On Performance Evaluation of Face Detection and Localization Algorithms

Vlad Popovici and Jean-Philippe Thiran
Swiss Federal Institute of Technology
Signal Processing Institute
CH-1015 Lausanne, Switzerland
Vlad.Popovici@epfl.ch, JP.Thiran@epfl.ch

Yann Rodriguez and Sébastien Marcel
Dalle Molle Institute of Perceptual Intelligence
Computer Vision Group
CH-1920 Martigny, Switzerland
rodrig@idiap.ch, marcel@idiap.ch

Abstract

When comparing different methods for face detection or localization, one realizes that just simply comparing the reported results is misleading as, even if the results are reported on the same dataset, different authors have different views of what a correct detection/localization means. This paper addresses exactly this problem, proposing an objective measure for the goodness of a detection/localization for the case of frontal faces. The usage of the proposed technique insures a fair and unbiased way of reporting the results, making the experiment repeatable, measurable, and comparable by anybody else.

1. Introduction

Human face detection became one of the most active research domains of computer vision. An impressive number of papers has been published during the last decade reporting various methods for face detection (FD) and localization (FL). Some of them are designed to be used in specific environments while others are thought to be robust enough for a general usage.

However, in evaluating and comparing their performances one needs not only the formal description of the algorithm, but also a strict experimental protocol and a clear definition of the performance criteria. Clearly, the performance criteria are problem–dependent and may cover diverse aspects of practical importance like complexity of the algorithm, hardware requirements, scalability and so on. But the most important ones remain the detection rate and localization precision. In order to have a fair comparison of the results, not only the test images must be the same, but also an objective measure of the *goodness* of the detection/localization is needed. While most of the published methods use publicly available datasets, others report results on data that is not so easily available. A number of databases have emerged as standard testbeds for the face detection (e.g. the combined test sets from CMU [2]) and face localization (e. g. XM2VTS [3], Banca [1] or BioID [6]) algorithms. While establishing a common pool of data is an important step forward, there are still a number of issues that are not generally agreed upon and that may bias the comparisons. For example, in the CMU database there are some hand-drawn faces. Should they be considered as 'real' faces, or not?! Or, and arguably the most important issue, what does a *good face detection/localization* mean?

Most of the papers generally only provide detection and error rates to show the quality of their system, but rarely mention the way they count the detections and the errors to compute those rates. A *good* detection for someone may appear as not sufficient for someone else. In general, two kind of methods are used to count the detections: manual and automatic. In the first case, the faces are manually identified by humans, like in [5]. Besides being tedious, this technique is above all very subjective. In the second case, people usually consider the difference between the detected eye positions and the groundtruth positions. A correct detection is accounted if this difference is under a given threshold. As we will explain, there are a number of problems with this approach, basically due to the subjectiveness of the measurement or to some geometric issues (like scale-dependence).

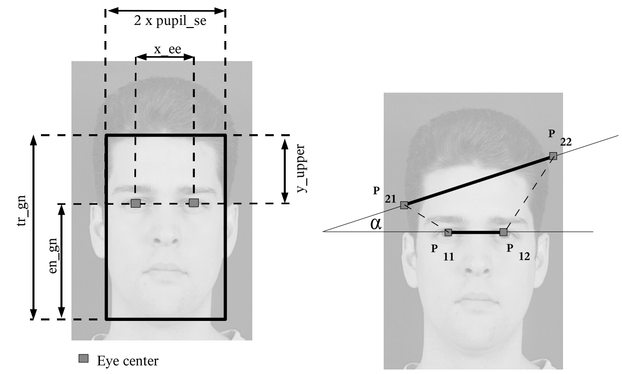
Jesorsky et al. [6] recently introduced a relative error measure. They used the maximum of the distances between the true and the estimated eye center positions divided by the distance between the expected eye centers (scale independence). A region is considered as being a face if the relative error is less than a given threshold. The drawback of this approach is that it is not possible to differentiate errors in translation, rotation and scale.

The main contribution of this paper is to propose a general, objective measure for assessing the performances of the FD and FL algorithms. The proposed measure is flexible enough to allow adaptation to different interests by tuning the weights of specific types of errors. We will present also its applicability for a real face detector and show how it can be used for assessing the performances of the method.

The remaining of the paper is organized as follows: section 2 gives a short overview of the parameters used for modeling a face, section 3 introduces the face detection scoring function and describes its application. In section 4 we give a brief description of the detection method used as example, and we present the results obtained on a standard database (XM2VTS [3]). Finally, we draw some conclusions.

2 Anthropometric Face Modeling

The first step to any face processing is to choose a face model. This model will be used to collect faces according to the groundtruth for training purposes. Usually, it is represented by a bounding box. The face bounding box is determined using face/head anthropometry measures [4] according to a face model (Fig. 1(a)). The face bound-



(a) Face model (b) Detection parameters

Figure 1. Face modeling: 1(a) The face model using eyes' centers coordinates and facial anthropometry measures. 1(b) The relative position of an estimated and a true position of the eyes.

ing box w/h crops the physiognomical height of the face. The width w of the face is given by zy_zy/s where s = 2·pupil_se/x_ee and x_ee is the distance between eyes in pixels. In this model, the ratio w/h is equal to $15/20$. Thus, the height h of the face is given by w·20/15 and y_upper = h·(tr_gn - en_gn) / tr_gn. For the constants pupil_se (pupil-facial middle distance), en_gn (lower half of the craniofacial height), tr_gn (height of the face), and zy_zy (width of the face) we use the values $33.4, 117.7, 187.2,$ and 139.1 respectively, from [4].

In practice, there are two commonly used approaches for face detection: one is based on components where one searches for different parts of the face (eyes, nose, and so on) and then decides if a face is present or not, and one based on holistic models (like eigenfaces) where the face is treated as whole object. While in the first case, the position of the eyes is explicitely sought, in the second case this is implicit. Indeed, one can estimate an average position of the eyes from the training set and use this estimation for the detections. Moreover, using the approach described above, the position of the eyes is fixed within the bounding box, so there is no need for postprocessing the detection results.

3 Evaluating Face Detection and Localization

In the following we will construct a scoring function for evaluating the results of FD and FL algorithms, that is adapted for frontal face case. As the position and the size of the bounding box of a face can be determined knowing the coordinates of the eyes, we will consider them as the relevant attributes of a detection/localization. The goal is to build a scoring function that assigns high scores of 1.0 (or close to 1.0) to the good detections and 0.0 (or close to 0.0) to the bad ones. In the same time, the function must possess the following properties: (1) it has to be continuous and smooth; (2) it has to be invariant to translations, scalings and rotations (TSR-invariant); (3) it has to accommodate some degree of uncertainty that are inherent in practice.

3.1 Scoring function

Let now x denote the criterion (e.g. $\cos x$ or the others defined below) we want to score and let ψ be the scoring function. It is clear that the requirements above are general enough and they do not uniquely identify a function. As such, we have chosen the following form for the scoring function:

$$\psi(x;\gamma,\delta,\mu) = \begin{cases} e^{-\gamma^2((x-\mu)+\delta)^2}, & \text{if } x \le \mu-\delta \\ 1, & \text{if } \mu-\delta < x < \mu+\delta \\ e^{-\gamma^2((x-\mu)-\delta)^2}, & \text{if } \mu+\delta \le x \end{cases} \tag{1}$$

where γ,δ,$\mu \in \mathbb{R}$ and $\delta \ge 0$, are some suitably chosen parameters of function ψ. In the following we will denote by $\theta = (\gamma, \delta, \mu)$ the set of parameters when we will not need to address them individually. Figure 2 shows the plot of ψ for different combinations of the parameters. Normally, one would choose μ such that it is the correct/expected value for x. On the other hand, δ defines the width of the constant region of value 1 and corresponds to the degree of tolerance one accepts in the precision of x. Finally, γ controls the slope of the two branches and must be set to a suitable value. For all these parameters we will present some values that are sensible for our application.

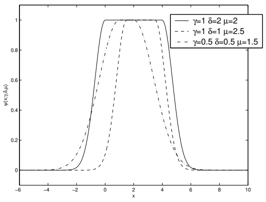


Figure 2. Plots of ψ function for different combinations of parameters.

3.2 Goodness of a dectection

Let us now define the criteria by which we evaluate the goodness of a detection, where we compare the *detected* position of a face with its *groundtruth* position. As explained before, they must be TSR-invariant. Consider the situation in Figure 1(b), where we let $T(P_{11}, P_{12})$ be the true groundtruth position of the eyes and $D(P_{21}, P_{22})$ be the detected position. It is clear that the angle between the two support lines ($\overline{P_{11}P_{12}}$ and $\overline{P_{21}P_{22}}$) is a first important parameter (and is TSR-invariant). It accounts for errors due to the estimation of the rotation angle of our detection. In practice we will use the cosine of the sharp angle formed by the two support lines (denoted hereinafter by $\cos\alpha$). The other criteria we consider are defined in terms of distances, which normally are scale-dependent, but by dividing them with a normalizing term (defined by the distance between the correct positions of the eyes), we will obtain the TSR-invariance:

$$d_1 = \frac{\|P_{21}P_{22}\|}{\|P_{11}P_{12}\|}, d_2 = \frac{\|P_{11}P_{21}\|}{\|P_{11}P_{12}\|}, d_3 = \frac{\|P_{12}P_{22}\|}{\|P_{11}P_{12}\|}. \quad (2)$$

All these four criteria are TSR-invariant and do completely describe the relative positions of the two couples of points.

For every possible detection, we will use the function $\psi(\cdot;\theta)$ to score each corresponding parameter individually (with properly chosen values for θ), obtaining four parameter–scores, $\psi_{1,\ldots,4}$ and we will take as the final score of the detection D when the groundtruth is T, the weighted sum of these individual scores:

$$\Psi(D, T) = \sum_{i=1}^{4} \omega_i \psi_i, \qquad \text{with} \sum_{i=1}^{4} \omega_i = 1 \qquad (3)$$

If one wishes to control the importance of a given kind of errors then one has just to adapt the values of ω_i to reflect his/her wishes. However, in this paper, we use an uniformly weighted sum, i.e. we take as the final score the average value.

3.3 Choice of γ, δ, and μ

Before using the scoring function, one has to choose appropriate values for the parameter set θ for each criterion considered. Two issues are considered in our choice: the set of allowable values for the (above defined) four criteria, and the tolerance level we set when deciding what is a good detection or localization. From this perspective, we consider the face localization process as being a more precise detection, i.e. we tolerate smaller deviations of the estimated position from the true position. This means that, once we have the results produced by a system (face detector or localizer), given as a list of estimated positions of the eyes, we can easily evaluate its performances from both perspectives, detection or localization, by simply changing the value of θ.

Examining the plots in Figure 2 and using Eq.(1), it can be seen that the region of value 1 has a width of 2δ, being centered at μ. The value of γ defines the slopes of the two branches of the ψ function and we choose it such that the values outside the acceptable range will score at most 0.001. We now set the performance criteria that will define the values of these parameters. In the case of face *detection*, allowing for $\pm 10°$ error in orientation estimation, $\pm 10\%$ in scale and position estimation leads to the following:

- $\psi(\cos\alpha) = 1$ for $|\alpha| \in [0, \pi/18]$ and $\psi(\cos\alpha) < 0.001$ for $|\alpha| > \pi/12$, $\Rightarrow \theta_{\cos\alpha} = (139.2, 0.0152, 1)$;
- $\psi(d_1) = 1$ for $d_1 \in [0.9, 1.1]$ and $\psi(d_1) < 0.001$ for $d_1 < 0.75$ or $d_1 > 1.25$, $\Rightarrow \theta_{d_1} = (17.52, 0.1, 1)$;
- $\psi(d_{2,3}) = 1$ for $d_{2,3} \in [0, 0.1]$ and $\psi(d_{2,3}) < 0.001$ for $d_{2,3} > 0.6$, $\Rightarrow \theta_{d_{2,3}} = (5.26, 0.1, 0)$.

The corresponding performance criteria for a *localization* are more restrictive ($\pm 5°$ error in orientation estimation, $\pm 2.5\%$ in scale and 5% in translation estimation):

- $\psi(\cos\alpha) = 1$ for $|\alpha| \in [0, \pi/36]$ and $\psi(\cos\alpha) < 0.001$ for $|\alpha| > \pi/18$, $\Rightarrow \theta_{\cos\alpha} = (230.81, 0.0038, 1)$;
- $\psi(d_1) = 1$ for $d_1 \in [0.975, 1.025]$ and $\psi(d_1) < 0.001$ for $d_1 < 0.95$ or $d_1 > 1.05$, $\Rightarrow \theta_{d_1} = (2.84, 0.025, 1)$;
- $\psi(d_{2,3}) = 1$ for $d_{2,3} \in [0, 0.05]$ and $\psi(d_{2,3}) < 0.001$ for $d_{2,3} > 0.3$, $\Rightarrow \theta_{d_{2,3}} = (10.51, 0.05, 0)$.

Finally, *a good detection/localization* is considered to be any detection/localization that scores at least $\Psi_0 = 0.5$.

These values ($\theta_{\cos\alpha}, \theta_{d_{1,2,3}}, \Psi_0$) should be considered as reference values and any reported results for face detection should be based on them in order to have a common comparison basis. All the results reported in section 4 are using these specific values.

3.4 Error rates

Having defined what a good FD/FL means is not enough: usually we are given a set of positions in the image that are the outcomes of an application of a detector/localizer, and a set of groundtruth positions and we have to estimate the detection/localization rate and the false alarm rate.

We proceed as follows: having a list of detected positions $\{D_1, \ldots, D_n\}$ and a list of ground truth positions $\{T_1, \ldots, T_m\}$ we put in correspondence each true position T_i with a detection D_j by searching that D_j that has the highest score $\Psi(D_j, T_i)$:

$$j = \arg \max_{k=1,\ldots,n} \Psi(D_k, T_i), \tag{4}$$

and we add the pair (D_j, T_i) to a list of good detections if the score is higher than Ψ_0. If for a given T_i there are more D_j's or vice–versa, we solve the ties by randomly selecting only one correspondence. We end up with a list r pairs detection/localization–true position, $\{(D_{j_k}, T_{i_k}) | k = 1, \ldots, r\}$. Finally, we define the *detection/localization rate* to be r/m and the *false alarm rate* to be $1 - r/n$.

4 Experiments

Viola and Jones [7] recently proposed a real-time state-of-the-art frontal face detector. Instead of directly using pixel information, they used a set of simple fast-to-compute features, named Haar-like. A variant of AdaBoost [8] selects relevant features and combines linearly weak classifiers into a strong one. By assembling such strong classifiers in a cascade, Viola and Jones improved the detection performance while reducing the computation time.

We use the face detection algorithm mentioned above to detect the faces from the XM2VTS database [3]. Before performing the evaluation we have selected the performance criteria, corresponding to the detection scenario, presented in section 3.

Figure 3 presents the histogram of the scores obtained by the d_2, d_3 parameters (the other two parameters, $\cos\alpha$ and d_1, obtained scores higher than 0.9 in 99% of cases). As it can be noted, the detector is less accurate in estimating the position (the offset) of the face. Note that in interpreting the distribution of scores in Figure 3 one has to take into account also the constraints imposed; in fact, 98.31% of detections scored more than 0.001 for the d_2 parameter, meaning that the corresponding error was less than 0.4.

Finally, the detections scores are obtained by means of Eq.(3), with $\omega_i = 0.25$. The distribution of scores can be seen in Figure 4. Counting as good detections only those detections that obtained a score higher than 0.5 leads to the conclusion that the detector has an accuracy of 93.22%.

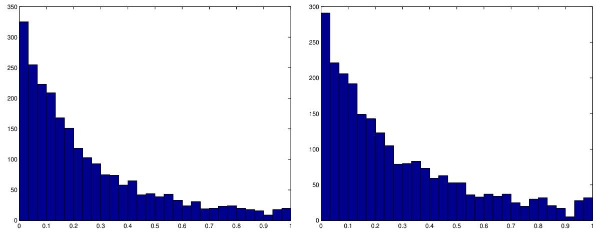


Figure 3. Scores for individual parameters (d_2, d_3) obtained on XM2VTS database.

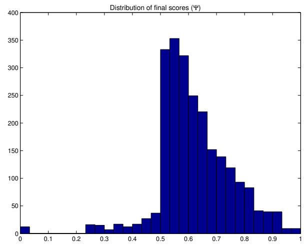


Figure 4. Final score distribution on the XM2VTS database. 93.22% of detections obtained a score higher than 0.5.

If one is interested in evaluating the performances of this method from a localization perspective, then one has to use the second set of constraints. In our specific case, this means that only 48.14% of the localization are considered as good localizations.

5 Conclusions

In this paper we have presented an objective measure for evaluating the face detection/localization results. The measure is independent of the method employed and provides the basis for an unbiased comparison of different techniques. Moreover, it unequivocally defines what a good detection or localization means, removing the ambiguity present in a large number of reported experiments.

It has to be emphasized the flexibility one has in using the scoring function and in interpreting the results from either a detection or a localization perspective. Also, the fact that each type of error can be individually analyzed provides useful informations for improving the performances of the system. As face detection or localization is rarely the final goal in face processing, analyzing the scores may provide useful hints in tuning the upper levels of processing.

Based on the scoring function, we have defined the detection/localization rate and the false alarm rate. What is still left to be done is to define a strict experimental protocol that should be accompanied by a significant data corpus so that anybody will be able to compare his/her results with the other available techniques.

It is our intent and hope that other researchers will adopt this measure when reporting their experiments, making the results obtained more meaningful for the whole research community.

Acknowledgments

The authors want to thank the Swiss National Science Foundation for supporting this work through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)".

References

[1] Bailly-Baillière, E. and Bengio, S. and Bimbot, F. and Hamouz, M. and Kittler, J. and Mariéthoz, J. and Matas, J. and Messer, K. and Popovici, V. and Porée, F. and Ruiz, B. and Thiran, J.-P. The BANCA Database and Evaluation Protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, 2003.

[2] H. Rowley, S. Baluja and T. Kanade. Rotation Invariant Neural Network-Based Face Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 38–44, 1998.

[3] K Messer and J Matas and J Kittler and J Luettin and G Maitre. XM2VTSDB: The Extended M2VTS Database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999.

[4] L.G. Farkas. *"Anthropometry of the Head and Face"*. "Raven Press", 1994.

[5] M.-H. Yang, D. Roth and N. Ahuja. A SNoW-Based Face Detector. In *Advances in Neural Information Processing Systems, S. A. Solla, T.K. Leen and K.-R. Muller (eds), pp. 855-861, MIT Press*, 2000.

[6] O. Jesorsky, K. Kirchberg and R. Frischholz. Robust Face Detection Using the Hausdorff Distance. In *Proceedings of Audio and Video based Person Authentication*, pages 90–95, 2001.

[7] P. Viola and M. Jones. Robust Real-time Object Detection. In *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, 2001.

[8] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the IEEE International Conference on Machine Learning*, pages 148–156, 1996.