

A landmark paper in face recognition

G.M. Beumer, Q. Tao, A.M. Bazen, and R.N.J. Veldhuis
University of Twente, EEMSC, Signals and Systems
P.O. box 217, 7500 AE, Enschede, The Netherlands
g.m.beumer@utwente.nl

Abstract

Good registration (alignment to a reference) is essential for accurate face recognition. The effects of the number of landmarks on the mean localization error and the recognition performance are studied. Two landmarking methods are explored and compared for that purpose: (1) the Most Likely-Landmark Locator (MLLL), based on maximizing the likelihood ratio [2], and (2) Viola-Jones detection [14]. Both use the locations of facial features (eyes, nose, mouth, etc) as landmarks. Further, a landmark-correction method (BILBO) based on projection into a subspace is introduced.

The MLLL has been trained for locating 17 landmarks and the Viola-Jones method for 5. The mean localization errors and effects on the verification performance have been measured. It was found that on the eyes, the Viola-Jones detector is about 1% of the inter-ocular distance more accurate than the MLLL-BILBO combination. On the nose and mouth, the MLLL-BILBO combination is about 0.5% of the inter-ocular distance more accurate than the Viola-Jones detector. Using more landmarks will result in lower equal-error rates, even when the landmarking is not so accurate. If the same landmarks are used, the most accurate landmarking method will give the best verification performance.

Keywords: *face registration, face recognition, landmarking, likelihood ratio, Viola-Jones, landmark correction*

1. Introduction

Riopka et al. [10], Cristinacce et al. [6] and Beumer et al. [4] have shown that precise landmarks are essential for a good face-recognition performance. Cristinacce [5] investigated landmark locators based on correlation, orientation maps and Viola-Jones detection [14].

In this paper¹ we propose an improvement on earlier work by Bazen et al. [2] and a Viola-Jones based landmark finder. Both will be compared to each other and to groundtruth data. Their performances will be quantified by the RMS value of the error with respect to the groundtruth data. The equal-error rates (EERs) measured in a verification experiment measured by will be presented.

2. Landmark detection

The first step in face recognition is to locate the face in the image. In the methods proposed here, this is done by a Viola-Jones detector [14], obtained from the OpenCV library [8]. It is assumed that there is only one face per image. When it is found, a region of interest (ROI) is selected for each landmark. In this ROI we search for the landmarks using one of the two algorithms explained in the following two subsections.

2.1. Most Likely Landmark Location

MLLL treats landmark finding as a two-class classification problem: a location in an image is either the landmark or it is not. The texture values in a region surrounding a landmark are the features for the classification. For each location in the ROI the likelihood ratio -for that location to be the landmark- is calculated. The most likely location, i.e. the one with the highest score, is taken to be the landmark. The 17 landmarks that the MLLL searches for are shown in Figure 1.

Outliers due to errors by the MLLL, can sometimes be corrected by a shape correction. Both the landmark detection and the shape correction are discussed below.

Likelihood-ratio-based landmark finder The MLLL calculates a similarity score, derived from the

¹The authors wish to emphasize that the title merely intends to express that the paper is about landmarking.

likelihood ratio for a landmark at each position in the ROI. The gray-level intensities in the neighbourhood of a candidate location (u, v) are arranged into a vector $x_{u,v}$. The likelihood that $x_{u,v}$ is the neighbourhood of the landmark is expressed by the likelihood ratio

$$L_{u,v} = \frac{p(x_{u,v}|L)}{p(x_{u,v}|\bar{L})}, \quad (1)$$

with $p(x_{u,v}|L)$ the probability density function of $x_{u,v}$, given that it is the neighbourhood of a landmark and $p(x_{u,v}|\bar{L})$ the probability density function of $x_{u,v}$, given that this is not the case. The location of the landmark is chosen at the point (u, v) in the ROI at which $L_{u,v}$ is maximum. It is assumed that the probability density functions of $x_{u,v}$ in (1) are normal. Therefore, rather than the likelihood ratio an equivalent similarity score

$$S_{u,v} = -(y_{u,v} - \mu_L)^T \Sigma_L^{-1} (y_{u,v} - \mu_L) + (y_{u,v} - \mu_{\bar{L}})^T \Sigma_{\bar{L}}^{-1} (y_{u,v} - \mu_{\bar{L}}) \quad (2)$$

is calculated. In (2), $y_{u,v} = T(x_{u,v} - x_{0,u,v})$, with T a transformation matrix reducing the dimensionality of $x_{u,v}$ to manageable proportions [13], and $x_{0,u,v}$ the global sample mean. The matrices Σ_L and $\Sigma_{\bar{L}}$ are the covariance matrices for the landmark and non-landmark templates, respectively. The landmark and non-landmark means are denoted by μ_L and $\mu_{\bar{L}}$, respectively.

The matrix T , the covariance matrices, the global sample mean, and the averages are all obtained from training. The transformation matrix T is chosen such that Σ_L is the identity matrix and $\Sigma_{\bar{L}}$ is diagonal. It reduces the dimensionality while trying to optimize the discriminability between the landmark and non-landmark distributions. The method applied is known as Approximate Maximum Discrimination Analysis [1].

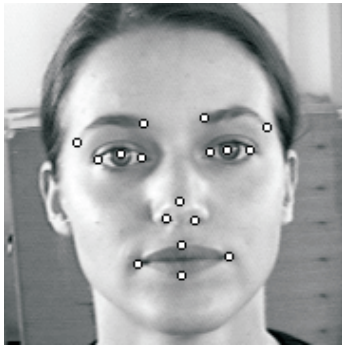


Figure 1. Landmarks detected by the MLL.

The location of the landmark is chosen at the point (u, v) in the ROI at which $S_{u,v}$ is maximum. In this way all 17 landmarks are located.

Shape correction Sometimes landmarks locations are incorrect. The aim of shape correction is to detect and correct these errors. A shape is the collection of the coordinates of a set of landmarks, arranged into a vector. Correct shapes are assumed to be in a subspace of \mathbb{R}^{2d} with d the number of landmarks, here $d = 17$. Incorrect shapes, containing one or more erroneous landmarks, are assumed to be outside this subspace. A basis $(u_1 \dots u_n) = U$, with $n < 2d$ of this subspace is determined by means of principal component analysis (PCA) applied to a training set of correct shapes. A shape x is projected *there and back again* (BILBO), resulting in a modified shape $x' = UU^T x$. coordinate space. The effect of this is that all landmarks will have changed: the correct ones only slightly, but the erroneous ones significantly and (hopefully) in the direction of the correct location. Therefore, the landmarks of which the location changed significantly using BILBO, are considered to be wrong, and their new location is taken as a better one. This procedure is repeated until convergence has been reached. This is usually after 5 iterations.

BILBO is trained on a set of shapes, taken from the groundtruth data, arranged as the columns of a matrix X . The training consists of the following steps:

1. Register all shapes in X to the average shape used for registration.
2. Apply -limited and random- rotation, translation and scaling to all shapes in X in order to model variations encountered in the images. The rotation angle, translation vector and scaling factor have normal zero-mean distributions. The translation has a standard deviation of 5 pixels. The scaling has a standard deviation of 5%. The rotation has a standard deviation of 3 degrees.
3. Perform a singular value decomposition $X = USV^T$.
4. Reduce the dimensionality of the subspace by taking only the first $n < 2d$ columns of U .

To correct a shape the following algorithm is used:

1. Estimate the shape after transformation, $x' = UU^T x$.
2. Determine the Euclidian distance D_i per landmark between x and x' .

- Determine the threshold

$$\tau = RC \frac{1}{d} \sum_{i=1}^d D_i, \quad (3)$$

with C a constant and R the run number.

- Replace the coordinates of which $D_i > \tau$ in x by the corresponding coordinates in x' : $x_i = x'_i$.
- Repeat steps 1 to 4. Once for a landmark $D_i < \tau$ stop updating it until no $D_i > \tau$ for all i .
- Repeat step 1 to 5 changing all coordinates until $R = 5$.
- Transform the coordinates back to the original scale.

Figure 2 shows an example of the result of landmark correction. The circles are the raw landmarks and the triangles their new locations after BILBO. In this example $n = 3$ and the constant C in (3) is chosen as $C = 1.5$.

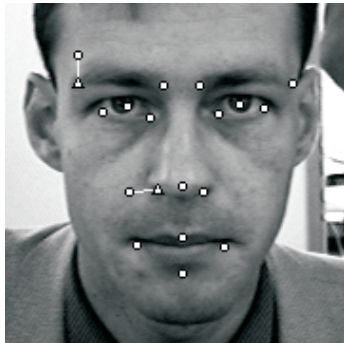


Figure 2. Original and corrected landmarks (triangles).

2.2. Viola-Jones based landmark localization

The second method for landmark localization is the Viola-Jones detector [14], which uses a combination of Haar-like features to represent the texture information in an image. A detailed description of this method and the training method -Adaboost- can be found in [14].

We developed detectors for 5 landmarks: two eyes (size 28×14), one nose (size 28×14), and two mouth corners (size 20×20). Only 5 of the 17 landmarks

have been chosen for the Viola-Jones based method, because the other landmarks did not result in fast and compact cascades for detection. For simplicity, the face region is first detected as a ROI for the localization landmarks in face.

Figure 3 shows the results of applying the Viola-Jones method for localizing the face and the landmarks. This method does a multi-scale search and chooses the

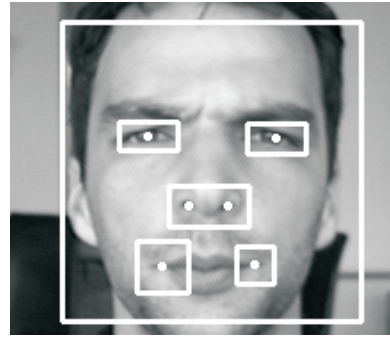


Figure 3. The landmarking result by Viola-Jones method

facial landmark candidates through thresholding [14]. There is the possibility of multiple candidates (multi-size and/or different position), or a missing candidate for one facial landmark. For the multi-size case, we choose the candidate with the largest size, because smaller-size candidates seem to be less reliable. For candidates of the same size but with difference positions, we calculate their relative locations in reference to the face ROI, and choose the one nearest to its average position. In case of missing landmarks, we use fewer landmarks for registration. It is also possible to restore the missing landmarks according to the statistics of their geometrical distribution, e.g. [12].

The authors have tried to develop Viola-Jones detectors for all the landmarks obtained by the MLLL method, but found that Adaboost training of landmarks other than the eyes, the nose and the mouth corners failed to converge.

3. Experiments and results

The two landmarking methods have been evaluated in two ways. The RMS errors with respect to groundtruth data have been computed and the landmarks found have been used for registration in a face-verification experiment. The EERs measured in this experiment serve as a benchmark for the quality of the landmarking.

The FRGC database [9] was used to compute the RMS errors. It was also used to train and test the verification system. The FRGC database consists 5658 images. Two third, 3772, of the database are high-quality images with a low variety in pose, lighting and scale and having around 300 pixels inter eye distance. For our tests we used only the high-quality part of FRGC database images.

3.1. Training

The BioID database [7] was used to train both landmark-detection methods as well as the shape correction. It consists of 1521 images, which vary in pose, scale and lighting conditions, but which are mainly frontal. All images have been manually landmarked.

MLLL The positive templates were all selected using hand-labeled groundtruth data. The negative samples were all taken around the landmark at a minimal distance of half the size of the template. All templates are either 40x40 pixels or 60x40 pixels in size and have a zero mean and a standard deviation equal to one.

Viola-Jones The positive training samples are obtained from all 1,521 images in the BioID groundtruth data. The negative training samples are randomly chosen from samples that do not contain landmarks. The code of the Adaboost training was taken from the Intel OpenCV library [14]. In our work each detector has been trained with 3,000 positive and 6,000 negative samples.

Shape correction An image in the BioID database has 20 landmarks 17 of which we used for training: the eyes, inner and outer eye corners, eyebrow ends, mouth corners, upper lip, lower lip, nostrils and the nose. In Figure 1 these landmarks are shown. All 1521 sets of 17 landmarks were used.

3.2. Landmarking accuracy

In order to be able to evaluate the landmarking methods a well-defined error measure is required. Since the images in the databases are of various scales, a straightforward root mean square (RMS) error could be used. In order to calculate a meaningful measure a simple method was used:

1. Translate, scale and rotate the groundtruth data so that the eye landmarks are on a horizontal line at a 100-pixels distance.

2. Register the shape found to the corresponding groundtruth shape.
3. Calculate the Euclidian distance between each landmark and its groundtruth equivalent.
4. Remove the bias caused by the different labeling policies in the databases, i.e. tip of the nose (BioID) versus a point between the nostrils (FRGC).
5. Calculate the RMS value of the remaining difference between the found shape and the groundtruth shape. This is now a percentage of the inter-ocular distance.

In the FRGC database the center of the mouth is labeled, whereas our methods label the mouth corners. Therefore, prior to calculating the error an estimate of the center of the mouth was obtained by computing the midpoint of the mouth corners found.

Results The average errors are presented in Table 1. Application of BILBO for shape correction improves the MLLL results significantly. On the eyes, the Viola-Jones detector is about 1% of the inter-ocular distance more accurate than the MLLL-BILBO combination. On the nose and mouth, it is the MLLL-BILBO combination that is about 0.5% of the inter-ocular distance more accurate than the Viola-Jones detector.

FRGC	right eye	left eye	nose	mouth
Viola-Jones	3.2	3.3	6.3	4.1
MLLL	6.7	7.2	13.0	7.3
MLLL+BILBO	4.2	4.6	5.8	3.7

Table 1. RMS errors as a percentage of the inter-ocular distance, obtained on the FRGC database.

In Figure 4 the cumulative error distribution is shown as function of the RMS pixel error. This is the percentage of the landmarks with an RMS error less than the RMS pixel error. For the nose and the mouth the lines for the MLLL+BILBO and the Viola-Jones methods cross over. This means that the Viola-Jones method makes more large errors than the MLLL-BILBO combination. It is, again, clear that BILBO under all the conditions evaluated increases the performance of the MLLL algorithm significantly. Since BILBO looks at shape deviations from a statistical model -the subspace- it corrects the larger errors. This suggests that the Viola-Jones method could also benefit from a shape-correction method. Because of the smaller number of landmarks, 5 instead of 17, this cannot be a subspace method, but

rather a restoration method based on minimizing a expected restoration error. e.g. [12].

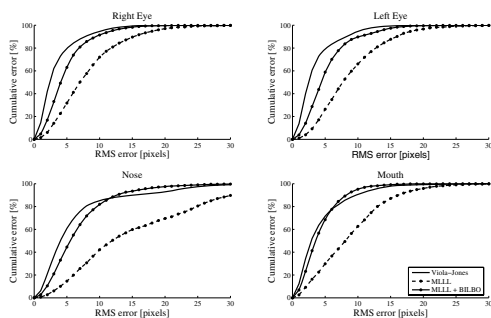


Figure 4. The cumulative error distribution function of the RMS pixel error.

3.3. Impact on verification

In this section, the effect of the to landmarking methods on the performance of face verification is discussed. The images in the FRGC database are registered based on the landmarks obtained by the methods discussed here. The experiment consisted of ten trials. In each trail the FRGC data base was randomly split in a training and a testing set, each containing 50% of the data, or 1886 images. In each trail genuine and impostor matching scores were collected. The final EER was computed from the total sets of genuine and impostor matching scores [4].

In order to investigate the impact of the number of landmarks on the registration we also ran the experiment using the MLLL method with only the 5 landmarks that are found by the Viola-Jones method.

The verification method used employs a standard combination of PCA and LDA for the reduction of the dimensionality of the image data, followed by a likelihood-ratio classifier. It is described in, for instance, [13] for hand-geometry recognition. It has not been optimized and parameter tuning might improve its performance. However, this is not necessary in order to compare results.

After registration, the images containing the faces are cropped to 256 pixels high and 256 pixels wide. The centres of the eyes in the reference image are at (78,101) and (178,101). The upper left corner is at (1,1). From this image, a fixed region of interest (ROI), containing most of the face, is taken. All grayscale values in the ROI are arranged into a feature vector x . The ROI is

visualised in Figure 5. In each trial these images are randomly split into a training set and testing set.



Figure 5. Region of Interest.

Training/Enrollment The training [13] is done using a combination of the Eigenfaces [11] and Fisherfaces [3] methods:

- First apply PCA on the training data after subtracting the mean. After a subsequent dimension reduction, the number of features is twice the number of classes.
- Then apply a linear discriminant analysis (LDA), making the total covariance matrix, Σ_T , unity. After a subsequent reduction, the number of features is the number of classes minus one. Store the within class covariance matrix, Σ_W , total average, μ_T , and the transformation matrix, T .

During the enrollment phase, the class averages, $\mu_{W,i}$, are stored as templates.

Testing In the testing phase, a feature vector x , containing all pixel values in the ROI, is projected onto the reduced feature space by premultiplying it with the transformation matrix, i.e. $y = T(x - \mu_T)$. The extracted feature vector, y , is then compared to class i by calculating a log likelihood based matching score S :

$$S_{y,i} = -(y - \mu_{W,i})^T \Sigma_W^{-1} (y - \mu_{W,i}) + y^T \Sigma_T^{-1} y - \log |\Sigma_W| + \log |\Sigma_T| \quad (4)$$

Results The results are shown in Table 2. Face verification based on groundtruth data gives good results (EER = 0.45%). The proposed methods result in higher EERs between 3.6% and 6.1%. Therefore, face recognition using automatic landmarking methods seems still not as good as facerecognition using handlabeled landmarks.

Using more landmarks results in lower EERs, even if the landmarking is not so accurate: MLLL with 17

FRGC	EER [%]	std(EER) [%]
Ground truth data	0.45	0.03
Viola-Jones	4.9	0.1
MLLL	4.0	0.1
MLLL+BILBO 17 landmarks	3.6	0.1
MLLL+BILBO 5 landmarks	6.1	0.1

Table 2. Results of the verification experiment

landmarks gives a better verification performance than Viola-Jones with only 5, even though the RMS error of MLLL is much higher, cf. Table 1. The fact that BILBO does improve the results obtained by MLLL, but not as significantly as it improves the registration error shown in Table 1, seems to confirm this. If the same landmarks are used, the most accurate landmarking method will give the best verification performance.

4. Conclusions

Two landmarking methods for face registration have been proposed: MLLL, which is likelihood-ratio based and a method based on Viola-Jones detection. The version of MLLL that was presented is capable of detecting 17 landmarks. The Viola-Jones detector detects 5 landmarks. There are some indications that it could be problematic to devise a Viola-Jones method for a much larger number of landmarks. The MLLL method can be enhanced with a shape-correction method, BILBO, that can substantially improve its accuracy.

The accuracy of both methods has been investigated as well as their impact on the face-verification performance. It was found that application of BILBO for shape correction improves the MLLL results significantly. On the eyes, the Viola-Jones detector is about 1% of the inter-ocular distance more accurate than the MLLL-BILBO combination. On the nose and mouth, the MLLL-BILBO combination is about 0.5% of the inter-ocular distance more accurate than the Viola-Jones detector. It was also found that using more landmarks results in lower EERs, even when the landmarking is not so accurate. On the same landmarks, the most accurate landmarking method will give the best verification performance.

It seems that good face recognition relies on accurate registration and that (1) accurate (rigid) registration can be achieved by accurate landmarks and (2) can be improved by increasing the number of landmarks.

5. Acknowledgements

The work presented here was done in the contexts of the IOP-GenCom project BASIS and the Freeband-BISIK project PNP2008.

References

- [1] A. Bazen and R. Veldhuis. Detection of cores in fingerprints with improved dimension reduction. In *Proc. SPS 2004*, pages 41–44, Hilvarenbeek, The Netherlands, apr 2004.
- [2] A. Bazen, R. Veldhuis, and G. Croonen. Likelihood ratio-based detection of facial features. In *Proc. ProRISC 2003, 14th Annual Workshop on Circuits, Systems and Signal Processing*, pages 323–329, Veldhoven, The Netherlands, nov 2003.
- [3] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *ECCV 2*, 1996.
- [4] G. Beumer, A. Bazen, and R. Veldhuis. On the accuracy of eers in face recognition and the importance of reliable registration. In *SPS 2005. IEEE Benelux/DSP Valley*, April 2005.
- [5] D. Cristinacce and T. Cootes. A comparison of shape constrained facial feature detectors. In *6th International Conference on Automatic Face and Gesture Recognition 2004, Seoul, Korea*, pages 375–380, 2004.
- [6] D. Cristinacce, T. Cootes, and I. Scott. A multi-stage approach to facial feature detection. In *15th British Machine Vision Conference, London, England*, pages 277–286, 2004.
- [7] HumanScan. Bioid face db. <http://www.humanscan.de/>
- [8] Intel. Open computer vision library. <http://sourceforge.net/projects/opencvlibrary/>
- [9] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005*.
- [10] T. Riopka and T. Boulton. The eyes have it. In *Proceedings of ACM SIGMM Multimedia Biometrics Methods and Applications Workshop.*, pages 9–16, Berkeley, CA, 2003.
- [11] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, pages 71–86, 1991.
- [12] R. Veldhuis. *Restoration of Lost Samples in Digital Signals*. Prentice Hall, New York, 1990.
- [13] R. Veldhuis, A. Bazen, W. Booi, and A. Hendrikse. Hand-geometry recognition based on contour parameters. In *Proceedings of SPIE Biometric Technology for Human Identification II*, pages 344–353, Orlando, FL, USA, March 2005.
- [14] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002.