Impact of Eye Detection Error on Face Recognition Performance

Abhishek Dutta[a], Manuel Günther[b], Laurent El Shafey[b],
Sébastien Marcel[b], Raymond Veldhuis[a] and Luuk Spreeuwers[a]

[a]University of Twente, Faculty of EEMCS, P.O.-Box 217, 7500 AE Enschede, Netherlands.
{a.dutta,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl
[b]Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, CH - 1920, Martigny, Switzerland. {manuel.guenther,laurent.el-shafey,marcel}@idiap.ch

Abstract

The locations of the eyes are the most commonly used features to perform face normalization (i. e., alignment of facial features), which is an essential preprocessing stage of many face recognition systems. In this paper, we study the sensitivity of open source implementations of five face recognition algorithms to misalignment caused by eye localization errors. We investigate the ambiguity in location of the eyes by comparing the difference between two independent manual eye annotations. We also study the error characteristics of automatic eye detectors present in two commercial face recognition systems. Furthermore, we explore the impact of using different eye detectors for training/enrollment and query phases of a face recognition system. These experiments provide an insight into the influence of eye localization errors on the performance of face recognition systems and recommend a strategy for designing training a test set of a face recognition algorihtm.

1 Introduction

The normalization of a facial image for scale and rotation is a common preprocessing stage of many face recognition systems. This preprocessing stage, often called geometric normalization of the face, ensures that facial features like nose or eyes occupy similar spatial positions in all images. The locations of at least two facial landmarks are required to normalize a facial image for translation, scale and in-plane rotation. Most commercial and open source face recognition systems use the centers of the eyes as landmarks for face normalization because the inter-ocular distance can be used to correct scale, while the orientation of the line between the eyes allows correction of in-plane rotation [15], as indicated in Figure 1.

A face normalization scheme based on the centers of the eyes is known to contribute to decrease face recognition performance if supplied with inaccurate eye locations [13, 16, 15, 22]. So far, this observation was based on results of a small number of face recognition

This paper is a postprint of a paper submitted to and accepted for publication in IET Biometrics and is subject to Institution of Engineering and Technology Copyright. The copy of record is available at IET Digital Library.

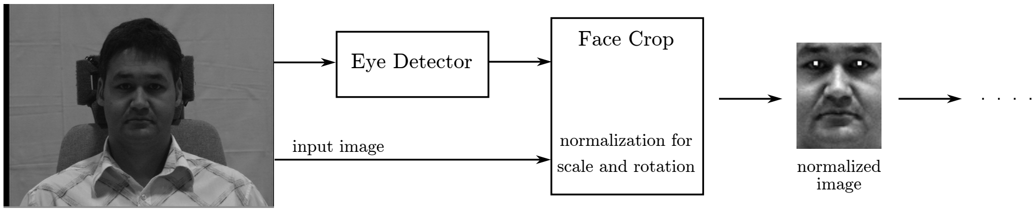


Figure 1: Basic facial image preprocessing pipeline used by most face recognition systems.

system operating on a limited facial image data set. However, face recognition systems have different tolerances to misalignment of facial features. Therefore, in this paper, we study the impact of misalignment that is caused by eye localization errors on the performance of the following five open source face recognition systems, which are spanning a wide range of popular and state-of-the-art systems: Eigenfaces [18], Fisherfaces [2], Gabor-Jet [7], Local Gabor Binary Pattern Histogram Sequence (LGBPHS) [25] and Inter-Session Variability modeling (ISV) [20]. All methods are evaluated using two evaluation metrics, the (unbiased) Half Total Error Rate (HTER) and the (biased) Area Under Curve (AUC).

It is common practice to consider manually located eye coordinates as ground truth when assessing the accuracy of automatically detected eye coordinates. We investigate the merit of this practice by analyzing the difference in manual eye annotation performed by two independent institutions. We also analyze the accuracy of automatic eye detectors present in two commercial face recognition systems. Based on a novel experiment methodology, we confirm that "eye locations are essentially ambiguous" [17]. Furthermore, we investigate the impact of using different types of annotations (manual, automatic) for the training/enrollment and the query phases on face recognition.

Our experiments provide an insight into the limits of geometric face normalization schemes based on eye coordinates. Our aim is to motivate face recognition system designers to build algorithms that are robust to a minimum amount of misalignment that is unavoidable due to the ambiguity in eye locations. The major contributions of this paper are:

1. We reveal the sensitivity of five open source face recognition systems towards eye localization errors causing various types of misalignment as translation, rotation and scaling of the face. We also show that unbiased and biased evaluation techniques – HTER and AUC – have different characteristics in evaluating query images with inaccurate eye locations.

2. We explore the inherent limitations of facial image alignment schemes based on eye coordinates. To the best of our knowledge, this is the first study to analyze the difference between two independent manual eye annotations carried on same set of images. Our study shows an ambiguity of four pixels (with an average inter-ocular distance of 70 pixels) in both horizontal and vertical location of manual eyes center annotations in frontal facial images.

3. We show that the automatic eye detection system included in a commercial face recognition system achieves the eye detection accuracy of manual eye annotators. Furthermore, we show that such a fairly accurate automatic eye detector can help to achieve

recognition performance comparable to manually annotated eyes, given that the same automatic detector is used for annotating all training, enrollment and query images.

4. Our work lays the foundation for reproducible research in this avenue of research by allowing the extension of our experiments with additional face recognition systems and other facial image data set. The results of all experiments presented in this paper can be reproduced by using open source software[1] including a detailed description on how to regenerate the plot and tables. Such an open platform encourages researchers to pursue similar studies with other face recognition systems or image databases. We also aim to achieve reproducible research to allow the research community to reproduce our work on public databases and open source software hence excluding commercial systems that are not available to the large majority of researchers. With our framework other researchers can test or proof their claimed stability against eye localization errors by simply re-running the experiments presented in this study using their algorithms.

This paper is organized as follows: We review related studies about face recognition with eye localization errors in Section 2. In Section 3, we describe our methodology to study the impact of misalignment on face recognition performance and to assess the accuracy of manual and automatic eye detectors. Sections 4.1 and 4.2 describe the experiments designed to study the influence of misalignment (translation, rotation and scale) on face recognition performance. In Section 4.3, we analyze the ambiguity in manual and automatic eye annotations. Section 4.4 studies the impact of different sources of eye annotations for training/enrollment and query phases of a face recognition system. In Section 5, we discuss the results from our experiments and, finally, Section 6 concludes the presented work.

2 Related Work

Face normalization is a critical preprocessing stage of most 2D face recognition systems because subsequent processing stages like feature extraction and comparison rely on an accurate spatial alignment of facial features. Therefore, other researchers have investigated the influence of this preprocessing stage on face recognition performance.

The impact of translation, rotation and scaling on the similarity value of an Eigenfaces [18] algorithm was studied in [11]. There, the analysis was limited to a small data set (seven human subjects and one synthetic face) and a single face recognition algorithm. Moreover, only the variation in the raw similarity score was studied, and not the actual face recognition performance.

The authors of [22] compare the performance of one Eigenfaces based and one commercial system for automatically detected and manually labeled eye annotations. They also compare the accuracy of automatic eye detection under controlled and uncontrolled lighting conditions by considering the manual annotations as ground truth. Their experimental results show that image quality (like illumination) affects face recognition performance in two ways: by contributing to misalignment caused by error in eye detection and by directly influencing a classifiers ability to extract facial features from misaligned images. We do not include image quality variations in our study, but we perform experiments on a larger number of face recognition systems and use two automatic eye detectors.

[1] The scripts to reproduce the experiments that are described in this paper can be downloaded from `http://pypi.python.org/pypi/xfacereclib.paper.IET2015`

In [13], the authors perturb initial eye locations to generate multiple face extractions from a single query image and select a match using nearest neighbor scheme. At the expense of an increased computational cost, they show that this approach consistently outperforms face recognition systems based on both manually located and automatically detected eye coordinates. Hence, manual eye annotations do not always guarantee optimal face recognition performance.

The authors of [17] show that the performance of a Fisherfaces [2] based face recognition system drops drastically already under small misalignment. They argue that "eye locations are essentially ambiguous" and even the manual eye annotations have large variance. Therefore, they put forward a case to explicitly model the effects of misalignment during the training of the classifier. To make a Fisherfaces based system robust to small misalignment, they deliberately introduce misaligned samples corresponding to each facial image during training. They show that such a system has a higher tolerance towards misaligned query images.

The sensitivity of Eigenfaces, Elastic Bunch Graph Matching (EBGM) and a commercial system's face recognition performance to misalignment was investigated by [15]. They systematically perturb manual eye annotations and report performance variation for three face recognition systems. They find that misalignment caused by scaling have a higher influence on face recognition performance than rotation or translation.

The authors of [23] propose an automatic eye detection system and report that their automatically detected eye coordinates achieve face recognition performance comparable to manually annotated eye coordinates on the FRGC 1.0 data set. However, it is not clear which eye coordinates, manual or automatically detected, were used for training the face recognition systems. Furthermore, their analysis is only limited to Eigenfaces and Fisherfaces baseline face recognition systems. In this paper, we report performance variations for five face recognition systems, which span a wide range from classic systems to state-of-the-art systems of different kinds.

Based on the analysis of similarity scores, the authors of [24] build a model to predict recognition performance for a set of probe images. They use this model to predict the performance corresponding to probe sets built from different perturbed eye coordinates. This allows them to adjust the image alignment in order to attain best possible recognition performance. They conclude that manually located eye coordinates do not necessarily provide the best alignment and that performance prediction systems can be used to select the best alignment that can outperform systems based on manually located eye coordinates. However, their analysis is limited to a single Eigenfaces based face recognition system. Furthermore, it is not clear which eye coordinates (manual or automatic) they use for training.

Some face recognition systems are more tolerant to misalignment. Therefore, the authors of [16] argue that performance of a face localization system should be evaluated with respect to the final application, e. g., face verification. They build a model that directly relates face localization errors to verification errors. It is possible to build such a model using the data reported in this paper. However, in this paper, we only aim to compare the tolerance of different face recognition systems towards misalignment. Therefore, we do not build such a model and only report verification errors parametrized for the following two types of localization errors: (a) errors involving only translation and rotation (without scaling) and (b) errors belonging to a normal distribution of landmark locations with variable variances.

More recently, there have been efforts to build face recognition algorithms that are more robust to misalignment. In [19], the authors highlight the importance of image alignment for correct recognition and propose a modification to the sparse representation-based classification algorithm such that it is more robust to misalignment. Similarly, by extracting features from parts of the facial image independently, the authors of [20] achieve natural robustness to misalignment – a claim that is also validated by the experiment results presented in this paper. In the pursuit for robustness against misalignment, several works like [4], [17] and [13] train the classifier on multiple instances of same image cropped using perturbed eye coordinates. At the expense of increased computational cost, they achieve some performance improvement in handling misaligned images.

In this paper, we study the impact of misalignment caused by errors in eye localization on the verification performance of face recognition systems. Similar studies carried out in the past were either limited by the number of face recognition systems or the size of facial the image database. Our study is based on five open source face recognition algorithms using an unbiased verification protocol based on a much larger data set containing images of 272 subjects (208 in training and 64 in development set). Furthermore, our experiments are solely based on open source software and our results are reproducible and extensible to include more algorithms and facial image databases in the future.

3 Methodology

We want to study the influence of eye annotation error on face recognition performance. Therefore, we keep all other parameters (like training and enrollment images, algorithm configuration, etc.) fixed and only perturb the manual eye annotations of the query images. This allows us to analyze the contribution of misalignment caused by eye annotation error in degrading the performance of a face recognition system.

We study the difference between manual eye annotations performed independently at two institutions. Furthermore, we compare the accuracy of automatic eye detectors by considering manual eye annotations as ground truth. This analysis reveals the ambiguity present in the location of the eyes for both manual annotators and automatic eye detectors.

3.1 Face recognition systems

We evaluate the performance of five open source face recognition algorithms implemented in the `FaceRecLib` [8], which itself is built on top of the open source signal-processing and machine learning toolbox `Bob` [1]. In our experiments, we extend the `FaceRecLib` to allow systematic perturbation of the, otherwise fixed, eye coordinates in the normalized images. The five systems considered are listed and succinctly described below. These algorithms were chosen for the following two reasons: (a) the recognition performance of these algorithms span from baseline performance (Eigenfaces) to state-of-the-art face recognition performance (ISV), and (b) their open source implementation is available in a single stable package called `FaceRecLib`, which allows to reproduce and extend the results presented in this paper. In our experiments we use the stock implementations of the algorithms with their default configuration, we do not adapt any parameter to the database or the experimental setup. For a more detailed description of the algorithms, please refer to the cited papers, or have a look into their implementations.[2]

The first algorithm we investigate is the well-known Eigenfaces [18] algorithm. It uses raw pixel gray values as features by concatenating all pixels of the normalized image to one

[2] The latest stable version of the `FaceRecLib` can be downloaded from `http://pypi.python.org/pypi/facereclib`

vector. It extracts a basis of eigenvectors from a training set of samples using a Principal Component Analysis (PCA). For feature extraction, the dimensionality of samples is reduced by projecting them into this basis. Finally, classification in this lower dimensional space is performed using the Euclidean distance.

Fisherfaces [2] use a similar approach to face recognition. This system first extracts a basis of eigenvectors from a training set of samples using a combination of PCA and Linear Discriminant Analysis (LDA) techniques. As for Eigenfaces, the dimensionality of the samples is reduced by projecting them into this basis, and classification is performed in the exact same way.

A more complex strategy for recognizing faces is given in the Gabor-Jet [7] algorithm. Here, responses of several Gabor wavelets are computed on a set of points located on a grid on the original image. At each point, a descriptor is obtained by concatenating the response values, which is referred to as a Gabor-Jet. The comparison of these Gabor-Jets is then performed using a similarity function that is based on both absolute values and phase values of Gabor wavelet responses. Please note that this algorithm requires no training.

The Local Gabor Binary Pattern Histogram Sequences (LGBPHS) [25] algorithm combines two different feature extraction techniques. It first extracts Gabor filtered images from the original sample, before applying the Local Binary Pattern (LBP) operator on them. Each of this filtered images is decomposed into overlapping blocks, from which local LGBP histograms are gathered. Finally, all local histograms are concatenated to form the LGBPHS, and classification is performed using the histogram intersection measure. As for Gabor-Jet, this algorithm requires no training either.

A completely different approach to face recognition is given by the Inter-Session Variability modeling (ISV) [20]. This generative parts-based algorithm models the distribution of local DCT block features using Gaussian Mixture Models (GMMs), where many DCT features are extracted independently from overlapping image blocks. At training time, a Uniform Background Model (UBM) is learnt from a set of samples from several identities, as well as a subspace that describes the variability caused by different recording conditions (session variability). At enrollment time, for each client a specific GMM is computed by adapting the UBM to the enrollment samples of the client. In particular, ISV enrolls these models by suppressing session-dependent components to yield true session-independent client models. Finally, classification relies on the computation of log-likelihood ratios.

3.2 Image database and evaluation protocol

For all the experiments discussed in this paper, we chose the training and development set according to the protocol[3] `M` of the Multi-PIE data set [6]. The full Multi-PIE data set contains images of 337 subjects captured in four sessions with various variations in pose, illumination and expression.

This unbiased face verification protocol `M` is defined as follows: The training set contains those 208 subjects that do not appear in all four sessions. The development and evaluation sets contain 64 and 65 disjoint subjects, respectively, all of which are not included in the training set. The enrollment images are taken from session 01, while query images stem from sessions $\{02, 03, 04\}$, the pose is fixed to the frontal camera `05_1` with no flash `00`. Hence, the training set consists of 515 images, while the development set contains 1 enrollment image per subject and 256 query images of the same 64 subjects, where all enrolled models

[3] An open source implementation of the protocols for the Multi-PIE database is available at `http://pypi.python.org/pypi/bob.db.multipie`

are compared with all query samples. To keep our investigations unbiased, we report results only for the development set and do not perform any experiment on the evaluation set.

3.3 Performance measures

To evaluate the face verification performance, we use the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). For a given similarity score threshold s_t, these metrics are defined as follows:

$$\mathrm{FAR}(s_t) = \frac{|\{s_{imp}|s_{imp} \geq s_t\}|}{|\{s_{imp}\}|}, \quad \mathrm{FRR}(s_t) = \frac{|\{s_{gen}|s_{gen} < s_t\}|}{|\{s_{gen}\}|}, \tag{1}$$

where s_{gen} and s_{imp} denote genuine (same source comparison) and impostor (different source comparison) scores, respectively. In this paper, we use two evaluation metrics, both of which are based on FAR and FRR. The first metric is the Half Total Error Rate (HTER). Let s_t^* denote the threshold for development set, without any eye perturbations, such that:

$$s_t^* = \arg\min_{s_t} \frac{\mathrm{FAR}(s_t) + \mathrm{FRR}(s_t)}{2}\,, \tag{2}$$

then HTER with perturbed eye locations is defined as:

$$\mathrm{HTER}_{(\theta,t_X,t_Y)} = \frac{\mathrm{FAR}_{(\theta,t_X,t_Y)}(s_t^*) + \mathrm{FRR}_{(\theta,t_X,t_Y)}(s_t^*)}{2}\,, \tag{3}$$

where (θ, t_X, t_Y) are the rotation and translation parameters as defined in Section 4.1. Note that a perfect system has an HTER of 0.0, while the HTER of a random system is 0.5.

The second evaluation metric is the Receiver Operating Characteristics (ROC), where the Correct Acceptance Rate (CAR) with CAR = 1.0 − FRR is plotted against the FAR. Usually, we are interested in the CAR values at low FAR values and, therefore, we plot the FAR axis in logarithmic scale. Additionally, the ROC can be characterized by a single number called the Area Under Curve (AUC), which – as the name implies – measures the region covered by the ROC curve. The AUC can be approximated as:

$$\mathrm{AUC} = \sum_{i=1}^{n-1} (F[i+1] - F[i]) \left(\frac{C[i] + C[i+1]}{2} \right), \tag{4}$$

where $C[i]$ denotes the CAR value corresponding to FAR of $F[i]$ and the n values in F are sorted in ascending order. A perfect verification system has an AUC of 1.0.

Though both measures HTER and AUC are based on the same FAR and FRR, they have different characteristics. The AUC measures performance directly using the perturbed scores, which makes this measure biased. A more realistic and unbiased measure is given by the HTER, which defines a threshold s_t using clean conditions, but measures performance in perturbed conditions.

3.4 Measures of misalignment

The authors of [10] have proposed the Jesorsky measure of eye annotation misalignment, which is defined as follows:

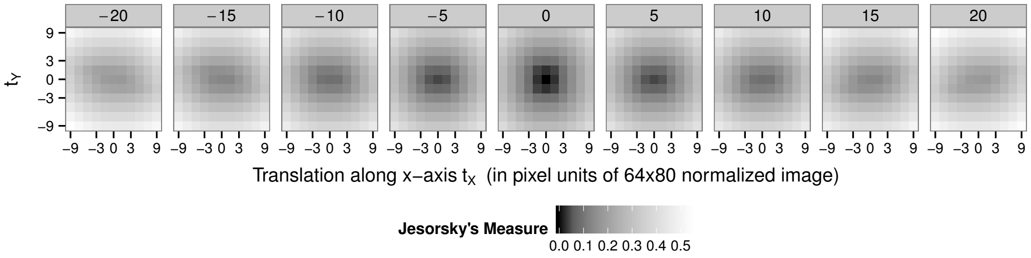


Figure 2: Relationship between the Jesorsky measure and annotation transformation (θ, t_X, t_Y) when same transformation is applied to both eye coordinates for the rotation θ varying between $-20°$ and $20°$ and translation t_X, t_Y between -9 and 9 pixels.

Table 1: List of symbols

(x, y)	cartesian coordinate in original image space (also in subscripts)
(X, Y)	cartesian coordinate in normalized image space (also in subscripts)
p	position vector $[x,\ y]$ in original image space
P	position vector $[X,\ Y]$ in normalized image space
c	center between the two eyes in original image space
C	center between the two eyes in normalized image space
$*^{\{m,d\}}$	superscripts to denote *m*anually and automatically *d*etected eyes
$*_{\{l,r\}}$	subscripts to denote *l*eft and *r*ight eye

$$J = \frac{\max\{\|p_l^m - p_l^d\|, \|p_r^m - p_r^d\|\}}{\|p_l^m - p_r^m\|}, \tag{5}$$

where $p^m_{\{l,r\}}$ denote manually annotated left and right eye coordinates, while $p^d_{\{l,r\}}$ denote automatically detected eye coordinates (as defined in Table 1). In Figure 2, we show the correspondence between the Jesorsky measure J and the transformation parameters (θ, t_X, t_Y) (see next section), when same transformation is applied to both eye coordinates. This measure of misalignment cannot differentiate between errors caused by translation, rotation or scale. This is also evident from Figure 2, which shows that multiple transformation parameters map to same J value. Therefore, in this paper we quantify the amount of misalignment in the normalized image space using translation (t_X, t_Y) and rotation θ parameters.

Misalignment in the original image space is defined as the difference between the manual eye annotations and automatically detected eye locations. However, we report misalignment in units of the normalized image because misalignment in the original image depends on the inter-ocular distance, which in turn varies with image resolution. For the normalized image, the inter-ocular distance remains fixed and, therefore, our results are not affected by resolution of the original image. Here, we establish the relationship between misalignment in the original image space and misalignment in the normalized image space. Such a relationship allows us to express eye detection errors in units of the normalized image space and, therefore, a comparison with the results presented in Section 4.1 and 4.2 is possible.

Geometric normalization of facial images involves scaling and rotating the original image space p such that the manually annotated left and right[4] eyes in the original image $p^m_{\{l,r\}}$ get transformed to a predefined location $P^m_{\{l,r\}}$ in the normalized image:

$$P = \frac{1}{s} R_{-\alpha}(p - c) + C \tag{6}$$

where scale s, rotation angle α and rotation matrix R_α are defined as:

$$s = \frac{\|P^m_r - P^m_l\|}{\|p^m_r - p^m_l\|}, \qquad \alpha = \tan^{-1}\left(\frac{y_r - y_l}{x_r - x_l}\right), \qquad R_\alpha = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}. \tag{7}$$

Using the coordinate space transformation of (6), we transform the original image and extract a predefined region around $P^m_{\{l,r\}}$ to obtain the final geometrically normalized facial image. For our experiments, we use $P^m_l = [15, 16]$ and $P^m_r = [48, 16]$ in a normalized image of dimension 64×80. An exemplary normalized image including the locations of the normalized eye positions $P_{l,r}$ can be found in Figure 1.

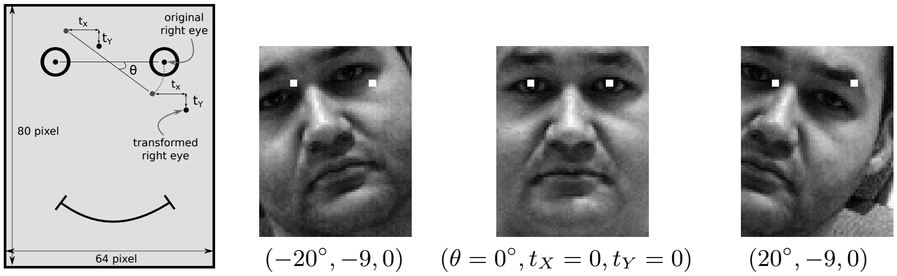
In this paper, we also study the accuracy of automatic eye detectors with respect to manual eye annotations. We report face recognition performance results parametrized by eye position error in the normalized image space. The method to convert eye detection errors from pixel units in the original image to errors in the normalized image is as follows: We first compute scale s and rotation α using the manually annotated eye coordinates p^m in the original image and predefined eye locations P^m in the normalized image as given in (7). Using (6), we transform the automatically detected eye coordinate p^d in original image space to obtain its position P^d in the normalized image space. We do not differentiate between errors in left and right eye coordinates and, therefore, define the eye detection error in normalized image space as:

$$\Delta X = X^d - X^m \qquad\qquad \Delta Y = Y^d - Y^m, \tag{8}$$

where $(\Delta X, \Delta Y)$ denote the difference between manually annotated and automatically detected eye coordinates in the normalized image space.

4 Experiments

In our experiments, we quantify misalignment in units of normalized image space $(\Delta X, \Delta Y)$ as defined in (8) and evaluate their influence on face recognition performance in Sections 4.1 and 4.2. Note that we only perturb the eye coordinates of query images, while for training and enrollment images, we use the manually annotated eye coordinates provided by [5].[5] In Section 4.3, we study the variability between two independently performed manual eye annotations. Furthermore, in Section 4.4, we study the accuracy of automatic eye detectors by considering manual eye annotations as ground truth for eye locations. In Section 5, we present a list of key observations from all these experiments.



$(-20^\circ, -9, 0)$ $(\theta = 0^\circ, t_X = 0, t_Y = 0)$ $(20^\circ, -9, 0)$

Figure 3: Rotation (about center between the eyes) followed by translation of query images in the normalized image space. The two white dots denote the untransformed location of left and right eyes in the normalized image.

4.1 Impact of Translation and Rotation

In our first experiment, we systematically rotate (θ) and translate (t_X, t_Y) the hand-labeled eye coordinates $P^m_{\{l,r\}}$ in normalized image (depicted as two white dots in Figure 3) to simulate misalignment. We apply the same transformation (θ, t_X, t_Y) to both eye coordinates $P^m_{\{l,r\}}$ and, hence, the size of the faces in the misaligned facial images is not varied. The perturbed eye coordinates $\mathcal{P}^m_{\{l,r\}}$ in the normalized image are computed as follows:

$$\mathcal{P}^m_{\{l,r\}} = T(t_X, t_Y)T(C)R(\theta)T(-C)P^m_{\{l,r\}} \tag{9}$$

where C denotes the coordinate of the center of two eyes $P^m_{\{l,r\}}$, while T and R denote the translation and rotation operator, respectively.

In Figure 4, we report the recognition performance for all possible variations of (θ, t_X, t_Y):

$$\begin{aligned} \theta &\in \{-20^\circ, -15^\circ, -10^\circ, -5^\circ, 0^\circ, 5^\circ, 10^\circ, 15^\circ, 20^\circ\} \\ t_X, t_Y &\in \{-9, -7, -5, -3, -1, 0, 1, 3, 5, 7, 9\} \end{aligned} \tag{10}$$

in terms of HTER and AUC. For HTER, the threshold s^*_t is computed on the basis of untransformed images, i. e., for $\theta = 0, t_X = 0, t_Y = 0$.

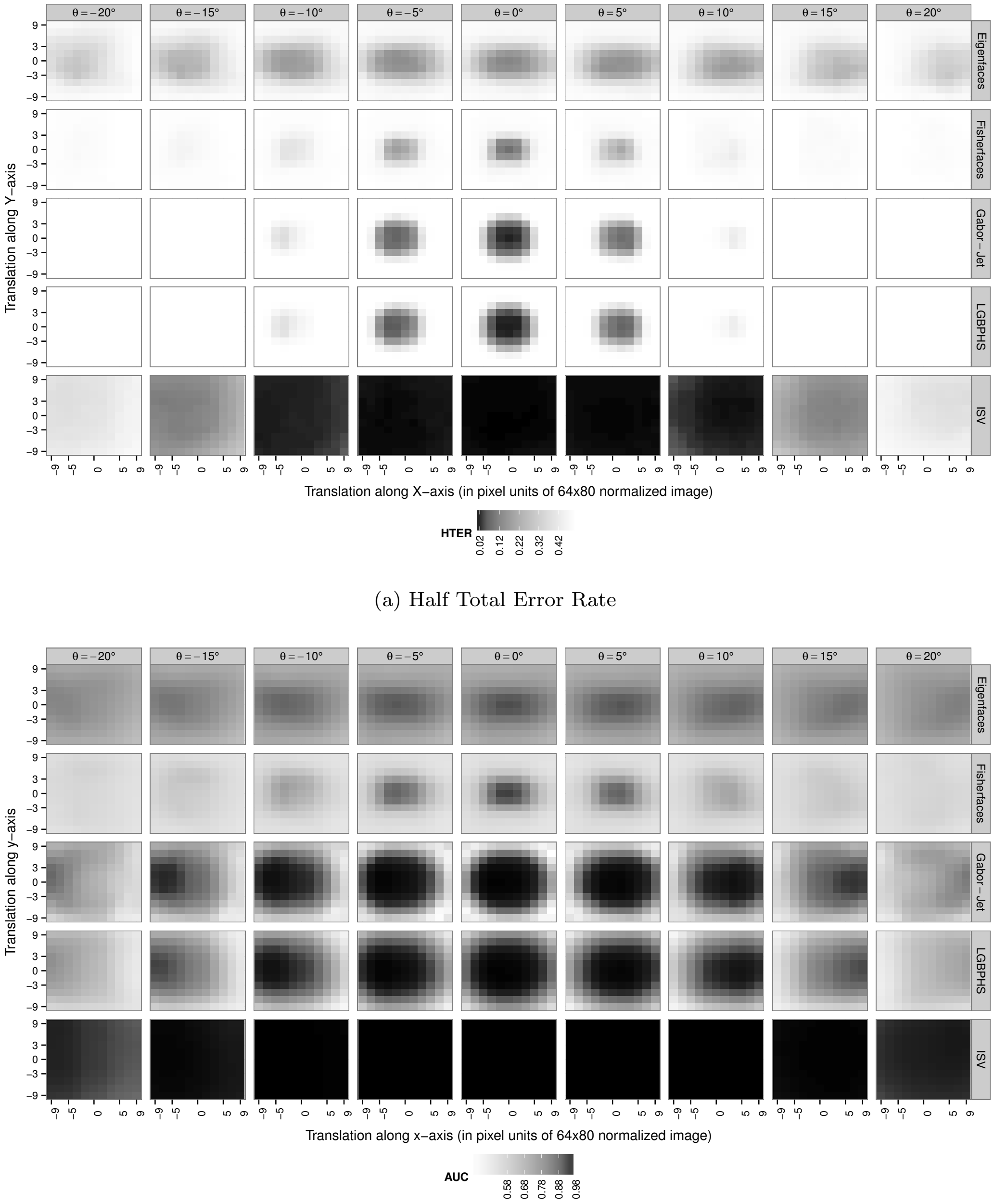
For the Eigenfaces system, which is the first row in Figure 4, translation along vertical direction (t_Y) has more impact on performance as compared to horizontal translation (t_X). For small angles $-5^\circ \leq \theta \leq +5^\circ$ and small translations $-5 \leq t_X \leq +5$ and $-3 \leq t_Y \leq +3$, the Eigenfaces algorithm has stable (though comparably low) performance.

Fisherfaces has better recognition performance as compared to Eigenfaces for aligned facial images. However, its performance drops more rapidly for small misalignment. For instance, the HTER of Fisherfaces increases from 0.08 (AUC = 0.95) to 0.39 (AUC = 0.80) for a horizontal translation of $t_X = 3$ (with $t_Y = 0, \theta = 0$). For the same misalignment, the HTER of Eigenfaces increases from 0.12 (AUC = 0.94) to 0.30 (AUC = 0.87). The authors of [12] have shown that "when the training data set is small, PCA [Eigenfaces] can outperform LDA [Fisherfaces]". Our results highlight another property of Eigenfaces, that it is more robust to misalignment in the input image.

Both Gabor-Jet and LGBPHS achieve an HTER of 0.01 (AUC > 0.99) for properly

[4] with respect to the viewer

[5] The manual eye coordinates can be downloaded from `http://www.idiap.ch/resource/biometric`, they are also included in our source code package `http://pypi.python.org/pypi/xfacereclib.paper.IET2015`.



(a) Half Total Error Rate

(b) Area Under Curve

Figure 4: Impact of same rotation and translation applied to both left and right eye coordinates on performance of five face recognition systems (along rows). Each cell denotes recognition performance for the query images misaligned by the applying the transformation of (θ, t_X, t_Y) to manually annotated eye coordinates.

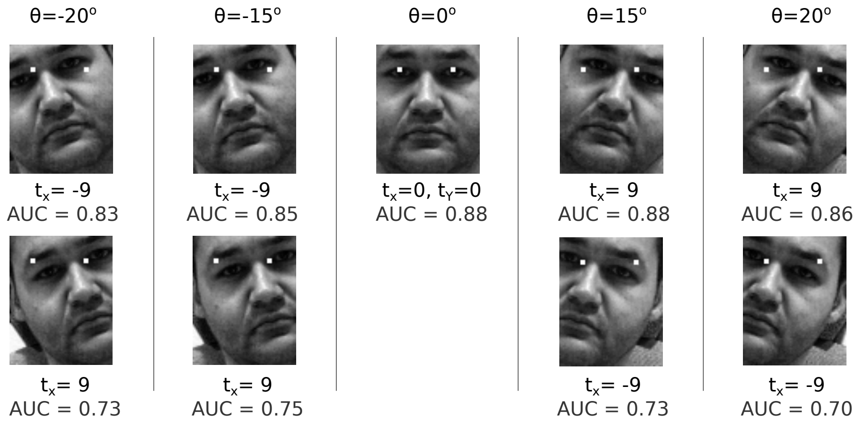
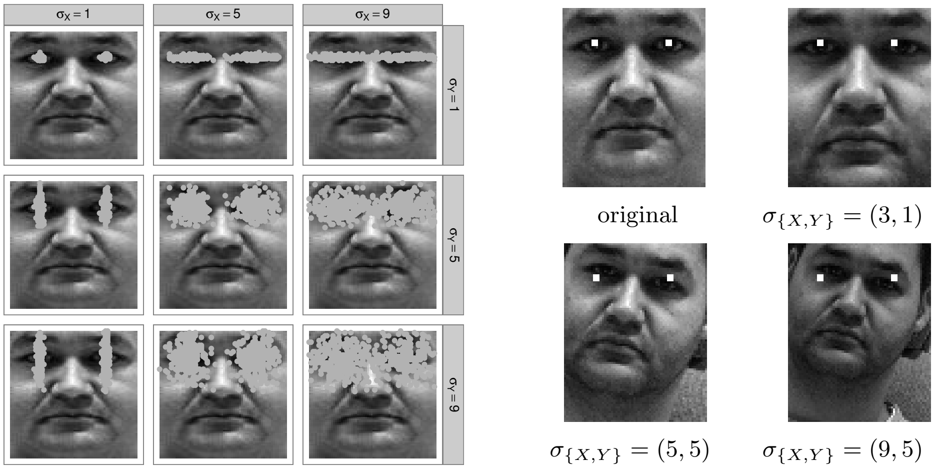


Figure 5: Examples of transformations with large (t_X, t_Y, θ) parameters that cause low (first row) and high (second row) misalignment of facial features. The AUC values of the Eigenfaces algorithm for those transformations are added to the plots.

aligned images. For a horizontal misalignment of 3 pixels ($t_X = 3, t_Y = 0, \theta = 0$), the HTER of Gabor-Jet increases to 0.08 and of LGBPHS to 0.04, while the AUC is stable (AUC > 0.99). From Figure 4, it is clear that both Gabor-Jet and LGBPHS have similar tolerance towards misalignment and both of them have higher recognition performance and a better robustness towards misalignment than Eigenfaces and Fisherfaces.

From the five recognition systems included in this study, ISV clearly has the best tolerance towards misalignment for all possible combinations of rotation and translation. For properly aligned images, ISV achieves HTER of 0.00018 (AUC > 0.99). For a 3 pixel horizontal misalignment ($t_X = 3$, $t_Y = 0$, $\theta = 0$), the HTER increases only to 0.00217 (AUC > 0.99). In ISV, features are independently extracted from each part of the facial image and, therefore, this approach is naturally robust to face misalignment, occlusion and local transformation. It also explicitly models and removes the effect of session variability, e. g., changes in environment, expression, pose and image acquisition. Only for an extreme misalignment of ($t_X = 9, t_Y = 9, \theta = 20°$), the HTER grows to 0.39 (which is close to chance level 0.5), but the AUC = 0.98 still shows very good discrimination abilities. From this effect one can infer that the similarity values change with the transformation, but for both genuine and impostor accesses in the same way. One way to improve the unbiased HTER in this case is given by categorical calibration [9].

For larger rotations ($\theta \geq \pm 10°$), we can observe that the highest recognition performance deviates from the translation center ($t_X = 0, t_Y = 0$). This effect can best be seen in the AUC plots for Gabor-Jet and LGBPHS in Figure 4. To investigate this behavior, we display exemplary images of a subject cropped under these eye perturbations in Figure 5. We observe that some transformations, which are shown in the first row of figure, cause less misalignment of facial features like nose tip, mouth center, etc. and, hence, the performance drop is small. On the contrary, some transformations introduce large amount of misalignment, which even might lead to facial features being outside of the cropped image, as shown in the second row of Figure 5. For those transformations, the recognition performance severely degrades.



(a) Sampled eye perturbations

(b) Images after Transformation

Figure 6: Random transformation applied to left and right eye coordinates independently, where random samples are drawn from a normal distribution with $\mu_{\{X,Y\}} = 0$.

4.2 Impact of Translation, Rotation and Scaling

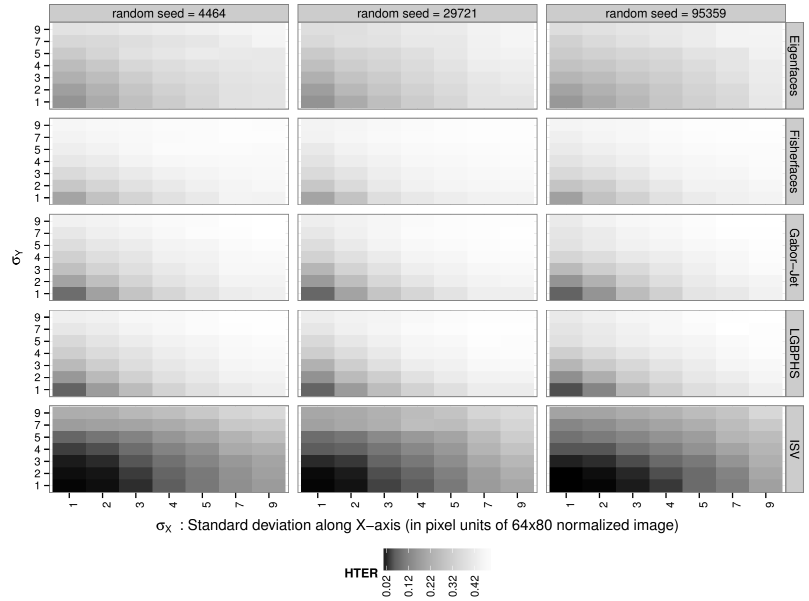
In the experiments of Section 4.1, we apply the same transformation to both eyes in the normalized image. This eye coordinate transformation strategy can only simulate rotation and translation error, but not the error caused by scaling. Typically, in a practical automatic eye detector, all three types of transformations are present. Therefore, in order to simulate the misalignment caused by an automatic eye detector, we independently apply random translation to the left and right eye coordinates as follows:

$$\mathcal{X}^m = X^m + \epsilon_X \,, \qquad\qquad \mathcal{Y}^m = Y^m + \epsilon_Y \,, \tag{11}$$

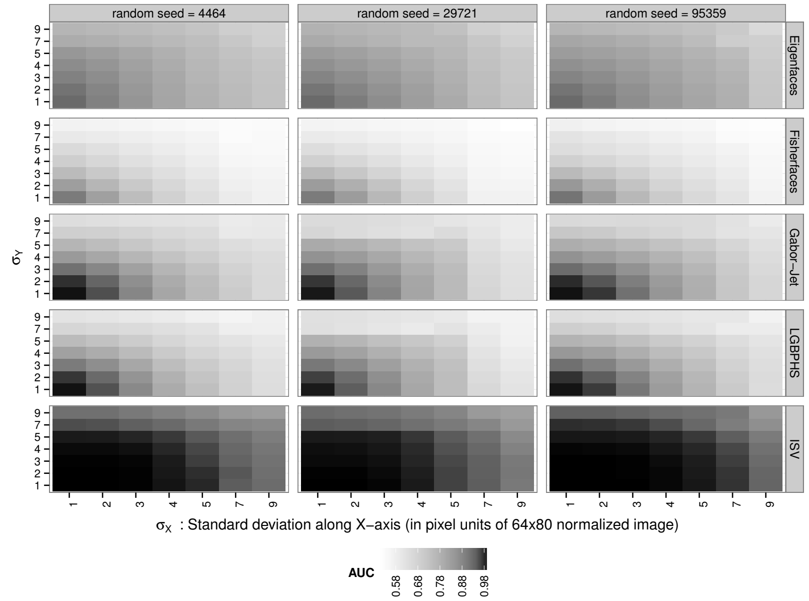
where $\epsilon_{\{X,Y\}}$ follows the normal distribution $\mathcal{N}(\mu = 0, \sigma_{\{X,Y\}})$, and $\mathcal{P}^m = (\mathcal{X}^m, \mathcal{Y}^m)$ denotes the perturbed eye coordinates of the normalized image. During random sampling, we discard all samples that move the eye coordinate location beyond the boundary of the normalized image (64×80). In Figure 6, we show the randomly perturbed eye locations superimposed on a sample facial image. Additionally, we display exemplary normalized images obtained from such a random eye transformation scheme. It clearly shows the misalignment caused by all three types of transformations: translation, rotation and scaling.

In Figure 7, we report the face recognition performance corresponding to each possible combination of $(\sigma_X, \sigma_Y) \in \{1, 2, 3, 4, 5, 7, 9\}$ (in pixel units of the normalized image) for three different sets of random positions. We obtain consistency in performance variations of systems across three random seeds, which shows that our random samples and, hence, our random eye perturbations are not biased.

When changing random eye perturbations from $\sigma_{\{X,Y\}} = 1$ to $\sigma_{\{X,Y\}} = 3$, the HTER of Eigenfaces increases from 0.14 (AUC = 0.93) to 0.26 (AUC = 0.86), whereas the HTER of Fisherfaces experiences a more drastic increase from 0.12 (AUC = 0.93) to 0.37 (AUC = 0.71). In the previous section, we observed that Eigenfaces, as compared to Fisherfaces, is more robust to misalignment caused by translation and rotation. The results from the



(a) Half Total Error Rate



(b) Area Under Curve

Figure 7: Impact of random eye perturbations applied to left and right eye coordinates independently on performance of five face recognition systems (along rows). Random perturbations are sampled from a normal distributions with $(\mu = 0, \sigma_{\{X,Y\}})$.

random eye perturbation experiment show that Eigenfaces is more robust towards all types of misalignment.

The performance drop for both Gabor-Jet and LGBPHS are similar for change in random eye perturbations from $\sigma_{\{X,Y\}} = 1$ to $\sigma_{\{X,Y\}} = 3$. The HTER of Gabor-Jet increases from 0.05 (AUC $>$ 0.99) to 0.33 (AUC = 0.87), while that of LGBPHS increases from 0.04 (AUC $>$ 0.99) to 0.32 (AUC = 0.86). Comparing this large drop in performance with results from the previous experiment (involving only translation and rotation), we can conclude that the performance of both Gabor-Jet and LGBPHS is more susceptible to scaling variations. Additionally, we can observe that both Gabor-Jet and LGBPHS drop performance similarly for misalignment of the same kind.

ISV has the best tolerance to misalignment involving translation, rotation and scaling. From Figure 7, this property of ISV is evident from the larger dark region (corresponding to good performance) as compared to the remaining four face recognition systems. Its HTER increases from 0.002 (AUC $>$ 0.99) to 0.045 (AUC $>$ 0.99) when the random eye perturbations change from $\sigma_{\{X,Y\}} = 1$ to $\sigma_{\{X,Y\}} = 3$. For larger perturbations of $\sigma_{\{X,Y\}} = 9$, the HTER still is 0.33 (AUC = 0.84). This shows that ISV experiences a significant drop in performance only for extreme misalignment.

4.3 Ambiguity in the Location of Eyes

With this experiment, we investigate the ambiguity in location of the eyes by comparing the manual eye annotations performed by two independent institutions. For the 1160 frontal images in the Multi-PIE M protocol, we possess the manual eye annotations from two independent sources: from the Idiap Research Institute (Switzerland) [5] and from the University of Twente - UT (Netherlands). In Figure 8, we show the distribution of the difference in x and y coordinate of the two eye annotations. In order to allow comparisons with results from Sections 4.1 and 4.2, we obtain a mapping from the original 640×480 pixel image space to the 64×80 pixel normalized image space by first computing s and α with (7) using the Idiap eye annotations as the base, and then transforming the UT eye annotations to the normalized space using (6) and computing the difference with the manual eye locations in the normalized image.

Most face recognition systems employ a carefully tuned automatic eye detector to obtain the location of the eyes. In the second part of this experiment, we investigate the accuracy of automatic eye detectors present in the commercial face recognition systems FaceVACS [3] and Verilook [14]. The correlation between automatically detected eye coordinates and manually annotated eye locations is shown in Figure 9, where the distribution of errors of the two automatic eye detectors are shown considering the manual eye annotation of Idiap [5] as ground truth. We transform this error distribution to normalized image space using the same procedure as for the UT annotations.

Note that the Idiap manual annotations and automatic eye annotations from FaceVACS and Verilook were carried out on no-flash images (i. e., under ambient illumination), while the UT manual annotations were performed on images captured using frontal flash (i. e., flash=07).

In Figure 9 it can be seen that out of 1160 image samples, Verilook generated eye detection errors $>$ 50 pixel for 15 images. Some of these are shown in Figure 10, which reveals that dark skin color combined with no-flash photographs contribute to large errors in automatic eye detection by Verilook. For visual clarity of Verilook's detection error histogram in Figure 11, we exclude those samples.

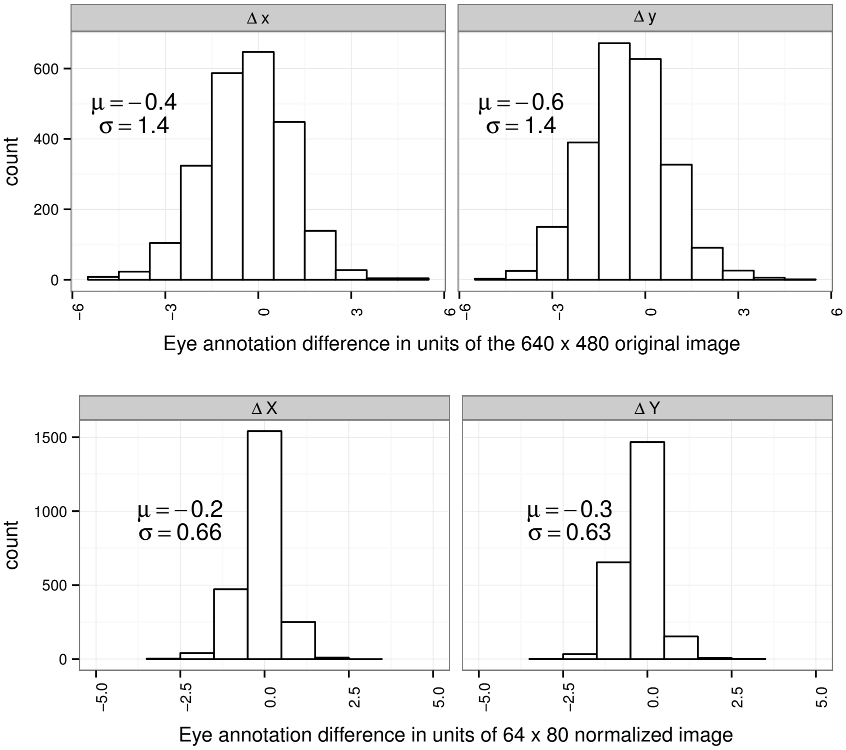


Figure 8: Statistical difference between manual eye annotations carried out independently at Idiap Research Institute (Switzerland) and University of Twente (Netherlands).

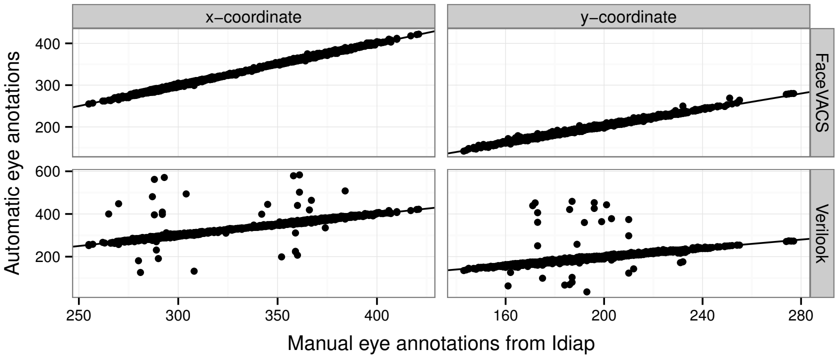


Figure 9: Correlation between manually annotated [5] and automatically detected eye coordinates in the pixels units of the original image space. The black solid line indicates a correlation of 1.

To see, whether the difference between the two independent manual eye annotations shown in Figure 8 follows a normal distribution, we plot the quantiles of this manual annotation difference (normalized such that $\mu = 0$ and $\sigma = 1$) against the quantiles of a standard normal distribution. If the manual annotation difference follows the normal distribution,

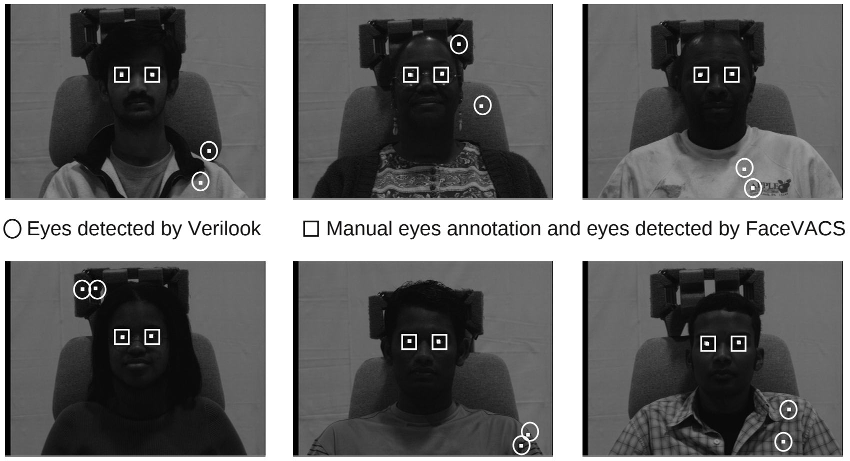

○ Eyes detected by Verilook
□ Manual eyes annotation and eyes detected by FaceVACS

Figure 10: Some sample images, for which the Verilook eye detector has an eye detection error > 50 pixels.

the points on this plot, which is also called a Quantile-Quantile (QQ) plot, should lie on the line $y = x$. The QQ plot of Figure 12 (top row), confirms that the manual annotation difference follows a normal distribution. We subtract the mean value from this distribution to remove systematic offset, if any, in the eye annotations and claim that:

$$\begin{aligned} Pr(-1 \leq \delta_x \leq 1) &= 0.51 & \quad Pr(-1 \leq \delta_y \leq 1) &= 0.53 \\ Pr(-2 \leq \delta_x \leq 2) &= 0.83 & \quad Pr(-2 \leq \delta_y \leq 2) &= 0.85 \\ Pr(-4 \leq \delta_x \leq 4) &= 0.99 & \quad Pr(-4 \leq \delta_y \leq 4) &= 0.99 \end{aligned} \tag{12}$$

where $[\delta_x, \delta_y] = p^{m,idiap} - p^{m,ut}$ denote the difference between Idiap and UT manual eye annotation in the original image space, and $Pr(-4 \leq \delta_x \leq 4)$ denotes the probability of four pixel difference in manual annotations along horizontal direction for frontal facial images, which in Multi-PIE have an average inter-ocular distance of 70 pixels. From this experiment, we have empirical evidence for an ambiguity of four pixels in the location of the eyes. Currently, we have access to only two independent sources of manual annotations for the Multi-PIE data set. However, we would need more independent sources of manual annotation to check if this conclusion generalizes to a larger population of manual annotators.

Now we investigate the error characteristics of automatic eye detectors as shown in Figure 11. The correlation plot of Figure 9 shows that the Verilook eye detector has large errors for many image samples, while FaceVACS is fairly accurate. Hence, we only include the error distribution of FaceVACS (as shown in Figure 11) in our further analysis. The QQ plot of Figure 12 (bottom row) shows that the error distribution of FaceVACS follows a normal distribution. For some large errors in vertical location (i. e. $(x^{m,idiap} - x^{d,fv}) < -3$), the distribution deviates from this normal distribution. Assuming normality and subtracting the mean from this distribution to remove systematic offset in eye detections, we can claim that:

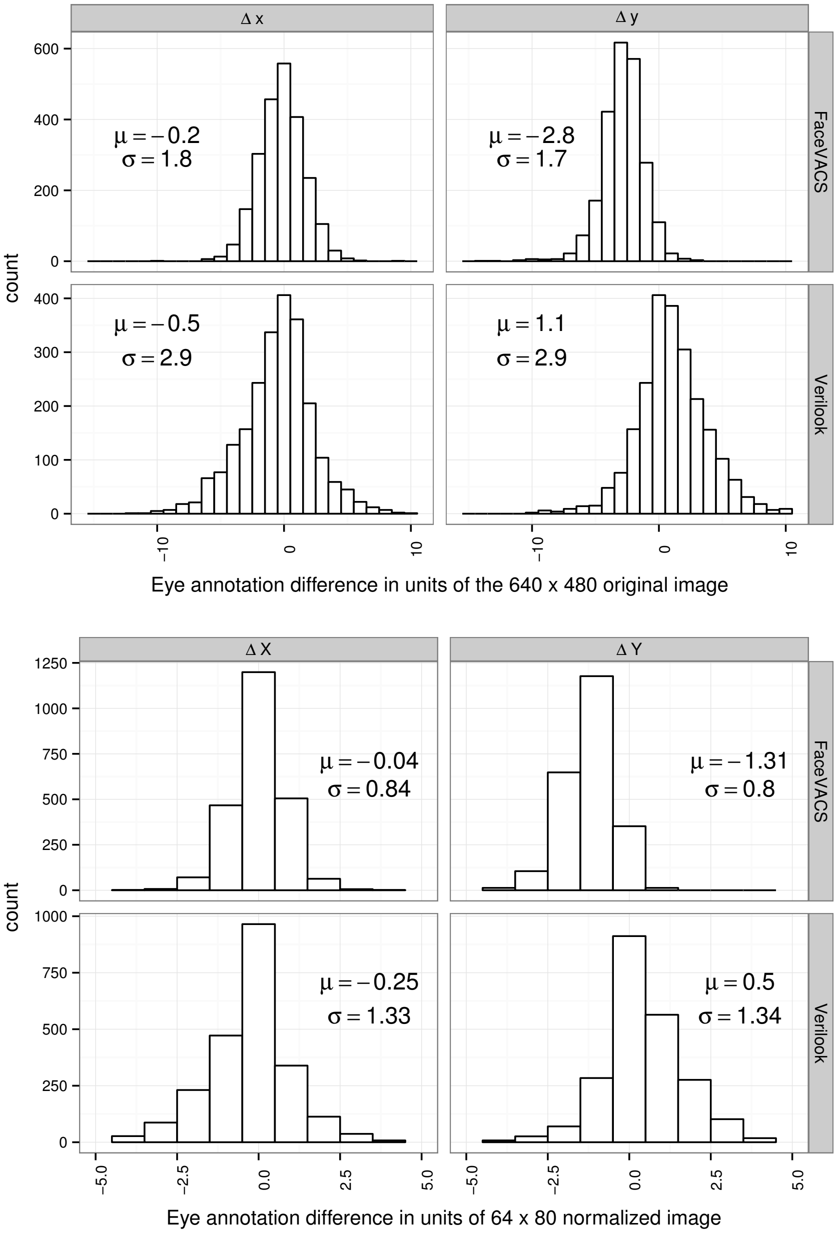


Figure 11: Difference in eye locations detected by the FaceVACS and Verilook eye detector with respect to manual eye annotations [5]. For Verilook, 15 samples with eye detection error > 50 pixels are excluded.

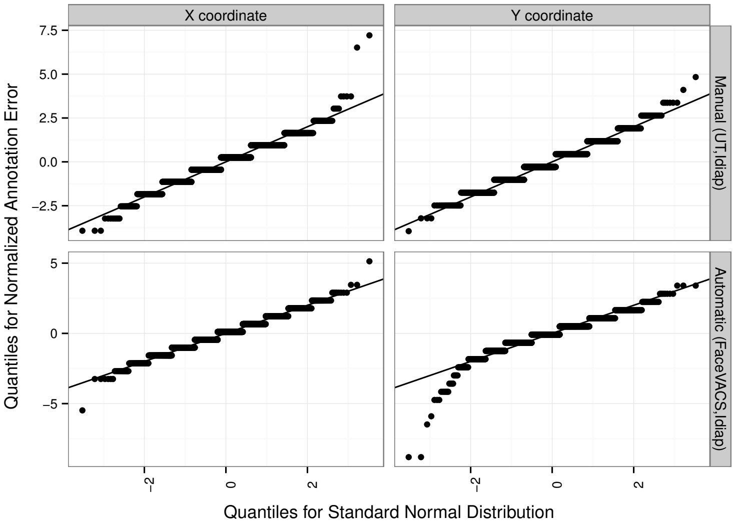


Figure 12: Quantile-Quantile plot of standard normal distribution ($\mu = 0, \sigma = 1$) and normalized manual annotation difference distribution ($\frac{x-\mu}{\sigma}$) shown in Figures 8 (UT) and 11 (FaceVACS only). Note that the staircase pattern is caused by discrete pixel location values.

$$\begin{array}{llll} Pr(-1 \leq \delta_x \leq 1) & = 0.42 & Pr(-1 \leq \delta_y \leq 1) & = 0.43 \\ Pr(-2 \leq \delta_x \leq 2) & = 0.73 & Pr(-2 \leq \delta_y \leq 2) & = 0.75 \\ Pr(-4 \leq \delta_x \leq 4) & = 0.97 & Pr(-4 \leq \delta_y \leq 4) & = 0.97 \\ Pr(-6 \leq \delta_x \leq 6) & = 0.99 & Pr(-6 \leq \delta_y \leq 6) & = 0.99 \end{array} \tag{13}$$

where $[\delta_x, \delta_y] = p^{m,idiap} - p^{d,fv}$ denote the difference between Idiap and FaceVACS eye annotations. Comparing (12) and (13), we observe that – after compensation for a systematic offset – the accuracy of FaceVACS eye detector comes very close to the accuracy of manual annotators. This observation is also evident in Figure 11, which shows that FaceVACS achieves a standard deviation of approximately 1.8 pixels, while manual eye annotators achieve a standard deviation of 1.4 pixels as shown in Figure 8 in both horizontal and vertical directions.

For the FaceVACS eye detector, we observe a systematic offset of around 3 pixels along the vertical direction as shown in Figure 11. This shows that the FaceVACS detector is trained with a different notion of eye center, revealing the lack of consistency in the definition of eye center in a facial image: Does the eye center refer to center of the pupil, or to center between the two eye corners or eyelids, or to something else?

4.4 Choice of Eye Detector for Training, Enrollment and Query

Most face recognition systems go through the following three phases of operations: *a*) a training phase to learn the representation of facial features, *b*) an enrollment phase to enroll models from facial images of known identities and *c*) a query phase to verify the identity in a given query image. All three phases exploit the location of the eyes for alignment of images. We consider the training and enrollment phase as one group because they are

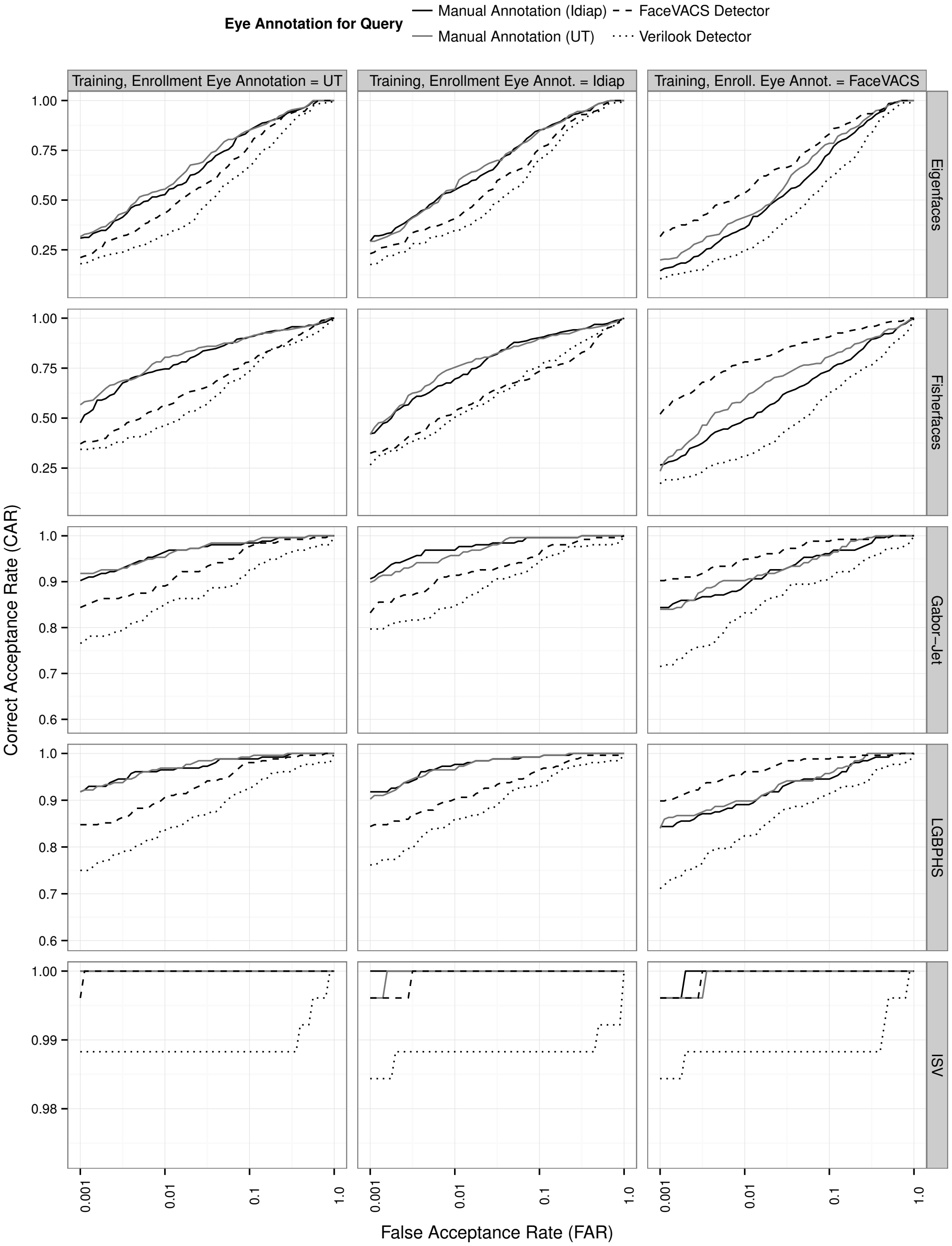


Figure 13: Face recognition performance with eye annotation provided by UT (manual), Idiap (manual), FaceVACS (automatic) and Verilook (automatic). Note that the range of CAR values are different in each row of the above plot.

defined during the off-line development of a face recognition system. In this section, we investigate the impact of using different eye detectors for on-line query phase and the off-line training/enrollment phase.

For the Multi-PIE M protocol images, we investigate the following combinations of eye annotations: *a*) training and enrollment images aligned using manual eye annotations from Idiap, UT or automatic eye locations from FaceVACS eye detector *b*) query images aligned using two types of manual annotations (Idiap, UT) or automatic eye annotations (FaceVACS and Verilook).

With this experiment, we investigate the impact of using different sources of eye annotations for the training/enrollment and the query phase of face recognition systems. In the first column of Figure 13, we show the recognition performance for the five open source systems when training and enrollment images are annotated using manual annotations from Idiap and the query images are labeled using manual (Idiap,UT) and automatic (FaceVACS, Verilook) eye annotations. Note that the range of CAR values are different in each row of this plot.

In Figure 13 (first and second column), we observe that both Idiap and UT manual annotations in the query set have similar recognition performance, independent on which of both were used in the training/enrollment stage. This shows that all of the investigated face recognition systems, when trained using manual eye annotations, are tolerant to small differences up to four pixels (with average inter-ocular distance of 70 pixels) in manual annotation caused by inter-observer variability. Moreover, when training/enrollment is done using manual annotations, the recognition performance for query images annotated using manual annotation is always higher than that obtained using automatic eye detectors. For the accurate eye detector of FaceVACS, we observe a large drop in performance, but this is primarily due to inconsistency in the definition of eye center. The Verilook eye detector is less accurate (as evident from Figure 11) and, therefore, all algorithms operating on facial images aligned using Verilook eye coordinates have poor performance.

The third column of Figure 13 shows the performance variation when training and enrollment images are labeled using FaceVACS automatic eye annotations. We observe that FaceVACS eye annotations for query achieves best recognition performance as compared to other sources of annotation for query images. The large drop in performance for manually annotated query images is, in turn, due to inconsistency in the definition of the eye centers. Moreover, for Eigenfaces and Fisherfaces, these performances are comparable to the performance of manually annotated query as shown in the first and second columns of Figure 13. For Gabor-Jet and LGBPHS, there is a slight drop in performance as compared to the manual annotations cases, but still the performance with the FaceVACS query images is best among the other three annotation sources. This shows that an accurate automatic eye detector (like that of FaceVACS) can help achieve recognition performance comparable to that obtained using manually annotated eyes, given that the same automatic detector is used for annotating the training, enrollment and query images.

The ISV algorithm seems to be unaffected by the source of eye annotations. Except for the complete misdetections of Verilook discussed in Section 4.3, the ROC curves are stable at a very high level in the last row of Figure 13. Not even the different definition of eye centers disturbs the recognition capabilities of ISV. Most probably this stability comes through the fact that facial features are extracted locally from the image, and the distribution of these features is modeled independently.

5 Discussion

For different types of misalignment in a 64×80 normalized image space, we evaluated the performance of following five open source face recognition systems: Eigenfaces, Fisherfaces, Gabor-Jet, LGBPHS and ISV. We simulated different types of facial image misalignment by scaling, rotating and translating manually annotated eye locations. We found that Eigenfaces is more robust to misalignment (caused by scaling, rotation or translation of eye locations) as compared to Fisherfaces. However, Fisherfaces has higher recognition performance (HTER=0.08, AUC=0.95) for properly aligned images (i. e., $t_X = 0, t_Y = 0, \theta = 0$) as compared to Eigenfaces (HTER=0.12, AUC=0.94). Furthermore, we found that Eigenfaces is more sensitive to vertical misalignment as compared to misalignment along horizontal direction. However, in any case the Eigenfaces and Fisherfaces algorithms showed performance inferior to the other investigated methods. Both Gabor wavelet based methods, Gabor-Jet and LGBPHS, have similar tolerance towards misalignment with both of them being more susceptible to misalignment caused by scaling. We found that ISV has the best tolerance for misalignment since it is able to maintain a consistent level of performance for a large range of misalignment ($\mu = 0, \sigma \leq 3$). ISV demonstrates such a natural robustness to misalignment because features from all parts of the facial image are modeled independently.

We investigated two different evaluation measures, AUC and HTER. While the former measure is biased since it evaluates performance after all scores of a certain database have been collected, the HTER is unbiased and more application oriented since it requires a threshold to be selected prior to evaluation. We have found that the biased AUC was often stable, while the HTER dropped drastically. This shows that the threshold for HTER cannot simply be selected based on hand-labeled eye locations when the query set is detected automatically. Hence, other strategies as score normalization [21] or calibration [9] need to be applied to the scores from the face recognition systems in order to make use of the face recognition algorithms in case of automatically detected eyes.

In practical face recognition systems, an automatic eye detector is used to localize the two eyes and then perform automatic registration of facial images. Therefore, we investigated the accuracy of automatic eye detectors present in two commercial face recognition systems: FaceVACS and Verilook. Additionally, we analyzed the difference between two independent sources of manual eye annotations for the frontal images of the Multi-PIE M protocol. This allows us to understand the inherent limitations of using the two eye coordinates as the landmarks for facial image registration. We found an ambiguity of around 4 pixels in manual eye annotation of frontal view images with an average inter-ocular distance of 70 pixels. Therefore, face recognition systems should be built to tolerate at least this amount of error in eye coordinates. Those 4 pixels translate to approximately 2 pixels in the normalized image, which means that the algorithms Gabor-Jet, LGBPHS and ISV still perform reasonably well, cf. Figure 7. The automatic eye detector of FaceVACS achieves a detection accuracy ($\sigma = 1.8$) that is close to the accuracy of manual annotators ($\sigma = 1.4$). However, the eye locations were detected on well-illuminated frontal images and, thus, this result needs to be verified in presence of illumination or non-frontal pose. We observed higher eye detection error in the automatic eye detector included in the Verilook system. We found that the FaceVACS eye detector showed a systematic offset of 3 pixels in vertical location of the eyes, which reveals lack of consistency in the definition of the eye center in frontal facial images.

We also explored the impact of using different sources (automatically detected or manually located) of eye annotations for training, enrollment and query phases of a face recognition system. We found that using manual eye annotations for training and enrollment while utilizing automatic eye annotations for query results in a large drop in performance, but we discussed that this is caused by the inconsistent definition of the eye centers. We also found that using eye annotations from a accurate automatic eye detector (like that of FaceVACS) for all training, enrollment and query images results in face recognition performance that is comparable to the performance achieved using manual eye annotations. Therefore, our results underline the importance of consistent definition of eye center in a facial image and also highlights the performance gain achieved by using same automatic eye detector for training, enrollment and query images. Furthermore, the performance of ISV remains consistently high for any combination of eye annotation sources. This shows that a combination of moderately accurate eye detector and a face recognition system that is naturally robust to moderate misalignment can potentially be a solution for practical applications.

One important fact of a face recognition algorithm is its complexity. The execution time of the five face recognition systems on a Intel i7 3.5 GHz (4 cores) machine for training, enrollment and scoring operation has been recorded as: ISV: 159.8 min., LGBPHS: 9.7 min., Gabor-Jet: 2.8 min., Fisherfaces: 1.8 min. and Eigenfaces: 1.7 min. Hence, the robustness of ISV towards misalignment comes at the expense of very high computational costs. The best trade-off between accuracy and complexity in our tests was achieved by the Gabor-Jet algorithm.

Another important point of this evaluation is that all experiments solely rely on open source software – except for the automatically detected eye locations, which were generated by third party software. Effectively, we provide the scripts and documentation to install the required software, to rerun all face recognition experiments presented in this paper, and to regenerate Figures 4, 7 and 13. Additionally, the source code can be easily adapted to run the same experiments using a different image database (for which at least the hand-labeled eye positions must be available) or to investigate the stability of other face recognition algorithms towards eye localization errors.

6 Conclusion

In this investigation, the aim was to determine the impact of misalignment caused by errors in eye localization on the performance of face recognition systems. Similar studies carried out in the past were either limited by the number of face recognition systems or the size of facial image database. Our study is based on five open source face recognition algorithms operating on a larger facial image database. One of the more significant findings to emerge from this study is the ambiguity in the definition of eye centers in a facial image. The two eye centers are widely used as the landmarks for registration of facial images. However, a commonly agreed definition of the "eye center" is still missing. This causes inconsistency in the eye locations detected by different automatic eye detectors thereby reducing performance when eye detectors and face recognition systems of different origin are mixed. Perhaps, this is the most serious limitation of using the two eye centers as the landmarks for facial image registration. If the same automatic eye detector is used for annotating the training, enrollment and query images, our experiment results show that it is possible to achieve recognition performance comparable to that obtained using manually annotated eyes, given that the facial images are well-illuminated and show a frontal pose.

We compared the manual eye annotations obtained from two independent sources to study the ambiguity in manual eye annotations. To the best of our knowledge, such a study has not been reported before. We found that there exists an ambiguity of four pixels in manual annotations of the two eyes when the frontal facial images have an average inter-ocular distance of 70 pixels. Therefore, assuming that humans are the best possible eye detectors, face recognition systems that use the location of the eyes for alignment should be built to handle at least this level of ambiguity, which was the case for the five investigated face recognition systems. Our results also show that the accuracy of FaceVACS automatic eye detector is very close to that of manual eye annotators.

It has been demonstrated that the Jesorsky measure is insufficient to distinguish between landmark localization errors that cause the normalized image to be shifted or rotated. On the other hand, we have shown that most algorithms have a higher tolerance towards translation and rotation than towards scaling. Hence, the Jesorsky error of an automatic eye detector has a limited correlation with the actual face recognition performance of a face recognition system.

The results reported in this paper reveal the nature of five open source face recognition algorithms towards misalignment caused by errors in eye localization. ISV demonstrates excellent tolerance towards large amount of misalignment caused by errors in eye localization. Its performance drops only for extreme misalignment of facial images, but this performance comes at a cost of long execution time and, hence, might not be usable under real-time requirements. Gabor-Jet shows good tolerance towards misalignment and has very low execution time as compared to ISV. Both Gabor-Jet and LGBPHS have similar tolerance towards misalignment and they have higher recognition performance and are more robust to misalignment as compared to Eigenfaces and Fisherfaces.

Due to the availability of independent hand-labeled sources of eye landmarks, the present study was performed on the Multi-PIE image database. A further study could include more face recognition systems or more challenging image conditions like different illumination, facial expressions and head pose. Since the tools used in this study are open source and released with this paper, it is possible to perform such a study with minimal effort. Also the impact of score normalization or calibration on the performance of the unbiased evaluation needs to be addressed.

Acknowledgement

- This work was supported by the BBfor2 project, which is funded by the European Commision as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008) under Grant Agreement number 238803.

- We would like to thank Cognitec Systems GmbH for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendor's maximum effort full capability result.

References

[1] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*, pages 1449–1452. ACM Press, October 2012.

[2] P.N. Belhumeur, J.P. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):711–720, 1997.

[3] Cognitec Systems GmbH. FaceVACS C++ SDK Version 8.4.0. Software Development Kit (SDK), 2010.

[4] H.K. Ekenel and R. Stiefelhagen. Face alignment by minimizing the closest classification distance. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, pages 1–6, Sept 2009.

[5] L. El Shafey, C. McCool, R. Wallace, and S. Marcel. A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(7):1788–1794, July 2013.

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 1–8, 2008.

[7] M. Günther, D. Haufe, and R.P. Würtz. Face recognition with disparity corrected Gabor phase differences. In A.E.P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm, editors, *Artificial Neural Networks and Machine Learning*, volume 7552 of *Lecture Notes in Computer Science*, pages 411–418. Springer Berlin, September 2012.

[8] M. Günther, R. Wallace, and S. Marcel. An open source framework for standardized comparisons of face recognition algorithms. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 547–556. Springer, 2012.

[9] M. I. Mantasari, M. Günther, R. Wallace, R. Saedi, S. Marcel, and D. Van Leeuwen. Score calibration in face recognition. *IET Biometrics*, 2014.

[10] O. Jesorsky, K.J. Kirchberg, and R.W. Frischholz. Robust face detection using the Hausdorff distance. In *Audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.

[11] J. Marques, N.M. Orlans, and A.T. Piszcz. Effects of eye position on eigenface-based face recognition scoring. *Image*, 8, 2000.

[12] A.M. Martinez and A.C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(2):228–233, 2001.

[13] Jaesik Min, KevinW. Bowyer, and PatrickJ. Flynn. Eye perturbation approach for robust recognition of inaccurately aligned faces. In T. Kanade, A.K. Jain, and N.K. Ratha, editors, *Audio- and Video-Based Biometric Person Authentication*, volume 3546 of *Lecture Notes in Computer Science*, pages 41–50. Springer Berlin Heidelberg, 2005.

[14] Neurotechnology Biometric SDK 4.2. Verilook 5.1. Software Development Kit (SDK), 2011.

[15] T. Riopka and T. Boult. The eyes have it. In *Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, WBMA '03, pages 9–16, New York, NY, USA, 2003. ACM.

[16] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24(8):882 – 893, 2006.

[17] S. Shan, Y. Chang, W. Gao, B. Cao, and P. Yang. Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 314–320. IEEE, 2004.

[18] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[19] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Towards a practical face recognition system: Robust alignment and illumination by sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2):372–386, Feb 2012.

[20] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Inter-session variability modelling and joint factor analysis for face authentication. In *International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2011.

[21] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Cross-pollination of normalisation techniques from speaker to face authentication using Gaussian mixture models. *IEEE Transactions on Information Forensics and Security*, 7(2):553–562, 2012.

[22] H. Wang and P.J. Flynn. Sensitivity of face recognition performance to eye location accuracy. *Biometric Technology for Human Identification II, Proc. SPIE 5779*, pages 122–131, 2005.

[23] P. Wang, M.B. Green, Q. Ji, and J. Wayman. Automatic eye detection and its validation. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 164–164, June 2005.

[24] P. Wang, Q. Ji, and J.L. Wayman. Modeling and Predicting Face Recognition System Performance Based on Analysis of Similarity Scores. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):665–670, 2007.

[25] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 786–791 Vol. 1, 2005.