Facial Landmarks Detection Based on Correlation Filters

Gabriel M. Araújo\textsuperscript{1}, Waldir S. S. Júnior\textsuperscript{1,2}, Eduardo A. B. Silva\textsuperscript{1} and Siome K. Goldenstein\textsuperscript{3}

\textsuperscript{1}PEE-COPPE/DELI-Poli, Federal University of Rio de Janeiro, POBox 68504, Rio de Janeiro, RJ, CEP 21945-970, Brazil.
\textsuperscript{2}DET/CETELI, Federal University of Amazonas, R. O. Octávio Jordão Ramos, 3000, Coroado I, Manaus, AM, CEP 69077-000, Brazil.
\textsuperscript{3}IC, State University of Campinas, POBox 6176, Campinas, SP, CEP 13084-971, Brazil.
e-mails: gmatos@lps.ufrj.br, waldirjr@ufam.edu.br, eduardo@lps.ufrj.br, siome@ic.unicamp.br

Abstract—The problem of locating facial landmarks is important in many applications such as security, 3D modeling and expression recognition. In this paper, we present a new facial landmarks detection system. The core of the proposed system is a cascade of a new detector based on correlation filters. This detector inherits from the correlation filters the tolerance to small variations of the desirable pattern. This detector is referred to as IPD (Inner Product Detector) and, different from the correlation filters, is suitable for features with a small number of dimensions. In our experiments we use cross-validation of 503 images from BioID database. We verify that the proposed method provides competitive performance when compared to Support Vector Machines.

I. INTRODUCTION

Recently, the scientific community has focused attention on the problem of facial landmarks detection, because many applications such as security, human-machine interfaces and 3D modeling, use landmarks. Current approaches can be grouped in two main categories: local and global methods [1]. The global methods are capable to detect more landmarks with robustness than the local ones, which can detect landmarks quickly [1].

Most global methods [1] use either ASM (Active Shape Models) [2] or AAM (Active Appearance Models) [3]. In the ASM case, the algorithm searches for the best match using a shape model, in AAM, the objective of the algorithm is to obtain the best match with a combined model using texture and shape.

In local methods, the algorithms, detect landmarks, like the corner of the eyes or the tip of the nose without using information from other parts of the face. We can find examples of local methods in [4], where a cascade is used to select features extracted by Gabor filters, and in [5], where the landmark detection is performed by feature extraction with Haar filters and a cascade of boosted classifiers.

In this paper, we propose a new facial landmark detection system to detect landmarks in human faces. It is a local method which consists of three steps: pre-processing, classification and post-processing. The classification step, that is the core of the system, is a cascade of classifiers using a detector based on correlation filters called IPD (Inner Product Detector). A linear SVM-based system (information on the SVM library used can be found in [6]), where the proposed IPD was replaced by the SVM, was used for performance comparison. Our method has shown competitive performance to the SVM-based ones.

The remainder of this paper is organized as follows. In section II, we present the proposed method and all steps of the method focusing on IPD. We describe the experiments and presents the obtained results in section III. In section IV we conclude and give some ideas for future works.

II. PROPOSED METHOD

The proposed method is a system to detect a set of landmarks in frontal faces. This system is comprised of three parts. In the first, the pre-processing step, we perform illumination correction, face detection, face rescaling and search region reduction. The second part is the core of the system, where there is a cascade of IPD detectors. The third one makes the final classification decision from the output of the cascade. A block diagram of the proposed method is shown in figure 1.

A. Illumination Correction

We use the illumination correction method proposed by [7]. It consists in a sequence of stages whose principal objective is to reduce the effect of illumination variations, like local shadows and highlights, without destroying the visual elements that are important to subsequent steps of the system [7]. It consists of gamma correction, DoG (Difference of Gaussians) filtering and contrast normalization. In figure 2, we can see a block diagram of this method, and in figure 3 we show examples of images with illumination correction.

In the remainder of this subsection we describe the illumination correction subsystem. It starts with the gamma correction. It is a nonlinear transformation whose objective is to enhance the dynamic range of the image in dark regions while compressing it in bright regions. For an image $I(x, y)$, the gamma correction is of the form $I^\gamma$ for $\gamma > 0$ or $\log(I)$ for $\gamma = 0$, where $\gamma \in [0, 1]$. In this work we use $\gamma = 0.2$, recommended by [7].

The second stage of the illumination correction is DoG filtering. It can be viewed as a bandpass filter. The objec-
C. IPD - Inner Product Detector

In correlation filtering, the detection is done by performing the cross correlation between the filter and an unknown signal [10]. The advantage of this technique is the tolerance to small variations of the pattern to be detected. Correlation filtering has been widely used to detect objects. In [11] and [12] it was used for human faces detection. The proposed method uses a new detector based on correlation filters, that we call IPD (Inner Product Detector). It inherits the tolerance to small variations from correlation filters, but has advantage to easily incorporate the statistics of the problem in the design of the detector. The IPD is described in remainder of this subsection.

Suppose an $N$ classes problem, whose classes are: $\{A_1, \cdots, A_n, \cdots, A_N\}$. We need to detect samples that belong to $A_n$ and reject all others. We want a detector $h_{A_n}$ whose inner product with an unknown signal $x$ has a large
value if \( x \in A_n \) and a small one otherwise. Mathematically, we can write
\[
h_{A_n} x = c, \tag{6}
\]
where ideally \( c = 1 \) if \( x \in A_n \) and \( c = 0 \) otherwise. Defining the classification squared error as
\[
\| e \|^2 = (h_{A_n}^t x - c) (h_{A_n}^t x - c)^2, \tag{7}
\]
the Least Squares solution is
\[
h_{A_n} = (E[xx^t])^{-1} E[xc]. \tag{8}
\]
We can write the terms \( E[xx^t] \) and \( E[xc] \) as functions of the training set moments as follows:
\[
E[xx^t] = \sum_{i=1}^{N} p(A_i) R_{A_i}, \tag{9}
\]
\[
E[xc] = p(A_n) \mu_{A_n}, \tag{10}
\]
where \( R_{A_i} \) is the autocorrelation matrix of the training samples from \( A_i \), \( \mu_{A_i} \) is the respective mean and \( p(A_i) \) is the probability of a sample being from \( A_i \). Replacing equations (9) and (10) in equation (8) we have:
\[
h_{A_n} = \left( \sum_{i=1}^{N} p(A_i) R_{A_i} \right)^{-1} p(A_n) \mu_{A_n}. \tag{11}
\]

In the above equation, the expression \( \sum p(A_i) R_{A_i} \) is a weighted sum of the correlation matrices of all classes. Again, in this case, the weights are the probabilities of the classes. Note that this term should be invertible. This implies that the number of different samples should be greater than the size of the vectors. This is particularly convenient in the case of features with a small number of dimensions, as is the case of landmarks. In the expression \( p(A_i) \mu_{A_i} \), the mean of the desired class is weighted by its probability.

D. Normalization Scheme

The inner product between two real vectors are real valued and equal to the projection of a vector on the direction of the other. Due to this, is natural that the inner product between the detector \( h_{A_n} \) and a sample \( x \) may fall out of the range \([0, 1]\). In addition, for the inner product of \( h_{A_n} \), with \( y \notin A_n \), if \( y \) is large enough, can take values greater than the inner product with \( x \in A_n \), which may lead to errors.

In this work we want to detect blocks of \((21 \times 21)\) pixels whose central pixel coincides with a manually annotated landmark and to reject all other blocks whose central pixel can be placed inside the elliptical search region. For training (and testing) we use vectors obtained by concatenating the columns of these blocks. In order to normalize the output of the IPD we add an extra dimension in sample coordinates so that all samples lie on the same hypersphere. Therefore, the vectors have dimensions \( 442 \times 1 \), where \( 441 \) are from the columns of the block and the last is added so that all vectors have the same norm.

The extra dimension is determined as follows. Suppose that the training samples are \( d \)-dimensional. At first, the largest norm vector is searched in the whole training set:
\[
E_{\text{max}} = 1.2 \max \{ \| x_i \|^2 \}, \quad i = \{1, \ldots, L\}, \tag{12}
\]
where \( L \) is the size of the training set. Then, an extra dimension added to each vector:
\[
\bar{x}_i = \left[ x_1 \quad \cdots \quad x_d \quad \sqrt{E_{\text{max}} - \| x_i \|^2} \right]^t. \tag{13}
\]

If we scale \( h_{A_n} \) to unit norm, and divide the inner product in the extended space by \( \sqrt{E_{\text{max}}} \), the output is the cosine of the angle between the detector \( h_{A_n} \) and the unknown sample. Its dynamic range is \([-1, 1]\). This makes the correlation between the detector and a sample \( y \) not to depend on its the norm. Therefore, the correlation of \( h_{A_n} \) with a sample from \( A_n \) tends to be greater than the one with other classes.

E. Cascade

The core of the proposed method is a cascade of IPD detectors. In the first stage of the cascade all the training samples inside the search region are used to design the detector (or to test, if the detectors were already designed). In subsequent stages only samples classified as positive by the previous stage are used. This way we can reject a large number of negative samples in the first stages while the last ones are concentrated in classifying the most difficult samples. The rejected samples in each stage are automatically labeled as negatives and only the samples that pass through all stages are labeled as positives. A diagram of the cascade is given in figure 4.

F. Post-Processing

Usually, the output of the cascade is not a single point. However, the output points tends to be grouped in small regions around the desired landmark. Due to this, in order to provide a single output, we use a simple post processing scheme. The automatic label is the average of all output points of the cascade.

III. Experiments and Results

In order to evaluate the proposed method we use 11 points as illustrated in figure 5, taken from 505 images from the BioID database [13]. Although this database is composed by 1521 gray level images at \((384 \times 286)\) resolution, we use only the frontal face images whose individuals do not wear glasses and do not have mustaches or beards. The training and test sets were built using k-fold cross-validation with 7 folds [14]. Therefore, we partition the database into 7 equally
sized subsets and perform 7 experiments. In experiment \( n \), subset \( n \) is used as the testing set and the remaining ones as the training sets, for \( n = \{1, 2, \ldots, 7\} \).

We assess our method by comparing it to a linear SVM-based system [6]. The SVM-based system differs from the proposed one in two points: the classifier inside the cascade and the post-processing method. The best result of the SVM-based method occurs when we merge the output points of the cascade that are less than 5 pixels apart and then take the most likely response (the one with the smallest Mahalanobis distance (see subsection II-B)).

The number of cascade stages for each fiducial point was empirically determined. The criterion used was add stages until the false positive rate (before post-processing) stabilizes and the false negative rate does not increase significantly.

A. Performance Evaluation

To assess the system performance we use the distance between the automatic and manual annotations. To standardize the measure, we express this distance as a percentage of the intra-ocular distance of the rescaled images. In more precise terms, supposing that \( p_l \) and \( p_r \) are the manually annotated positions of the left and right pupils, the adopted metric error \( d_l \) is

\[
d_l = \frac{|b_m - b_a|}{|p_r - p_l|},
\]

where \( b_m \) is the manual label and \( b_a \) is the automatic one.

All the curves plotted in next section depict hit rate versus the distance between the manual and the automatic labels as a percentage of the intra-ocular distance.

For the automatic labels, we consider less than 10% of the intra-ocular as acceptable.

B. Simulation Results

The simulation results are shown in figures 6, 7, 8, 9, 10 and 11. Due to space restrictions we only show the results for six points. Note that, since the face is symmetric, we show only results for the points on the right side of the face. The curves in the graphics give the average cumulative distribution over all folds. The best results of the IPD and SVM are for the center of the eye and the wing of the nose (points 2 and 7). The IPD outperforms SVM for three points (1, 2 and 8). For the point 7, the two methods have close results. The SVM outperforms IPD for points 3 and 10. Both methods do not work well in mouth points; the great variability of these points can be the reason for such atypical behavior.

Fig. 5. Landmarks used in this work.

Fig. 6. Cumulative distribution of the error measure to the outer corner of the right eye.

Fig. 7. Cumulative distribution of the error measure to the center of right pupil.

Fig. 8. Cumulative distribution of the error measure to the inner corner of right eye.
Fig. 9. Cumulative distribution of the error measure to the right wing of nose.

Fig. 10. Cumulative distribution of the error measure to the tip of nose.

Fig. 11. Cumulative distribution of the error measure to the right mouth corner.

IV. CONCLUSION

We have presented a novel approach to detect facial landmarks in frontal face images. The method consists of three basic steps: pre-processing, classification and post-processing. The pre-processing stage encompasses face detection, illumination correction and search region reduction. The second is a cascade of classifiers at which each individual classifier is designed based on the output of the previous one. The last stage is the final decision based on the output of the cascade. To evaluate our method, we use the BioID database and compare it to an SVM-based system.

The proposed method has competitive performance to SVM. This indicates that the introduced paradigm is worth pursuing. Future works include the evaluation of our method using other databases.

ACKNOWLEDGMENT

This research was supported by CAPES/CNPq and FAPEAM agencies.

REFERENCES