Asymmetric 3D/2D Processing: A Novel Approach for Face Recognition

Daniel Riccio[1] and Jean-Luc Dugelay[2]

[1] Universitá di Salerno, via Ponte Don Melillo, 84084 Fisciano, Salerno, Italy
driccio@unisa.it
[2] Institut Eurecom, CMM, 2229 route des Cretes,
B.P. 193, F-06904, Sophia Antipolis, Cedex
Jean-Luc.Dugelay@eurecom.fr

Abstract. Facial image analysis is very useful in many applications such as video compression, talking heads, or biometrics. During the last few years, many algorithms have been proposed in particular for face recognition using classical 2-D images. Face is fairly easy to use and well accepted by people but generally not robust enough to be used in most practical security applications because too sensitive to variations in pose and illumination. One possibility to overcome this limitation is to work in 3-D instead of 2-D. But 3-D is costly and more difficult to manipulate and then ineffective to authenticate people in most contexts. Hence, to solve this problem, we propose a novel face recognition approach that is based on an asymmetric protocol: enrolment in 3-D but identification performed from 2-D images. So that, the goal is to make more robust face recognition while keeping the system practical. To make this 3-D/2-D approach possible, we introduce geometric invariants used in computer vision within the context of face recognition. We report preliminary experiments to evaluate robustness of invariants according to pose variations and to the accuracy of detection of facial feature points. Preliminary results obtained in terms of identification rate are encouraging.

1 Introduction

Biometric technologies that currently offer greater accuracy such as iris and fingerprint, require, however, much greater cooperation from the user and are too much invasive in some cases. Face Recognition includes a good compromise between people acceptance and reliability (in controlled environments). In the last years, many strategies have been proposed in order to solve the recognition problem, mainly addressing problems such as changes in expression, pose and illumination. Recent works attempt to solve the problem directly on a 3D model of the face. Indeed, a 3D model provides more geometrical information on the shape of the face and is unaffected by illumination and pose variation. The development of 3D acquisition systems and then the 3D capturing process are becoming cheaper and faster too. This definitely makes the 3D approach more and more applicable to real situations out of the laboratories. However, unlike 2D face recognition, there are yet few works on 3D range images.

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 986–993, 2005.

In [3] Huang *et al.* develop a component-based recognition method based on a 3D morphable model. At first the 3D model is generated from two different views of the subject and then a number of synthetic views are rendered from the model changing pose and illumination. Yet the database consist of so few people (6 subjects) and training/testing images are generated from the same models. Bronstein *et al.* in [2] suggest the use of a canonical form, which consists of an isosurface of the face shape, with the flattened texture mapped on, and where principal component analysis is used to decompose the obtained canonical image. A method, that works on 2D, but using however a 3D morphable model for training/testing is shown in [1]. Despite of its performances in terms of recognition rate, the greatest drawback of this approach is its computational cost.

The most part of the proposed method applies only to the 3D range images, but even if the 3D acquisition is becoming cheaper, the problem of the sensitivity of the capturing process remains. This point out the usefulness of an approach, that profits by a 3D model based enrolment, but only needs of a 2D view of the model for testing. This is one of the main motivations for which a new framework for 3D/2D face recognition is introduced here. The proposed approach is based on 3D projective invariants, used for long time in computer vision, recognizing object with rigid surface. As the capacity of recognizing objects in a scene, regardless of their orientation, is an important goal in the computer vision from long time, several relevant papers have been published on this topic. They describe a lot of measures or ratio of distances that are invariant with respect to projective and/or perspective transformations. There are no geometric invariants for an unconstrained set of points in the space. However, interesting properties have been inferred when points in the 3D space are collinear, coplanar or their structure in the space is well described in a given way.

1.1 Geometric Invariants in Face Recognition

Given *inhomogeneous* coordinates $x = \left(x^1, x^2, \ldots, x^m\right)^t$, where $x \in R^m$, the correspondent *homogeneous* coordinates of the point are $z = \left(z^1, z^2, \ldots, z^m, z^{m+1}\right)^t$, where $x^l = z^l / z^{m+1},\ l = 1, \ldots, m,\ z^{m+1} \neq 0$. The homogeneous coordinates are a more general way to represent points, requiring the constraint that $\exists l \in \{1, \ldots, (m+1)\} \ni z^l \neq 0$. By means of this mapping, the projective transformation in R^m, can be easily managed as linear transformation in R^{m+1}. Thanks to this representation, the most of the ratios among distances in the space can be represented as ratio of determinants of the corresponding point coordinates.

The are two main categories of invariants. The first category, namely 2D image based invariants, does not require the 3-D object to be computed but constraints about the localization of feature points are important. On the contrary, the second category, namely 3D image based invariants, requires the 3-D object or at least 3 points of view but is very flexible about the repartition of the anchor points. Besides, to extract the invariants from the 3D model rather than from a given set of images is advantageous for two main reasons: 1) in some of the available images, the points to be selected could be occluded, while with

the 3D model, any point configuration can be considered; 2) On a set of views the points must be selected and their value is then affected by the localization error. On the contrary, on the 3D model a correct calculation of the invariants can be performed.

Given four collinear points $z_1, z_2, z_3, z_4 \in R^2$, the simplest invariant is their cross ratio that can be written as:

$$c(z_1, z_2, z_3, z_4) = \frac{M(1,3) \cdot M(2,4)}{M(1,4) \cdot M(2,3)} \text{ with } M(i,j) = \begin{vmatrix} x_i & x_j \\ 1 & 1 \end{vmatrix}. \quad (1)$$

This property can be extended to five points, which lie on the same plane $z_i \in R^3/(0,0,0), i = 1, \ldots, 5$, so that two functionally independent projective invariants hold:

$$c_1 = \frac{M(1,2,4) \cdot M(1,3,5)}{M(1,2,5) \cdot M(1,3,4)} \text{ and } c_2 = \frac{M(2,1,4) \cdot M(2,3,5)}{M(2,1,5) \cdot M(2,3,4)}. \quad (2)$$

At last Zhu *et al.* [6] demonstrated that given six point in a 3D space A, B, C, D, E and F, which lie on two adjacent planes, as shown in Fig. 1 (c), the cross-ratio of the areas of the corresponding triangles is a projective invariant, if no three of each set of four coplanar points are collinear:

$$I = \frac{P_{ABD} \cdot P_{FEC}}{P_{ABC} \cdot P_{FED}} \quad (3)$$

It is important to note that the goal of previous works using 3-D model based invariants was to discriminate objects that are rigid and obviously different.

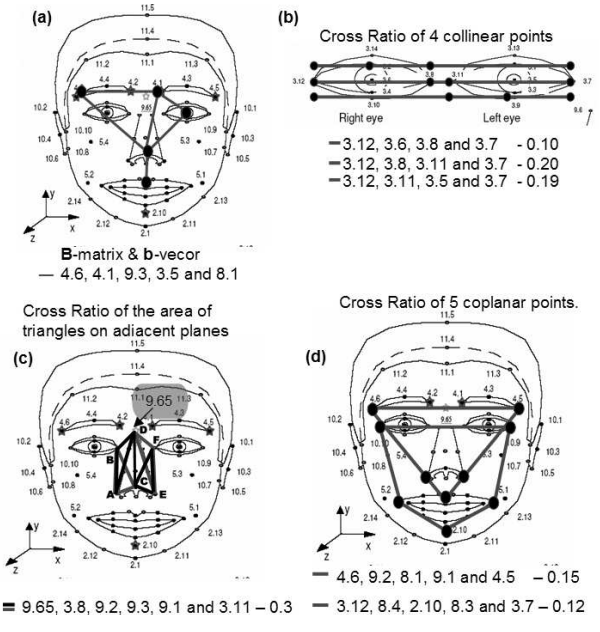

Fig. 1. Graphic representation of the control points and corresponding invariants

This is not the case for faces that are flexible and similar from one person to another one, so that even slight perturbations on the control points can result in a misclassification of the corresponding face. Therefore, a very important feature for a 3D model based invariant is then its sensitiveness to the noise on the control points. For this reason one of the more noise-insensitive invariant has been chosen. It has been proposed by Weinshall in [5].

Given an object in the 3D space, its representation is closely related to the reference frame. However choosing three points on the object, corresponding to three linear independent vector p_i, p_j, p_k, a new reference system can be defined, which make the representation of the object invariant to projective transformations. Every point $p_l \in R^3$ on the object can be written as a linear combination of this basis: $p_l = b_1^l p_i + b_2^l p_j + b_3^l p_k$. The vector $b^l = \left(b_1^l, b_2^l, b_3^l\right)$ represents an affine invariant. The Euclidean metric information on the basis points is represented by means of their inverse Gramian matrix $B = G^{-1}$.

In [5] the author also proposed two different kind of invariants, defined by means of a set of four/five non coplanar points and the inverse Gramian matrix B. Indeed, consider an object composed of four non-coplanar 3D points, where $\{P_l\}_{l=0}^3$ denote the 3D coordinates of the four points in some reference frame. Assume $P_0 = (0, 0, 0)$ without loss of generality and let $\{p_l\}_{l=1}^3$ denote the 3D vectors corresponding to the three remaining points. Given the image coordinates of the four points $(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3)$, where $x_0 = 0, y_0 = 0$ is assumed. Let $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$. The first rigid invariant is given by the equation 4, while retaining the same notation, the second invariant can be defined, combining the vectors of the basis $b = (b_1, b_2, b_3)$ and the image coordinates of five points, as shown by the equation 5.

$$f_B = \frac{\left|x^T B y\right| + \left|x^T B x - y^T B y\right|}{|x| \, \|B\| \, |y|} \tag{4}$$

$$f_b = \frac{\left|x_4 - \sum_{i=1}^3 b_i x_i\right|}{|x| \, |b|} + \frac{\left|y_4 - \sum_{i=1}^3 b_i y_i\right|}{|y| \, |b|}. \tag{5}$$

The value of the functions f_B and f_b is zero for all the views of the object, that the matrix B or the vector b describe. They are normalized by means of the norm of the vectors x, y and the matrix B or the vector b respectively, so that their value does not depend on the distance between the object and the camera.

2 The Proposed Approach

Works dealing with 3D invariants have been devoted to rigid objects, but face is flexible. This represents the first problem to be solved, when applying invariants in 3D face recognition. Indeed changes in expression modify the geometry of the face, more than anything else, jeopardizing the results. This point out the significance of choosing carefully the points, namely *control points*, used in next computations for the extraction of the invariants.

2.1 Feature Extraction

The 19 control points have been chosen as a subset of the Feature Points defined in MPEG-4 standard [4]. Fig. 1 highlights that the point 9.65 is not present in the MPEG-4 standard, but it has been inserted, so that two adjacent planes can be made up. Almost all the considered invariants impose some hypotheses on the configuration of the control points, such as collinearity or coplanarity. However the face is not a rigid surface and to find control points which both respect the required hypotheses and that are easy to locate, turns in a difficult task. Therefore it makes sense that the required conditions are tested for each candidate configuration, in order to assure a real good approximation of the theory. For the 2D image based invariants, that is all the ratios, the collinearity has been assessed computing the approximation error by means of a linear regression. In the same way, in order to test the coplanarity of each configuration of five points, the approximation error has been computed by means of a plane regression. Then the smaller the approximation error is, better the quality of the chosen configuration will be.

In Fig. 1 are reported the most part of the 2D image based invariants that have been used and the corresponding approximation error, computed on a large set of 3D face models. On the contrary, for the 3D model based invariants the only need is the non collinearity and/or non coplanarity of the control points, which is easy to achieve properly choosing each configuration. In order to optimize the choice of the control point configurations, opting for those providing the greatest discriminating power, the distribution of the control points with respect to their average position has been investigated. The models have been lined up with respect to the nose tip and the centroid has been drawn out for each of the 19 class of controls points, further the standard deviation of each class has been calculated with respect to its centroid. According to these results, for the f_B and f_b invariants eight configurations of four and five points have been chosen respectively.

2.2 Classification Process: Enrolment/Testing

When a new user has to be enrolled, the system acquires both the 3D shape and the 2D texture of his/her face. The control points are then located on the 2D texture of the face, while the corresponding 3D points are automatically retrieved on the 3D shape. All the 2D image based invariants (all the ratios) are computed by the (x, y) coordinates of the control points. They consist in scalar values (ratio of distances), so they can be stored in the first part of the feature vector. On the contrary, for each 3D model based invariant f_B the B matrix is computed and its $B_{i,j}$ items are then inserted in the second part of the feature vector. At last, the b vector is computed for each of the f_b invariant and the corresponding b_i values are stored in the last part of the feature vector.

The testing process is partitioned in two steps, in order to make this task efficient as well as effective. Let be F a query image submitted to the system. First of all, the 19 control points are located on F. Some of them are used to

compute all the cross ratios, as described in Section 1.1, forming the first part of the feature vector V_F for the face F. This vector is then used to query the system in order to retrieve a subset of only K of the N subjects in the database, which have to be further authenticated by the f_B and f_b invariants, according to a voting strategy, that is for each of the eight configuration of the control points, the corresponding f_B and f_b invariants votes for one of the K retrieved subjects, and that one receiving the most of votes is returned as the correct identity.

In other words, the proposed method performs in two sequential steps. The former is a pruning operation on the database, resulting a subset of the face database that retains best candidates, while the latter consists in the real identification task, in which the retrieved subjects are identified by means of the 3D model based invariants. In general, to reduce the number of the subjects to be identified allows a noticeable drop in the computational cost. Indeed, in this case the feature vectors are organized in a structured manner, such as a tree, then a subset of K good candidates can be retrieved by the screening operation in time $O(logN)$ and identified by the 3D model based invariants in time $O(K)$, intead of $O(N)$ of the full identification.

3 Experimental Results

Since the invariants are calculated only from the control points, which are detected on the image by hand at the moment, they are naturally robust against the illumination variations. The main problem that is faced in the experiments is therefore the sensitiveness of the algorithm with respect to the pose variations and inaccuracy of detection of the control points. The proposed method has been tested on a property database realized by Eurecom.

All the faces have been acquired by means of the Geometrix system [7], which uses two cameras (up and down), in order to extract the 3D shape of the face. The database consists of 50 people acquired in normal conditions of expression, pose and illumination. The age of the subjects ranges between 20 and 50 years, 40 of them are male and 10 females, further the most are Caucasians. Three models for each subject have been considered.

In the first experiment the goal is to investigate how the discriminant power of the 2D image based invariants drops respect to the parameter K and the pose changes. The database has been divided in two subsets, a probe and a gallery set, respectively. For each subject, one model has been inserted in the probe set and the remaining two in the gallery. In this way a manual localization of the control points is properly simulated. Furthermore K represents a tuning parameter for the system, ranging between 1 and N. In this case $K \in [1, 100]$, two models for each of the 50 subjects. Notice that all the models in the gallery are considered distinct. The results of this experiment are reported in Fig. 2 (a).

In the second experiment 50 models have been considered, in order to assess the performances of the system respect to the accuracy of locating the control points. The models in the probe and gallery set are the same, so there is no initial error on the control points, while K is fixed to 20. Five different pose

have been considered, while increasing noise is added to the coordinates of the control points. The noise is generated randomly in the range $[-2e_{rr}, 2e_{rr}]$ with mean $e_{rr} = 0.5, 1, ..., 5$ (results for $e_{rr} = 0$ has not been reported because they are all ones). In Particular, e_{rr} represents the pixel accuracy of locating the control points, with respect to a 256×256 image. The results are shown in Fig. 2 (b). The precision of the control point localization has been also drawn out on the models of the real database, measuring the mean error between the models in the probe and the corresponding ones in the gallery. An error of about 3.2 pixels has been estimated. Indeed, results marked with the ellipses confirm that when feature points are manually selected (ellipse in Fig. 2 (a)), the average error in localization is somewhat equivalent to a additional noise of 3 pixels in exact but artificial conditions (ellipse in Fig. 2 (b)).

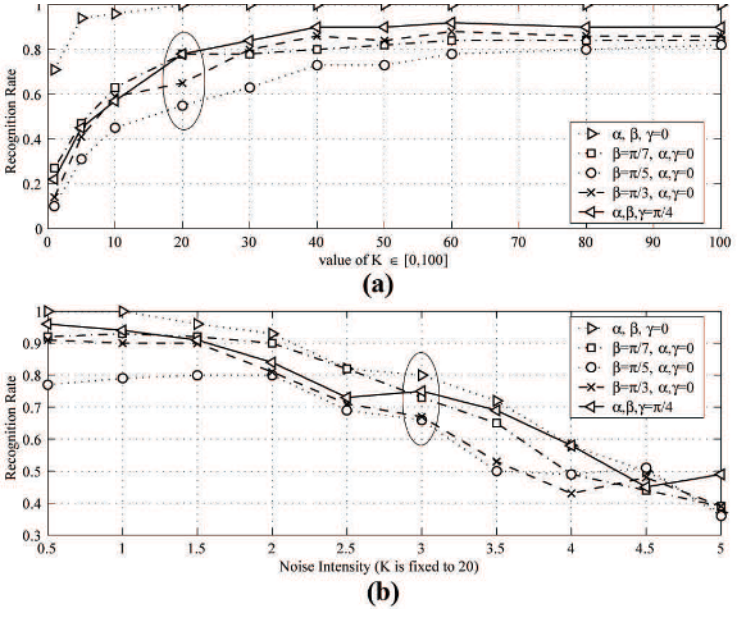

Fig. 2. (a) Recognition rate of the system, varing pose and the K parameter. (b) The probability that the correct subject is the first answer or is in the first 5 answers, when increasing random noise is added and $K = 20$.

The results of the first experiment suggest that the value of K must be proportional to the magnitude of pose variation, which can be estimated from the distribution of the control points. Therefore it makes sense that gometric invariants could play a role in a multimodal face recognition system, which also takes into account for the information provided by the texture image. At last both the experiments underline that 3D model based invariants are more powerful of the 2D image based geometric invariants, taking an interest in a more comprehensive study in this sense.

4 Conclusion and Remarks

An asymmetrical 3D/2D face recognition technique has been introduced. It is based on geometric invariants [5, 6], studied for the pose invariant object recognition problems for a long time. The crucial problem of choosing the control points, in case of faces, for the 2D image and 3D model based invariants has also been addressed.

The experiments have finally been made in order to assess the robustness of the method respect to the changes in pose and to the accuracy of locating the controls points. The results are encouraging in terms of recognition rate. Further works, can study the use of some other invariants, comparing their discriminating power and also integrating the texture information of faces within a multimodal framework.

References

1. Volker Blanz and Thomas Vetter, Face Recognition Based on Fitting a 3D Morphable Model. In IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pp. 1191–1202, 2002.
2. M. Bronstein, Michael M. Bronstein, and Ron Kimmel, Expression Invariant 3D Face Recognition. In Proc. of Audio & Video-based Biometric Person Authentication (AVBPA), Lecture Notes in Computer Science 2688, Springer, pp. 62–69, 2003
3. Jennifer Huang, Volker Blanz, and Bernd Heisele, Face Recognition with Support Vector Macines and 3D Head Models. In First International Workshop on Pattern Recognition with Support Vector Machines (SVM 2002), pp. 334–341, 2002.
4. Fabio Lavagetto, Roberto Pockaj, The Facial Animation Engine: Toward a High-Level Interface for the Design of MPEG-4 Compliant Animated Faces, In IEEE Trans. on Circuits and Systems for Video Technology, vol. 2, no.2, march 1999.
5. Daphna Weinshall, Model-based invariants for 3D Vision. In International Journal of Computer Vision, vol. 10, no. 1, pp. 27–42, 1993.
6. Y.Zhu, L. D. Seneviratne and S. W. E.Earles, A New Structure of Invariant for 3D Point Sets from A single View., In IEEE International Conference on Robotics and Automation, pp. 1726–1731, May 1995.
7. Geometrix, Introducing FaceVision-The New Shape of Human Identification, http://www.geometrix.com/, 13 February 2005