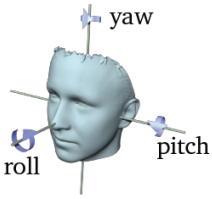


Fig. 1. The three rotation angles used to describe the pose of the head

Analyzing the data set used for training (AFLW) [13] and the effect of the observed face on the image regarding the three axis (Fig. 1), three separate classifiers are proposed:

1. SVM – pitch (nodding)
   An SVM (Support Vector Machine) classifier using four classes: three classes corresponding to faces with values for the pitch in the intervals of [-90°,-12°), [-12°,12°], (12°,90°] and a fourth class corresponding to images without faces.
2. SVM – roll (In plane rotation)
   An SVM classifier using six classes: five classes corresponding to faces with values for the roll in the intervals of [-90°,-30°), [-30°,-12°), [-12°,12°], (12°,30°], (30°,90°] and a sixth class with images without faces. (An assumption is made that the maximum in-plane rotation of the face is 90°. The detector can be applied on a 180° rotated test image if other faces are expected).
3. SVM – yaw (out of plane rotation)
   An SVM classifier using eight classes: seven classes corresponding to faces with values for the yaw in the intervals of [-108°,-60°), [-60°,-30°), [-30°,-12°), [-12°,12°], (12°,30°], (30°,60°], (60°, 108°] and a eighth class corresponding to images without faces.

The three classifiers are trained individually with different training sets. Next, the decisions and the probability estimates from all three classifiers are combined to form a final decision if the tested image should be classified as face or not. An ensemble of classifiers is also used [14]. First, they use three independently trained SVM classifiers for frontal, profile and semi-profile face. The probability estimations of the three classifiers are combined using SVM regression which should output 4 distinct values representing the four classes.

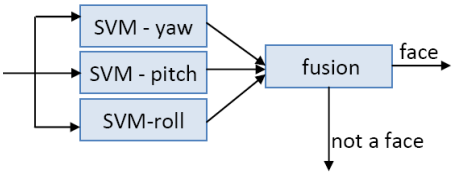


Fig. 2. Block diagram of the proposed face detection algorithm

In Fig. 2 a block diagram of the proposed algorithm is given. A block with a size of 36x36 is first extracted and the HOG is calculated. The feature vector is inputted to the three linear SVM classifiers. The outputs of the three classifiers are given to the decision fusion block which gives the final decision of the tested block being a face or not.

3.1 Training

Training of the linear SVM classifiers (for roll, pitch and yaw). For the first step of the proposed algorithm three linear SVM are trained using the LIBLINEAR [15] library.

Training set with images containing faces. For the training of the three linear SVM classifiers the training sets with images containing faces are created using the AFLW data set. The first 15.000 images of the data set are used for this step of the training. The face region is extracted from the image using a square region that contains all of the landmarks provided with the data set. The square region is than scaled to a block of size 36x36. Every block is than mirrored in order to increase the number of face images. The obtained training set is split into three equal subset, which are used for the separate SVM classifiers. In this way, every SVM classifier is trained with approximately 10.000 independent positive face samples.

Training set with images without faces. In order to obtain independent training and to find difficult samples without faces, the ILSVRC[16] data set is used. Only the images which are labeled that do not belong to the person class are used. A large number of feature vectors are extracted from one image in the following way: First, a feature vector is calculated for every 36x36 block in the image without overlapping. Than the image is scaled by a factor of two and again a feature vector is extracted for every

36x36 block. The procedure is repeated until the rescaling of the image produces an image with height or width less than 36.

The training set with images without faces is divided into several subsets which will be used separately. A subset of the training set is created by extracting feature vectors from the data set until the total number of negative feature vectors is above 20.000.

Choice of C- parameter for the SVM. The choice of the C -parameter controls the trade-off between complexity of decision rule and frequency of error [17]. If the parameter is too large, a high penalty for nonseparable samples is introduced and there is an increased chance of overfitting. If the parameter is too small, there is an increased chance of underfitting. That is why the choice of the C parameter is analyzed.

In order to obtain an optimal value for C parameter a grid search has been performed using 5-fold cross-validation. The C parameter is varied exponentially in the range from 2^{-5} to 2^{3}, a value 10 is also included as a value other authors used. An independent subset of the training set with images without faces is formed for the training of every SVM as explained above.

The results from the cross-validation show that a value of 2^{-4} is a good choice for every SVM classifier.

Mining for hard samples without faces. An iterative procedure in order to obtain hard samples without faces was performed. An SVM classifier is trained with a new subset of the training set with images without faces. After the training is completed the classifier is tested on the training set and the falsely classified as faces are extracted. This procedure is repeated for 15 iterations. After all of the iterations are finished the SVM is trained using another subset of the training set with images containing faces and all of the extracted hard samples.

Decision Fusion. Two methods are tested to fuse the final decision, using an SVM and using LR.

With the first method a final decision about the block is obtained with a new kernel – SVM classifier. As a feature vector the outputs of all of the three linear SVM are used. Every multiclass SVM is implemented using several binary SVM classifiers with the one-against-all approach. The margins outputted from every binary SVM are used as features for the training of the SVM, which will output the final decision. In that way, the feature vector has 18 features now, because 4 margins are obtained from the first linear SVM, 6 from the second and 8 from the third.

A separate training set is created using 5000 images from the AFLW data set, with the mirroring of the images about 10000 independent positive samples are obtained. (These images are exclusively used for the training of the fusing algorithm.)

Choice of Cost (C) parameter and gamma for the SVM. The optimal values of the C parameter and the gamma value for the radial basis function are estimated using grid search and 5-fold cross-validation. The range for the gamma parameter is from 2^{-15} to

2^3, and the range for the C parameter is the same as for the previous tests. The cross-validation accuracy yielded highest accuracy for C=4 and gamma=2^{-3}.

Mining for hard samples without faces. A similar iterative procedure is used to obtain hard samples without faces for the training of the kernel-SVM classifier. A new subset of the training set without faces is created for each iteration with a size of about 20.000 samples. The classifier is tested on the training set and the samples falsely classified as face are saved. Additionally, the support vectors corresponding to images without faces are also saved. The procedure is repeated for 50 iterations.

Final Training. After all the iterations have been finished a new classifier is trained with all of the extracted hard samples and a new subset of training set with images without faces. Again a grid-search and cross-validation is done in order to test the accuracy of the system for different values.

Another way to fuse the decision was also tested. For this alternative the decision is fused using logistic regression. The logistic regression classifier was trained using the same training set as for the final kernel-SVM classifier.

3.2 Classifier Test

In order to compare the two ways to fuse the decision an experiment has been made using the remaining images from the AFLW data set (about 2000 independent images, 4000 in total using mirroring) and a new subset from the ILSVR data set. The ROC curve is plotted and shown on Fig. 3. It can be seen that decision fusion with logistic regression outputs superior results compared to the kernel-SVM.

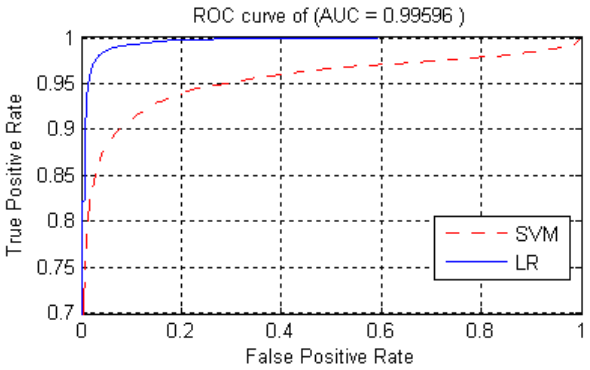


Fig. 3. ROC curve comparison of the two ways to fuse the decision

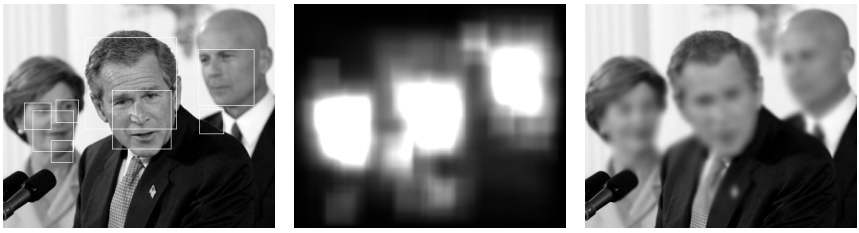
3.3 Face Detection

The face detection algorithm is implemented using a standard window-sliding approach. Every 36x36 block is tested in the image using the designed system.

The image is than rescaled and the same procedure is repeated. The regions with a probability higher than a threshold (0.99) are than grouped to form the final detection result. A probability map for the whole image is also created by adding all of the outputted probabilities. The procedure is repeated until the height or the width of the rescaled image is less than 36.
Fig. 4.a) and Fig. 5.a) show the output of the algorithm for the selected images. The images are part of the FDDB.

a) b) c)

Fig. 4. Results obtained with the proposed method



a) b) c)

Fig. 5. Results obtained with the proposed method

3.4 Face De-identification

The output of the designed face detection algorithm can be used for de-identification. To show that, the probability map outputted from the face-detection is used to do a space variant blurring. The blurring is not a good choice for a de-identification algorithm and serves only to show the benefits of simultaneous face detection and face de-identification. The PSF (Point Spread Function) is chosen to be Gaussian window

with a size of 19x19 and a standard deviation proportional to the output from the probability map. Fig. 4.c) and Fig. 5.c) show the output of the de-identification algorithm.

4 Conclusion

In this paper a review of the most important and most recent algorithms for face detection algorithm was done. A new approach was proposed for face detection in unconstrained condition using an ensemble of linear SVMs. The algorithm was tested on several images in order to evaluate the performance. Two algorithms were tested for the fusion of the decisions, SVM and logistic regression. The fusion with the logistic regression yield better results. The tests also showed that a better way should be used to group the rectangles with high probability of face. The accuracy of the detection of profile faces should also be improved.

The output of the algorithm is used to implement a simple de-identification algorithm. Although the proposed de-identification method is simple it shows the benefits of the simultaneous solution of the face detection and the face de-identification problem.

Acknowledgements. This work was partially done during a STSM (Short Term Scientific Mission), supported by the COST Action IC1206, hosted by TELECOM SudParis and prof. Dijana Petrovska-Delacrétaz.

References

1. Sweeney, E.N.L., Malin, B.: Preserving privacy by de-identifying facial images (2003)
2. Gross, R., Sweeney, L., Torre, F.D.l., Baker, S.: Model-based face deidentification. In: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on. pp. 161-161. IEEE (2006)
3. Dufaux, F., Ebrahimi, T.: A framework for the validation of privacy protection solutions in video surveillance. In: Multimedia and Expo (ICME), 2010 IEEE International Conference on. pp. 66-71. IEEE (2010)
4. Korshunov, P., Ebrahimi, T.: Using warping for privacy protection in video surveillance. In: Digital Signal Processing (DSP), 2013 18th International Conference on. pp. 1-6. IEEE (2013)
5. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. ACM Transactions on Graphics (TOG) 27(3), 39 (2008)
6. Zhang, C., Zhang, Z.: Boosting-based face detection and adaptation. Synthesis Lectures on Computer Vision 2(1), 1-140 (2010)
7. Li, S.Z., Zhu, L., Zhang, Z., Blake, A., Zhang, H., Shum, H.: Statistical learning of multi-view face detection. In: Computer Vision ECCV 2002, pp. 67-81. Springer (2002)
8. Wu, B., Ai, H., Huang, C., Lao, S.: Fast rotation invariant multi-view face detection based on real adaboost. In: Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on. pp. 79-84. IEEE (2004)

9. Osadchy, M., Cun, Y.L., Miller, M.L.: Synergistic face detection and pose estimation with energy-based models. The Journal of Machine Learning Research 8, 1197-1215 (2007)
10. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2879-2886. IEEE (2012)
11. Orozco, J., Martinez, B., Pantic, M.: Empirical analysis of cascade deformable models for multi-view face detection (2013)
12. Jain, V., Learned-Miller, E.G.: Fddb: A benchmark for face detection in unconstrained settings. UMass Amherst Technical Report (2010)
13. Kostinger, M.,Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 2144-2151. IEEE (2011)
14. Yan, J.: Ensemble svm regression based multi-view face detection system. In: Machine Learning for Signal Processing, 2007 IEEE Workshop on. pp. 163-169. IEEE (2007)
15. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871-1874 (2008)
16. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision pp. 1-42 (2014)
17. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273-297 (1995)