Pose Variability Compensation Using Projective Transformation for Forensic Face Recognition

Ester Gonzalez-Sosa, Ruben Vera-Rodriguez,
Julian Fierrez, Pedro Tome and Javier Ortega-Garcia
Biometric Recognition Group - ATVS, EPS, Universidad Autonoma de Madrid
Avda. Francisco Tomas y Valiente, 11 - Campus de Cantoblanco,
28049 Madrid, Spain
{ ester.gonzalezs,ruben.vera,julian.fierrez,pedro.tome,javier.ortega }@uam.es

Abstract: The forensic scenario is a very challenging problem within the face recognition community. The verification problem in this case typically implies the comparison between a high quality controlled image against a low quality image extracted from a close circuit television (CCTV). One of the downsides that frequently presents this scenario is pose deviation since CCTV devices are usually placed in ceilings and the subject normally walks facing forward. This paper proves the value of the projective transformation as a simple tool to compensate the pose distortion present in surveillance images in forensic scenarios. We evaluate the influence of this projective transformation over a baseline system based on principal component analysis and support vector machines (PCA-SVM) for the SCface database. The application of this technique improves greatly the performance, being this improvement more striking with closer images. Results suggest the convenience of this transformation within the preprocessing stage of all CCTV images. The average relative improvement reached with this method is around 30% of EER.

1 Introduction

Face biometric trait has been established in the biometric recognition field as one of the least intrusive biometric techniques [ARP04]. This is because it does not require any cooperation from the user. Face recognition can be applied to a wide range of different applications, which range from access control, commercial applications, government issued identity documents, up to law enforcement applications.

Although the problem of face recognition under controlled conditions has achieved great enhancements [YST14], there are still challenges to overcome.

The forensic scenario is one of the areas in which face recognition is involved. The crucial issue of this scenario is dealing with the differences of the images to be compared. The most challenging case within the forensic scenario implies a comparison between a high-resolution image, also known as mug shot, against a low-resolution image acquired from a CCTV device. While mug shot images are extracted under controlled conditions of pose, illumination and background, CCTV images are acquired unobtrusively. The CCTV camera is generally a low-resolution device that acquires video without focusing on the subject.

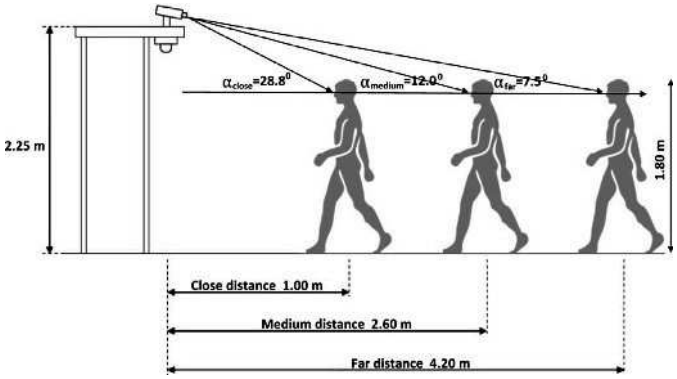

Figure 1: SCface database. There are three different acquisitions distances: close, medium and far. Acquisition angle of each distance calculated for a subject with mean height of 1.80 meters. Figure extracted from [RVRR13]

This fact leads to images with any kind of variation: illumination, pose, expression, occlusion etc. Also, CCTV cameras are commonly situated at ceilings or corners of ceilings, turning towards the floor. As subjects normally walk facing forward, we encounter that the majority of faces detected from the CCTV images suffer from pitch rotation, hindering even more the matching against a frontal image.

Bearing in mind that the last decision in a forensic case is normally done manually by a forensic examiner, any previous automatic procedure to make CCTV images look more similar to mug shot images would be useful to better carry out this manual comparison.

The work developed by *Klontz and Jain* [KJ13] shows a real application of a forensic scenario. In this case, after the terrible incident of the Boston Marathon bombing, an experimental work was carried out to show off the potential capabilities of automatic face recognition system to narrow down the search in this investigation. Results suggest that although state-of-the-art commercial face recognition systems are not yet ready to produce rank-1 results they would help hugely to reduce the number of subjects of the watchlist that are currently being compared manually.

Although there are approaches based on new and more challenging datasets such as [PPB13], neither of them is focused strictly under a forensic point of view. The SCface [KDG11] is a more suitable database for studying the forensic scenario. This database contains both mug shot and CCTV images from three different distances (far, medium and close) from 130 subjects. Fig. 1 shows the situation of the subject for the three different distances. The reader may notice the difference in angle deviation between the position of the camera and the head of the subject, which depends on the height of the subject and the distance to the camera. Hence, for an average height of 1.80 meters, close distance images suffer from a pitch rotation of $28.8°$; medium distance images from $12.0°$ and far distance images from $7.5°$.

The work carried out in [RVRR13] builds a system based on principal component analysis (PCA) and support vector machines (SVM) for the SCface database [KDG11]. Concretely,

three different systems are developed according to the distance to the camera. Among the different experiments carried out in this work, we focus our efforts on one of the most challenging cases in the forensic scenario, which consists of facing mug shot images against CCTV images.

Assessing the results obtained for the mug shot against CCTV images experimental protocol [KDG11], one may think that error should decrease when distance decreases since closer images possess better quality and resolution than images acquired at a far distance. According to those results, this is not the case. We demonstrate in this paper that this is due to the effect of pitch between the camera and the face, which produces errors that directly affect the performance of the system.

In this context the contribution of this paper is to prove the benefits of the use of a projective transformation before comparing mug shot and CCTV images suffering from pitch rotation. This technique leads to frontal images.

The strengths of the proposed technique rely on its simplicity and low computational cost. This technique could be easily used by forensic examiners similar to the work conducted in an *AFIS* (Automated fingerprint identification system [Kom05]). When working with an *AFIS*, forensic examiners first manually mark a set of minutiae points. Then, the *AFIS* system is used to match the feature template associated to this fingerprint sample against a stored database. Likewise, in a hypothetical forensic face recognition case, an examiner could mark a small set of points (e.g. the four points need to define the projective transformation matrix) to ease the task of any face recognition system.

Other related works have tried to compensate general variability sources using probabilistic techniques such as joint factor analysis (JFA) and intersession variability (ISV), reaching promising results [MMM11]. However, our aim with our approach is not to improve the state of the art on face recognition when dealing with general variability sources but to show off the potential of a simple technique to compensate the pitch rotation produced mainly in real forensic caseworks.

This paper is structured as follows. Section 2 presents related work regarding pose compensation and projective transformation. Section 3 describes in depth the SCface database. Section 4 features the preprocessing technique and the projective transformation put forward in this work. Section 5 addresses the experimental protocol followed in our experiments and Section 6 presents the major results obtained in this paper. Finally, Section 7 offers some brief conclusions and future work.

2 Related work

2.1 Pose compensation techniques

Pose compensation techniques are a matter of growing importance within the face recognition community (see [ZG09] for a survey). Different approaches have been proposed to overcome the difficulties of not having a frontal face. There are general techniques that in-

directly address the pose compensation problem. Approaches based on manifolds or deep belief neural networks are some examples.

As reported in the cited survey, there exist other methods that present algorithms designed specifically to compensate the pose either in 2D or in 3D. Regarding the 2D space, active appearance models and procrustes analyses address the alignment of faces through specific keypoints.

The approach presented in [MVN07] consists of creating a mosaic from frontal and semi profile face images. In this manner, they achieve a more representative subject model without the drawback of storing plenty of images.

3D imaging has produced noteworthy improvements in pose compensation. The most remarkable techniques are based on 3D face models, 3D morphable models and stereo matching.

2.2 Projective transformation

Projective transformation has been used in certain applications related to face recognition. The work developed by *Chen and Medioni* [CM01] builds a 3D human face model stemmed from two photographs.

In [HCD14], they manage to estimate the pose of a subject through the projective transformation of the features points of the 3D face model and video sequence. There are also projective transformation-based works to estimate the orientation of the face, useful for human-computer interaction applications [SPD11]. The convenience of using the projective transformation relies on its simplicity.

3 Database

This section describes the subset of the SCface database [KDG11] used in our experiments. SCface is a database of static images of human faces with 4.160 images (visible and infrared spectrum) of 130 subjects.

The dataset used in this paper is divided into 6 different subsets: *i)* mug shot images, which are high resolution frontal images and *ii)* five visible video surveillance cameras (CCTV). The images were acquired in an uncontrolled indoor environment with the people walking towards several video surveillance cameras (with different qualities). Further, the images were acquired at three different distances: 1.00 (Close), 2.60 (Medium) and 4.20 (Far) meters respectively (see Fig. 1).

This database is of particular interest from a forensic point of view because images were acquired using commercially available surveillance equipment and under realistic conditions.

There are several landmarks describing the most discriminative parts of the face: eyes,

nose, mouth, eyebrows, etc. In this work, landmarks acquired manually and automatically were used. To extract the landmarks automatically the commercial SDK Luxand Face 4.0 [1] was used resulting in a set of 13 points. For the manual approach of landmark detection, a set of 21 facial landmarks were manually tagged by a human bearing in mind the procedure followed by a forensic examiner [JF13].

For this study, the 5 available CCTV images per person and per distance (1950 images in total, 3 distances $\times$ 5 cameras $\times$ 130 persons) plus the 130 mug shot images are considered.

4 System description

4.1 Preprocessing

First, we obtain the grayscale version of the image. Then, we equalize the grayscale facial image. The face is normalised following the ISO standard [2] with an interpupilar pixel distance (IPD) of 75 pixels by using the eyes coordinates provided (computed either automatically or manually). This step eliminates variations in translation, scale and rotation in horizontal plane, and provides a normalized face in order to compare with a standard size for all faces considered.

4.2 Projective transformation

The projective transformation (often called homography) models the geometric distortion that is introduced in a plane when an image is taken with a perspective camera. Under a perspective camera, some geometric properties such as linearity are kept, whereas others such as parallelism are not.

A projective transformation is a two-dimensional transformation that maps two set of points that define a quadrilateral and that belong to two different projective planes.

A projective transformation between two planes is represented as a 3×3 matrix acting on homogeneous coordinates of the plane. The general projective transformation H from one projective plane, A, to another, B, is represented as:

$$\begin{bmatrix} b_1 \\ b_2 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ 1 \end{bmatrix} \quad (1)$$

where a_1, a_2, b_1, b_2 are the points of the projective plane A and projective plane B res-

[1] Luxand Face SDK, http://www.luxand.com

[2] ISO/IEC 19794-5:2011, Information Technology - biometric data interchange formats - part 5: Face image data

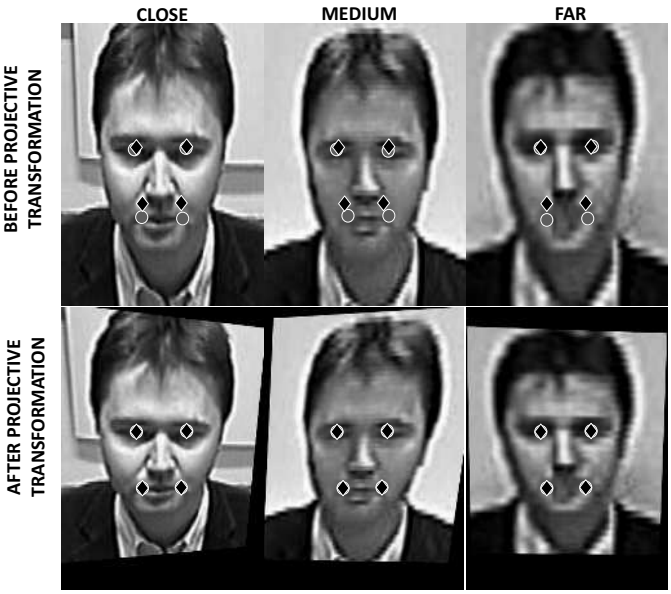

Figure 2: Example of applying the projective transformation to CCTV image from user 1 for all distances. Images presented followed the ISO format. Circles are the landmarks coordinates of the specific image and diamonds are the reference landmarks.

pectively and h_{ij} with $i = 1 : 3$ and $j = 1 : 3$ are the coefficients of the mapping transformation H.

The procedure followed to obtain the projective transformation was the following. First, we selected four landmarks: 2 eyes centres and 2 mouth vertices. We then average each of these landmarks for a set of mug shots, to obtain a general landmark position. A specific transformation for each CCTV image is then obtained to this reference positioning by solving (1) using digital image warping methods (specifically using the quadrilateral to quadrilateral mapping) [3].

Finally, we extracted the region of interest of the face from the projected images. Fig. 2 draws an example of the result of applying the projective transformation to the face image for the three different distances: far, medium and close. In the first row, images from the three different distances are plotted, marking with diamonds the reference coordinates and with circles the coordinates of the specific images. Notice the difference of situation between these two sets of points before the transformation. The projective transformation finds the transformation that maps circles to diamonds. The second row of Fig. 2 plots the resulting image after applying this transformation.

[3]The motivation of using mouth and eyes points relies on the fact that projective transformation works only in planar surfaces and, even though the human face is not planar, we may make the assumption that eyes and mouth points are coplanar.

4.3 Matching

In what concerns the recognition system itself, principal component analysis (PCA) is applied to the face image over the training set considering the first 200 principal components. Similarity scores are computed in this PCA vector space using a Support Vector Machine (SVM) classifier with a linear kernel.

5 Experimental protocol

The database is divided into 3 subsets based on the subject ID: Development (IDS in the range [1-43]), SVM Training (IDS in the range [44-87]), and Test (IDS in the range [88-130]). Each of the sets is comprised of mug shot and CCTV images.

The mug shot versus CCTV images scenario is common in forensic laboratories, and it is very challenging because of the difficulty in finding reliable similarities between probe CCTV images and gallery mug shot images from police records. For this reason, the results obtained in this scenario are especially helpful for the forensic practice.

In this case, each subject model is trained using a single mug shot image (SVM Training Clients) and impostors for the SVM are extracted from the SVM Training set. Then, Test images are taken from the 5 surveillance cameras at 3 different distances: close, medium and far (Test set).

Two different protocols have been defined: *distance-dependent* and *combined protocol*. For the *distance-dependent protocol*, we build the PCA-SVM system for each specific distance: close, medium and far. The analysis of these three configurations is also of great interest for forensics and face biometrics. Additionally, with the *combined protocol* we use jointly all CCTV images regardless of the distance. In this protocol, the PCA matrix transformation is estimated using all images from the three distances belonging to the Development Set. Likewise, the SVM model for each user is modeled having the mug shot image as the positive sample, and the rest of mug shot images and the CCTV images from all distances for the Training set. This latter protocol is more realistic than the *distance-dependent protocol* because a subject to camera distance should be estimated for the other three cases otherwise.

6 Experiments

We empirically proved that the projective transformation based on the coordinates of the eyes and the mouth seems to compensate better the pose deviation of the images. Table 1 compares the results obtained for the test set when applying this chosen projective transformation and the baseline system with the original images using manual landmarks. The relative improvement is 16.16%, 31.44% and 39.58% with respect to the baseline method for far, medium and close images respectively. As far, medium and close images suffer

Table 1: Equal Error Rates (EER in %) of the PCA-SVM system on the test set using **manual landmarks**.

Method	FAR	MEDIUM	CLOSE	COMBINED
No Pose Correct.	28.90	31.20	33.10	32.24
Pose Correct.	24.23	21.39	20.00	21.86

Table 2: Equal Error Rates (EER in %) of the PCA-SVM system on the test set using **automatic landmarks**.

Method	FAR	MEDIUM	CLOSE	COMBINED
No Pose Correct.	35.10	31.20	35.40	34.41
Pose Correct.	32.09	24.65	27.33	28.37

from an average pitch rotation of $7.5°$, $12.0°$, $28.8°$ respectively (for an average height of 1.80 meters), it is straight forward to think that the images with more deviation benefit more from this compensation of pose.

An additional experiment is carried out using automatic landmarks in order to assess the influence of the landmarks detection procedure in conjunction (manual or automatic) with the use of the projective transformation. Table 2 compares the result obtained between the baseline system and the projected images when automatic landmarks are employed. As can be seen, the relative improvement is 8.57%, 20.99% and 22.79% with respect to the baseline method for far, medium and close images respectively. Fig 3 draws the DET curves for the *distance-dependent* and *combined protocol* using manual landmarks.

Comparing results between the transformation defined by manual points and transformations defined by automatic points (Table 1 and Table 2) we conclude that, in both cases, the relative improvement increases when the distance is reduced. However, the improvement is always higher in transformations defined by manual points compared to transformations defined by automatic points. Hence, it may be deduced that the projective transformation is sensitive to the landmarks used. Although automatic points are essential for automatic face recognition systems, they are not so crucial for forensic applications in which the last decision is made by a forensic examiner.

As specified in Section 5, the *combined protocol* is defined with the aim of assessing the influence of this projective transformation in a more realistic scenario in which the distance between the subject and the camera is unknown. The last column of Table 1 refers to this protocol. Specifically, the performance of the original combined system is slightly worse than the average of the three distance-dependent systems. As it was expected, the influence of applying this projective transformation improves also the results of the *combined protocol*, having a relative improvement of 32.19% and 17.51% for manual and automatic points respectively. The average relative improvement for all protocols (far, medium, close and combined) is 29% and 17% for manual and automatic points respectively.

Paying attention now to the results obtained with the images projected, we clearly see

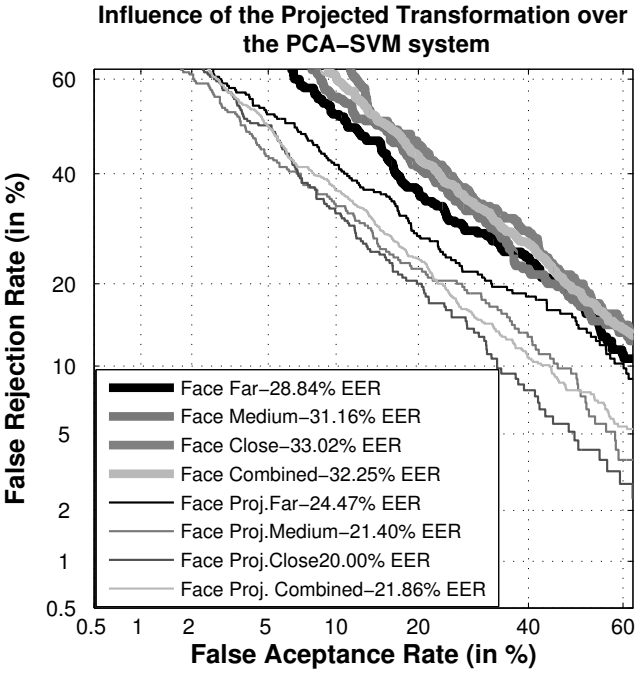

Figure 3: DET curves for the *distance-dependent protocol* and *combined protocol* before and after applying the projective transformation using manual landmarks.

now this improvement of equal error rate when reducing the distance of the subject to the CCTV camera.

7 Conclusions and future work

In this work, the specific case of pose compensation has been analysed. It must be noted that the aim with this approach was not to improve the state of the art on face recognition but to show the potential use of a simple technique to compensate the pitch rotation produced mainly in real forensic caseworks.

The relative improvement of this technique is greater for images that suffer higher pitch rotation, such as close images. Concretely, the application of the projective transformation may result in average relative improvements of 29% or 17% for the case of using manual or automatic points respectively. Hence, results suggest that the projective transformation may be used as a preprocessing stage for compensating pitch rotation of CCTV images, especially when comparing them to mug shot images in forensic scenarios. This projective transformation may be easily applied before using COTS face recognition systems, helping this way to narrow down even more the search of suspects to the forensic examiner.

For future experimental work, we aim to use other types of matching techniques and make comparisons with more general pose compensation techniques.

8 Acknowledgment

This work has been partially supported in part by Bio-Shield (TEC2012-34881) from Spanish MINECO, in part by BEAT (FP7-SEC-284989) from EU and in part by Cátedra UAM-Telefónica. E. Gonzalez-Sosa is supported by a PhD scholarship from Universidad Autonoma de Madrid.

References

[ARP04] A. Jain A. Ross and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2004.

[CM01] Qian Chen and Grard Medioni. Building 3-D Human Face Models from Two Photographs. *Journal of VLSI signal processing systems for signal, image and video technology*, 27(1-2):127–140, 2001.

[HCD14] C.Liu H. Cheng and A. Dasu. Scale Robust Head Pose Estimation Based on Relative Homography Transformation. *New Mathematics and Natural Computation*, 2014.

[JF13] R. Vera-Rodriguez F.J. Vega J Fierrez, N. Exposito. Analysis of the variability of facial landmarks in a forensic scenario. In *Proc. of IWBF*, 2013.

[KDG11] M. Grgic K. Delac and S. Grgic. SCface–surveillance cameras face database. *Multimedia tools and applications*, 2011.

[KJ13] J. Klontz and A. Jain. A Case Study on Unconstrained Facial Recognition Using the Boston Marathon Bombings Suspects. *Michigan State University, Tech. Rep*, 2013.

[Kom05] Peter Komarinski. *Automated fingerprint identification systems (AFIS)*. Academic Press, 2005.

[MMM11] R. Wallace M. McLaren, C.McCool and S. Marcel. Inter-session variability modelling and joint factor analysis for face authentication. In *Proc. of IJCB*, 2011.

[MVN07] R. Singh M. Vatsa, A. Ross and A. Noore. A mosaicing scheme for pose-invariant face recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 2007.

[PPB13] J. Beveridge P.J Phillips and D.S. Bolme. The challenge of face recognition from digital point-and-shoot cameras. In *Proc. of BTAS*, 2013.

[RVRR13] P. Tome R. Vera-Rodriguez, J. Fierrez and D. Ramos. Identification using face regions: Application and assessment in forensic scenarios. *Forensic science international*, 2013.

[SPD11] P. Petrov O. Boumbarov S. Panev, I. Paliy and L. Dimitrov. Homography-based face orientation determination from a fixed monocular camera. In *Proc. of IDAACS*. IEEE, 2011.

[YST14] X. Wang Yi Sun and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. of CVPR*, 2014.

[ZG09] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern Recognition*, 2009.