Robust Face Recognition Using 2D and 3D Data: Pose and Illumination Compensation

Sotiris Malassiotis and Michael G. Strintzis

Informatics & Telematics Institute,

1st Km Thermi-Panorama Rd. 57001 Thessaloniki, Greece,

Corresponding Author: S. Malassiotis, e-mail: malasiot@iti.gr, Fax: +30-2310-464164

Abstract

The paper addresses the problem of face recognition under varying pose and illumination. Robustness to appearance variations is achieved not only by using a combination of a 2D color and a 3D image of the face, but mainly by using face geometry information to cope with pose and illumination variations that inhibit the performance of 2D face recognition. A face normalization approach is proposed, which unlike state-of-the-art techniques is computationally efficient and does not require an extended training set. Experimental results on a large data set show that template-based face recognition performance is significantly benefited from the application of the proposed normalization algorithms prior to classification.

Key words: face recognition, range data, pose, illumination

1 Introduction

The increasing use of biometric technologies in high-security applications and beyond has stressed the requirement for highly dependable face recognition systems. Recent benchmarks [1] demonstrated that the accuracy of state-of-the-art algorithms (see [2] for a recent survey) is fairly high under constrained conditions, but degrades significantly for images exhibiting pose, illumination and facial expression variations. Current research efforts strive to achieve insensitivity to such variations following three main directions: a) introduction of novel classification and similarity measurement techniques, b) compensation of appearance variations, and c) augmentation of existing systems with additional modalities that are insensitive to these variations.

The paper describes and evaluates a complete face recognition system using a combination of 2D color and 3D range images captured in real-time. The main focus of the paper is to exploit 3D information to cope with pose and illumination variations. We present several novel techniques, which are capable, taking as input a pair of 2D and 3D images, to produce a pair of normalized images depicting frontal pose and illumination. The efficiency and robustness of the proposed system is demonstrated on a data set of significant size and is compared with state-of-the-art compensation techniques. On the following section we review the related work in this field highlighting the novelties of the proposed work. The employed 3D acquisition setup is briefly described in section 3. Then, pose and illumination compensation algorithms are presented in section 4 and 5 respectively. The performance of the algorithms is evaluated with extensive experiments in section 6 while section 7 concludes the paper.

2 Previous Work

2.1 3D face recognition

Although the 3D structure of the human face intuitively provides highly discriminatory information and is insensitive to environmental conditions, only a few techniques have been proposed employing range images. This is mainly due to the high cost of available 3D digitizers that makes their use prohibitive in real-world applications. Furthermore, these devices often do not operate in real time (e.g. time of flight laser scanners) or produce inaccurate depth information (e.g. stereo vision). The work presented in this paper is partly motivated by the recent development of low cost 3D sensors that are capable of real-time 3D acquisition [3].

A common approach adopted towards 3D face recognition is based on the extraction of 3D facial features by means of differential geometry techniques. Facial features invariant to rigid transformations of the face may be detected using surface curvature measures [4], that are subsequently used to extract higher level facial features [5–7]. Recently Point Signatures have been proposed for 3D surface feature extraction and have also been applied for 3D face recognition [8, 9]. A few techniques [9,10] employ grayscale images as well, mainly for the detection of features such as the eyes which are harder to detect on the range image. The most important argument against techniques using a feature-based approach is that they rely on accurate 3D maps of faces, usually extracted by expensive off-line 3D scanners. In practice, 3D digitizers produce range images containing noise and missing pixels (e.g. over areas with low reflection). The applicability of feature based approaches when using such data is questionable, especially if computation of curvature information is involved. Also, the computational cost associated with the extraction of

the features (e.g. curvatures) is significantly high. This hinders the application of such techniques in real-world security systems. The recognition rates claimed by the above techniques were estimated using databases of limited size and without significant variations of the faces. Only recently [11] an experiment was conducted with a database of significant size (275 persons) containing both grayscale and range images, and it produced comparative results of face identification using 2D and 3D eigenfaces and their combination, for varying image quality. This test however handled only frontal images captured under constant illumination conditions.

In a recent work by the authors [12] a combination of 2D color and 3D range images for face recognition was examined. A database of significant size containing several variations (pose, illumination, expressions, glasses, several recording sessions) was recorded and template based face classification was tested using 2D, 3D images and their combination. The face database was enriched with artificially generated examples depicting variations in pose and illumination. Although, considerable gains were achieved using the above scheme the overall performance of the system did not seem to be comparable with that of state-of-the-art 2D systems. This was attributed to the fact that template-based techniques require that the probe and gallery images be well aligned with each other. To avoid this problem in the present work a completely different approach is adopted where appearance variations are compensated before reaching the classifier. Also, in this paper face detection and localization is based on 3D information only while in [12] it relied also on the intensity image.

2.2 Pose Compensation

Several techniques have been proposed to recognize faces under pose variations. In [13] an active appearance model is proposed using a generic face model iteratively fitted to the input image. The estimated control parameters (excluding those associated with pose) are subsequently used for face classification. The idea is extended in [14] employing a high quality deformable model generated from a large number of pixel aligned 3D scans and corresponding texture images. An analysis-by-synthesis technique is used to fit the model to the input image. Another approach is through the generation of novel views resembling arbitrary poses from a previously stored gallery image. Classification is subsequently based on the similarity between the probe image and the generated view. [15] uses optical flow to estimate the deformation between the probe image and the corresponding frontal view gallery image. In [16] a face normalization procedure is described based on the detection of facial features such as the eyes and mouth and adaptation of a generic 3D model. An approach quite different from the above is based on building a pose varying eigenspace using several images of each person. Representative techniques are the view-based subspace of [17] and the predictive characterized subspace of [18].

The main difference of our approach from the above techniques is that pose estimation and compensation is based solely on 3D images and thus exhibits robustness under varying illumination and occlusion of facial features (beard and hair, glasses, etc.). Another advantage stemming from the availability of 3D data is that geometric artifacts which may be present after image warping are avoided. Comparing with state-of-the-art techniques such as [14] the proposed system is seen to be more robust (less than 1% failure rate for over 6000 images, while in [14] it is more than 10% for frontal images only) and computationally efficient (0.5 sec

compared with 40 minutes reported in [14] for the same hardware). Also, unlike subspace techniques, the proposed approach requires only a single gallery image for each person.

2.3 Illumination Compensation

Recent public face recognition benchmarks [1, 19, 20] demonstrated the seriously adverse effect of varying illumination in the performance of commercial face recognition systems. At the same time, the problem of coping with illumination variations is increasingly appreciated by the scientific community and significant progress has been achieved. Several techniques have been proposed in this area that may be roughly classified into two main categories.

The first category contains techniques seeking illumination insensitive representations of face images. Although it has been proven [21] that strictly speaking illumination invariants do not exist for Lambertian surfaces, several representations were seen to be relatively insensitive to illumination variability. For example the direction of the image gradient was recently [21] chosen to provide illumination insensitive representation, and similarly the sum of gradient of ratios between probe and gallery images for an illumination insensitive similarity measure [22]. [23] proposes the design of illumination insensitive correlation filters using a set of training images and varying illumination conditions. A similarity measure based on the difference between “prototype” images extracted from the corresponding probe and gallery images is proposed in [24], where the symmetry of the human face is also exploited. Recently [25] proposed the ”Quotient image” as an illumination insensitive representation computed from the original image and a bootstrapping image set.

The second approach relies on the development of generative appearance models, able to reconstruct novel gallery images resembling the illumination in the probe images. Exploiting the low dimensionality of the image space under varying illumination conditions and the Lambertian assumption [26] some of these techniques utilize a series of example images of the same person to reconstruct novel images [27]. Other approaches utilize a 3D range image and albedo map of the person's face to render novel images under arbitrary illumination [28], while others are based on a combination of the above [29, 30]. The shortcomings of this approach are a) the requirement in practice of large example sets to achieve good reconstructions, b) the increase in the dimensionality of the classification problem by the introduction of the illumination variability, c) the requirement of pixel wise alignment between probe and gallery images which necessitates pose compensation and d) the reliance to simplified reflectance models (e.g. convex surfaces, shadow free, Lambertian reflectance).

In parallel to these efforts a significant progress has been achieved by the Computer Graphics community for realistic image based rendering and relighting of faces and estimation of the reflectance properties of faces [31–33]. Some of these results have inspired computer vision work on illumination compensation.

The approach proposed in this paper may be classified in this latter category, since we try to relight the probe image so that it resembles the illumination in gallery images which is the inverse of the generative approach. The idea of face relighting is also proposed in [24, 31]. In [31] the illumination distribution of the scene is modelled as a linear combination of basis images computed on-line from the input range image, while we employ a simpler and much more efficient light source estimation algorithm using example-base regression. In [24] a ratio image that is indepenedent of the surface albedo is computed from the input image using the facial symmetry

assumption. This is subsequently used for light source estimation and generation of a symmetric frontally illuminated prototype image. In our approach the restrictive symmetry assumption is avoided, and a more efficient light source estimation technique is proposed exploiting the availability of 3D data.

3 Acquisition of 3D Face Data

The proposed system is based on real-time quasi-synchronous color and 3D image acquisition based on the color structrured-light approach [3]. The sensor is based on low cost devices, an off-the-shelf CCTV-color camera and a standard slide projector. In our experiments the system was optimized for an access control application scenario. For subjects located about one meter from the camera, and an effective working space of $60cm \times 50cm \times 50cm$ the average depth accuracy is about $0.5mm$. The spatial resolution of the range images is equal to the color camera resolution in the one direction while in the other direction it is dependent on the width of the color stripes of the projected light pattern (see fig. 1) and the bandwidth of the surface signal. For a low-bandwidth surface such as the face the resolution is close to the resolution of the color camera.

[Fig. 1 about here.]

Apart from image noise the acquired range images contain areas where no depth values was assigned. These are mainly over areas that cannot be reached by the projected light and/or are highly refractive (e.g. eye-glasses) or surfaces with low reflectivity (e.g. hair, beard). Techniques for dealing with missing pixels are discussed in the following sections.

Using the above setup two face databases were compiled following guidelines similar to those of FERET [19] and FRVT 2000 [20]. That is for each subject several images depicting different appearance variations were recorded: three facial expressions, three types of illumination (left/right side spot lights and overhead light), two pose variations (± 20 degrees), two images with and without glasses, and three frontal images.

The first database, recorded using a preliminary prototype of the 3D sensor, consists of 2500 images of 50 persons recorded in 5 sessions (with time lapse between recordings more than one week) and contains range images of low quality both in terms of accuracy and resolution but also in terms of the amount of missing pixels especially on the sides of the face and nose. The second face database contains 20 persons and was compiled in 2 recording sessions with time lapse between the recordings more than 10 days (2200 image pairs). Depth images in this data set are of improved quality with less noise and missing pixels.

4 Pose Compensation

The majority of appearance based face recognition techniques rely on rough alignment between facial features in the probe and corresponding gallery images to achieve high classification accuracy.

The aim of the pose compensation algorithm described in this section is to generate, given an arbitrary pair of color and depth images, novel corresponding color and depth images depicting a frontal, upright face orientation. Also the center of the face in the input image is aligned with the center of the face in the gallery images of the same person with single-pixel accuracy.

Many commercial face recognition systems also apply this step through the localization of prominent facial features such as the eyes and mouth on the input intensity image and geometric normalization. The proposed technique uses the range image only for face detection and pose estimation and therefore is expected to be more robust especially under difficult pose and illumination conditions. Indeed, the experimental results demonstrate that the pose compensation algorithm achieves equivalent or even better results than those obtained using manually specified feature points.

The detection of the face in the image is the first step of the algorithm. We shall rely on the assumption that there is single face in front of the camera inside the working volume of the sensor. This is a reasonable assumption since in practice this working volume is too tight to accommodate more than one person. Availability of 3D data makes also background segmentation straightforward. Segmentation of the head from the body relies on statistical modelling of the head - torso points using a mixture of Gaussians assumption. The parameters of the model are then estimated by means of the Expectation Maximization algorithm and by the incorporation of a-priori constraints on the relative dimensions of the body parts, as described in detail in [34]. The result of the above face detection procedure is an estimate of the center and orientation of the 3D blob corresponding to the head, which may be used to define an oriented bounding box on the image plane containing the face. We tested the algorithm on 5000 images and after visual inspection of the results only 3 cases of failure were found.

The estimation of 3D head pose is described in the following. In [34] we have proposed detecting the nose tip, assuming that this is the closest point to the camera, and use an appearance-based tracking approach to estimate 3D head orientation. Improved results in terms of robustness and accuracy were achieved in [35] which

also focused in real-time 3D face tracking. For the sake of completeness a brief overview of the algorithm in [35] is given below also highlighting a few improvements achieved after relaxing the real-time constraint.

[Fig. 2 about here.]

The localization of the nose tip is achieved using two additional constraints apart from closeness to the camera. First a set of candidate points is defined that includes all local depth minima inside the face bounding box. The first validation test exploits the fact that surface points around the nose tip define an approximately semi-spherical surface pattern which may be detected by means of pose invariant surface curvature measurement. The principal curvatures of the surface are computed over candidate points and thresholded to exclude invalid points. A second test is subsequently applied based on the observation that the nose tip lies on the bottom of the nose ridge. Since points lying on the nose ridge are characterized by high curvature perpendicular to the ridge direction, a suitable measure may be defined by:

$$C(\mathbf{p}) = \min_{\alpha} \sum_{\mathbf{q} \in L(\mathbf{p},\alpha)} K(\mathbf{q})^2 \tag{1}$$

where $L(\mathbf{p}, \alpha)$ is the set of points on the half-line starting from $\mathbf{p}$ with direction α and a constant length (equal to the average nose length in the image database) and $K(\mathbf{q})$ is the ratio of curvatures along and across L. The above minimization is performed by exhaustive search in a small set of orientation angles. Normally, small values of $C(\mathbf{p})$ indicate candidate points on the nose ridge while the rest are easily thresholded out. From the remaining candidate points the one with smallest depth value is selected. The algorithm was tested on both face databases and its failure rate was 1.5% for the first database (in the majority of these cases half or more of the pixels on the nose were missing) and 0.01% on the second database.

After the tip of the nose is localized a 3D line is fitted on the 3D coordinates of pixels on the ridge of the nose ($L(\mathbf{p}, \alpha)$ in eq. (1)). This 3D line Λ defines two of the three degrees of freedom of the face orientation. The third degree of freedom, that is the rotation angle ϕ around Λ, is then estimated by exploiting the approximate bilateral symmetry of the face. Given ϕ a cutting plane Π separating the face in two halves is defined. Then, the sought ϕ value is the one that makes these two halves bilaterally symmetric with respect to Π, and is estimated using an efficient hierarchical exhaustive search algorithm [35].

To test the accuracy of head pose estimation, several frontal images of faces were used to artificially generate novel views depicting ground truth 3D head pose (range of rotation angles ± 20 degrees, 2500 images in total). Then the head pose estimation algorithm was applied and the estimation error was measured. For the first database the average error was about 4 degrees while for the second database the error was less than 2 degrees.

Once the tip of the nose and the pose of the face have been estimated, a 3D coordinate frame aligned with the face is defined centered on the tip of the nose. A warping procedure is subsequently applied on the input depth image to align this local coordinate frame with a reference coordinate frame, thus bringing the face in upright orientation. This reference frame is estimated during the training phase once for each person in the database, using one of the pose compensated training images I_r, as we shall describe later. The transformation between the local and reference coordinate frames is further refined to pixel accuracy by applying the ICP [36] surface registration algorithm between the warped and the reference depth image I_r. Note that the dimensions of the warped image (100×120) are much smaller than the dimensions of the original images thus this final step is very efficient.

The rectified depth image contains missing pixel values that are interpolated using a series of steps. Some of the missing values are determined simply by copying corresponding symmetric pixel values from the other side of the face. In practice, these pixels belong to the sides of the face, thus inherent facial asymmetry normally present around the eyes and mouth is not violated. Remaining missing pixel values are linearly interpolated from neighboring points. Since the dimension and shape of the holes varies this is achieved by means of a 2D Delaunay triangulation.

In Fig. 2 the various steps of the surface interpolation procedure are illustrated. To compute the corresponding rectified color image, the inverse 3D transformation is first applied to the interpolated depth map. Then, the forward 3D transformation is applied to the color image using the back projected interpolated depth values.

For the training phase a similar but simpler pose compensation algorithm is applied. To obtain optimal results the pose of the face is estimated by manually selecting three points on the input image defining a local 3D coordinate frame. Then, the input color and depth images are warped to align this local coordinate frame with the coordinate frame of the camera, using the surface interpolation - back-projection algorithm described above. For one of the pose compensated depth images I_r of each person a simplified version of the automatic pose estimation algorithm above is applied (ϕ need not be estimated) thus estimating a reference coordinate frame. This last step is important since the slant of the nose differs from person to person.

The proposed pose compensation algorithm although apparently complicated is both very accurate as will be demonstrated in section 6 and also computationally efficient. In fact the total running time is less than 1 sec on a Pentium III 1 Ghz computer. Note that three quarters of this execution time is consumed by the image warping procedure that may be easily implemented in hardware using an off-the-

self 3D graphics card.

5 Illumination Compensation

The term illumination compensation is used here to denote the normalization of the input brightness image so that the effect of varying illumination conditions is diminished. In this section an algorithm is described that compensates illumination by generating from the input image a novel image relit from a frontal direction. Our approach is inspired by recent work on image-based scene relighting used for rendering realistic images, e.g. placing a virtual object in a real image [37]. The idea behind image relighting is very simple. The image irradiance (brightness) is a function of scene geometry, surface reflectance (bidirectional reflectance distribution function, BRDF), viewer position and illumination distribution. Given one or more images of the scene from different viewpoints, scene geometry and surface reflectance one may invert this function to recover the illumination distribution. Then the recovered illumination can be used either to render a virtual object or to relight the scene under novel illumination. In practice there are several difficulties in this process. Inverting the rendering equation is an ill posed problem, and the acquisition of BRDF (a 4 dimensional function) is complicated, especially for natural objects like the human face. Since the goal here is classification efficiency rather than realism, a few assumptions are made that greatly simplify the problem. However, as explained in the sequel, the proposed technique is not bound by these assumptions and may be very easily extended to incorporate more complex models.

We assume that the scene is illuminated by a single light source with direction $\mathbf{L}$. The characteristics of the light source (e.g. distance from subject) may be cho-

sen depending on the application [1]. For each subject in our database we use the reference pose compensated depth image I_r to render N virtual views of the face illuminated from different directions. The set of light source directions is uniformly sampled from a section of the positive hemisphere. This set of images is used as bootstrap set for learning the nonlinear function that relates the image brightness distribution with the direction of the illuminant. To decrease the dimensionality of the problem, from each rendered image a feature vector is extracted containing locally weighted averages of image brightness over M preselected image locations ($M = 30$ in our experiments). The sample locations are chosen so as to include face areas with similar albedo (i.e. the skin). Feature vectors $\mathbf{x}_i,\ i = 1, \ldots, N$ extracted from all the images, normalized to have zero mean and unit variance, are then used as samples of the M-dimensional illuminant direction function $\mathbf{L} = \mathbf{G}(\mathbf{x})$. An approximation of this function $\tilde{\mathbf{G}}$ using the samples is a regression problem that is efficiently solved using Support Vector Machines (SVM) [38]. SVM regression is advantegeous over other machine learning techniques such as neural networks because: a) it always provides a global minimum of the objective function, b) it is insensitive to outliers in the training data, c) it admits regularization by direct control of the accuracy - model complexity trade off.

Assume now that we want to compute the similarity between a pose compensated probe image and gallery images of a person j in the gallery. A feature vector $\mathbf{x}$ is computed from the probe image as described previously. Then an estimate of the light source direction is given by $\tilde{\mathbf{G}}_j(\mathbf{x})$ i.e. the SVM regression function computed for the person j during the training phase.

[1] With our particular acquisition setup the main illumination of the face is contributed from the slide projector. Thus, distant light sources have a very small effect and the image irradiance is significantly affected only by spot lights close to the face. Therefore a spot light at short distance pointing to the face is assumed.

[Fig. 3 about here.]

Given the estimate of the light source direction $\mathbf{L}$ relighting the input image with frontal illumination $\mathbf{L}_0$ is straightforward. Let I_C, I_D be respectively the input pose compensated color and depth images and $\tilde{I}_C$ the illumination compensated image. Then the image irradiance for each pixel $\mathbf{u}$ is approximated by,

$$I_C(\mathbf{u}) = A(\mathbf{u})R(I_D, \mathbf{L}, \mathbf{u}), \quad \tilde{I}_C(\mathbf{u}) = A(\mathbf{u})R(I_D, \mathbf{L}_0, \mathbf{u}) \tag{2}$$

where A is the unknown face albedo or texture function (geometry independent component) and R is a rendering of the surface with constant albedo. When $R(\mathbf{u}) = \mathbf{n}(\mathbf{u})^T \cdot \mathbf{L}$, with $\mathbf{n}(\mathbf{u})$ denoting the surface normal at $\mathbf{u}$, we get the well-known Lambertian reflection model. Equation 2 is written

$$\tilde{I}_C(\mathbf{u}) = I_C(\mathbf{u}) \frac{R(I_D, \mathbf{L}, \mathbf{u})}{R(I_D, \mathbf{L}_0, \mathbf{u})}$$

i.e. the illumination compensated image is given by the multiplication of the input image with a ratio image. Figure 3 illustrates the relighting of a side illuminated image.

The same relighting procedure is applied also on the training images. Then it is expected that illumination compensated probe and gallery images of the same person will ideally differ up to a scale factor since the intensity of the light source may not be recovered. This scale factor is cancelled by taking the logarithm of the images, which makes the factor additive instead of multiplicative, and subsequently subtracting the mean value.

Although the description of the above relighting algorithm refers to a single channel image, color images may be handled equally well by applying the same procedure

(illuminant estimation and relighting) separately for each color channel.

An important advantage of the previously described algorithm is the flexibility in coping with complex illumination conditions by adaptation of the rendering process. For example, accounting for attached shadows may be simply achieved by activating shadowing in the rendering engine. Similarly, by using more complex rendering algorithms, one may exploit prior knowledge such as previously measured BRDFs of faces and pre-calibrated global illumination distributions. On the other hand good results may also be obtained with relatively simple renderings. In our experiments the average face reflectance is modelled as a weighted combination of a diffuse term and a specular term (Phong model). The parameters controlling the reflectance were estimated rather crudely by visual comparison between rendered images and original color images. In a practical application these parameters may be easily obtained during the enrollment of a subject e.g. using a pre-calibrated recording environment.

6 Experimental Results

We have conducted several experiments to evaluate the effectiveness of the proposed pose and illumination compensation algorithms in face recognition. An Embedded Hidden Markov Model (EHMM) classifier [39] was applied in order to estimate the similarity between the rectified probe images and a set of training images. Apart from the EHMM classifier tests with the Bayesian [40] and Elastic Graph Matching [41] algorithms were also performed, but results will not be presented here since the differences were not significant. [2]

[2] We have used the "Open Source Computer Vision Library" implementation for EHMM (http://www.intel.com/research/mrl/research/opencv/), and for Bayesian and Elastic Graph Matching the "CSU Face Identification Evaluation System"

For reasons of efficiency, the red color component was used for classification of color images. This yielded similar results as with gray-scale images and also with using all three spectral components as independent experts. Another EHMM classifier is used for depth images.

Each EHMM classifier is trained using a small set of training images. This set contains 2-5 images per person in the database depicting different facial expressions, thus achieving relative robustness against expression variations. The training images are rectified as described in sections 4 and 5. Testing was performed using images from a recording session different from the one used for training. For tests on the first database we used only the first 20 persons to make the results comparable with the second database. For the same reason, the images used for training and the images used for testing had a recording time lapse of 10 days.

Given a probe image and a claimed identity the output of the classifier is a similarity score between the probe image and images of the person with the given identity. In the case of face recognition this test is performed for all subjects in the database and the one with greater similarity is selected. The similarity scores S_C, S_D obtained by the color and depth classifiers respectively are subsequently normalized in the range $[0, 1]$. Fusion of color and depth cues is performed by computing a combined similarity score given by

$$log_{10}(S_{C+D}) = w_c \cdot \log_{10}(S_C) + w_d \cdot \log_{10}(S_D) \tag{3}$$

where w_c and w_d are relative weights corresponding to color and depth modalities respectively, chosen experimentally ($w_c = w_d = 0.5$ was used in our experiments). The fusion technique described above was shown elsewhere to give the best results

(http://www.cs.colostate.edu/evalfacerec/)

[12, 42].

We used the above experimental methodology to evaluate the effectiveness of the proposed algorithms and perform comparison with other techniques as detailed in the sequel.

6.1 Pose Compensation

Face detection, localization and pose estimation-compensation algorithms described in section 4 are applied to the set of test images. Then, pose compensated images are used for face classification as mentioned above. We compared the resulting recognition rates with those obtained using: a) manual affine pose normalization, b) manual 3D pose normalization, c) no normalization, d) database enrichment with artificial views. Manual normalization is performed by manually selecting three points on the face (e.g. on the eyes and mouth) that are sufficient for affine warping of the input image to a rectified pose. We have included this test since the majority of face recognition results in the literature assume the images have been previously normalized using this scheme. In the case that a 3D model of the face is available, 3D warping may be performed which does not suffer from distortions present in affine warping. The three manually specified points and associated depth values define a 3D coordinate system which is used to calculate the 3D pose of the face. Results obtained using this scheme may be considered close to the best results that one may obtain using pose compensation. In the case denoted as "no normalization" above we have used the face detection algorithm in [39] to detect, approximately localize and crop a sub-image containing the face. Results obtained using this approach correspond to a worst case scenario. Finally, we also list results from [12] where the original training database was enriched with novel artificially generated views

depicting pose variations (more than 100 views per image).

[Table 1 about here.]

[Table 2 about here.]

Table 1 shows face recognition results for different variations that may be affected by pose compensation and table 2 shows comparative face recognition results. From these results it is obvious that significant gains are obtained by using the proposed pose compensation algorithm. Its performance is better than manual affine warping and very close to the optimal manual 3D warping approach.

Worse results are obtained on images depicting large rotation angles. The main reason for this is not the accuracy of 3D pose estimation but occlusion and foreshortening of one part of the face that leads to distortions of the warped image. We are currently investigating a model-based surface interpolation scheme that is expected to solve this problem. Finally when the probe images contained faces wearing glasses and the corresponding gallery images the subjects are not wearing glasses (or vice versa), color-based recognition was bad. This is not the case however with depth-based recognition where the glasses surface may not be captured by the 3D sensor. Color-based recognition of “glass” images is improved by including such images in the training set as we did with the first database.

6.2 Illumination Compensation

Illumination compensation is applied on the pose compensated images automatically obtained in the previous test. Note, that since the structured light approach is used for the acquisition of range images, all images are affected by varying illumi-

nation since even small rotations of the head affect the image appearance.

The proposed illumination compensation scheme was compared with a) the robust gradient-based similarity measure of [21], b) the quotient image method [25], c) the symmetric shape from shading approach [24] and d) no illumination compensation. We chose the above techniques since unlike other compensation techniques (e.g. [27]) they do not require a training set containing images of the same person under varying illumination conditions. In [21] the weighted sum of differences of image gradient angles between the probe and gallery images is used as a similarity measure that is robust to illumination variations. The weighting function reflects the probability distribution of gradient angles on the image and is empirically estimated using a set of bootstrap images. The quotient image method [25] creates an illumination invariant representation of the original image (the quotient image) by dividing it with an illumination variable term estimated from the input image and a set of bootstrap images. The technique in [24] exploits bilateral symmetry of the face to estimate the light source direction and uses this information to create a symmetric "prototype" image using the input image and surface gradient information. Since no 3D information is available, the surface gradients are estimated from a global 3D face model (the average of all depth images is used in our experiments). Finally, for the results obtained with no illumination compensation original images are log-transformed and normalized to zero mean and unit variance. Without this step the results are much worse.

[Table 3 about here.]

[Table 4 about here.]

Table 3 demonstrates face recognition performance after illumination compensation. In table 4 a comparison of results obtained with different methods is presented.

By examining these results it is evident that the proposed illumination compensation algorithm outperforms state-of-the-art techniques. The results corresponding to the gradient difference [21] and quotient image approach were shown to have worse to, at best, equal performance with the case where no compensation is applied. For the first algorithm these results were more or less to be expected since it is a kind of robust image correlation measure. Correlation techniques were shown to be very sensitive to small misalignment between the probe and gallery images. The misalignment between images in the bootstrap set and the test set is the main reason for the bad performance of the quotient image method. For example our test set contains faces with glasses, mustaches and facial expressions which is not the case with images on the bootstrap set. Comparably good performance is achieved by the symmetric shape-from-shading algorithm [24]. The symmetry assumption widely used in this approach is the main drawback also leading to worse classification results compared with our approach. In [24] perfectly rectified images showing neutral expression are assumed which makes the symmetry assumption valid. However, in less than ideal situations this assumption leads to distortions apparent on reconstructed “prototype” images.

For the case of images with strong side illumination the improvement provided by our algorithm is smaller. Although the estimation of the light source direction was found to be globally accurate (as a linear combination of the projector and the external light source directions), locally the single light source assumption is not valid and therefore the rendering model used in our experiments was not sufficient to correctly recover areas on the sides of the nose and face. Range image artifacts also apparent over these areas further contribute to this problem. To cope with these issues in the future we plan a) using a pre-computed 3D model of each person’s face for the renderings in eq. (2) instead of the input range images and b) using

pre-computed BRDFs of faces and projector light source calibration data estimated during the training phase.

6.3 Effect of Depth Quality

The pose estimation algorithm was shown to be fairly robust, failing only for a few images where almost half of the face surface was missing. The accuracy of face rectification degraded with increased noise and missing pixels on the depth images. However, we have noticed that after filtering the input depth images with a Gaussian mask the results improved, thus demonstrating a graceful degradation of performance with the amount of sensor noise. The same observation holds for illumination compensation since depth noise and artifacts are reflected to the rendered images used for image relighting. This is apparent on results obtained with the 1st face database where the illumination compensation leads to practically no improvement. Again, simply filtering the depth images demonstrated a marked improvement of the results.

The dependence of recognition performance on the quality of 3D data should be judged positively; As 3D sensing technologies come to maturity better results should be expected.

7 Conclusions

This paper contributes to the efforts for robust recognition of faces. Robustness is achieved not only by introducing an additional cue, a 3D image of the face, but mainly by using this information to cope with pose and illumination variations that inhibit the performance of 2D face recognition. We proposed a “black box” ap-

proach capable of producing normalized images, which unlike state-of-the-art techniques is computationally efficient and does not need an extended training set. As demonstrated by extensive experimental evaluation the proposed algorithms lead also to superior face recognition rates.Having examined techniques to cope with appearance variations our aim in the future is to investigate in more depth the 3D face classification problem and optimal fusion of color and depth information.

8 Acknowledgement

This work has been supported by the EU FP6 integrated project BIOSEC “Biometrics and Security ” (IST-2004-001766).

References

[1] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, J. Bone, Face recognition vendor test 2002, evaluation report, Tech. rep. (March 2003).

[2] W. Zhao, R. Chellappa, A. Rosenfeld, , P. Phillips, Face recognition: A literature survey, Tech. rep., University of Maryland, ftp://ftp.cfar.umd.edu/TRs/CVL-Reports-2000/TR4167-zhao.ps.gz (October 2000).

[3] F. Forster, P. Rummel, M. Lang, B. Radig, The hiscore camera: a real time three dimensional and color camera, in: Proc. Int. Conf. Image Processing, Vol. 2, 2001, pp. 598–601.

[4] E. Trucco, A. Verri, Introductory Techniques for 3-D Computer Vision, Prentice-Hall, NJ 07458, 1998.

[5] J. C. Lee, E. Milios, Matching range images of human faces, in: Proc. IEEE Int. Conf. on Image Processing, 1990, pp. 722–726.

[6] G. Gordon, Face recognition based on depth and curvature features, in: Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR ’92, 1992, pp. 808–810.

[7] H. T. Tanaka, M. Ikeda, H. Chiaki, Curvature-based face surface recognition using spherical correlation. principal directions for curved object recognition, in: Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition, 1998, pp. 372–377.

[8] C.-S. Chua, F. Han, Y.-K. Ho, 3D human face recognition using point signature, in: Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2000, pp. 233–238.

[9] Y. Wang, C.-S. Chua, Y.-K. Ho, Y. Ren, Integrated 2D and 3D images for face recognition, in: Proc. 11th Int. Conf. on Image Analysis and Processing, 2001, pp. 48–53.

[10] S. Tsutsumi, S. Kikuchi, M. Nakajima, Face identification using a 3D gray-scale image-a method for lessening restrictions on facial directions, in: Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition, 1998, pp. 306–311.

[11] K. Chang, K. Bowyer, P. Flynn, Face recognition using 2D and 3D facial data, in: Proc. Multimodal User Authentication Workshop, Santa Barbara, 2003, to appear.

[12] F. Tsalakanidou, S. Malassiotis, M. G. Strintzis, Face authentication using color and depth images, IEEE Trans. Image Processing (submitted).

[13] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. and Mach. Intell. 23 (6) (2001) 681–685.

[14] V. Blanz, S. Romdhami, T. Vetter, Face identification across different poses and illuminations with a 3D morphable model, in: Proc. of Int. Conference on Automatic Face and Gesture Recognition, 2002, pp. 202–207.

[15] D. Beymer, T. Poggio, Face recognition from one example view, in: Proc. IEEE Int. Conf. Computer Vision, 1995, pp. 500–507.

[16] T. S. Jebara, 3D pose estimation and normalization, Ph.D. thesis, McGill University (1996).

[17] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, in: Proc. Conf. Computer Vision and Pattern Recognition, 1994.

[18] D. B. Graham, N. M. Allinson, Face recognition from unfamiliar views: Subspace methods and pose dependency, in: Proc. of Int. Conference on Automatic Face and Gesture Recognition, 1998, pp. 348–353.

[19] P. J. Philips, H. Moon, P. J. Rauss, S. Rizvi, The feret evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. and Mach. Intell. 22 (10) (2000) 1090–1100.

[20] D. Blackburn, J. Bone, P. Phillips, Face recognition vendor test 2000, evaluation report, Tech. rep. (February 2001).

[21] H. F. Chen, P. N. Belhumeur, D. W. Jacobs, In search of illumination invariants, in: Proc. Conf. Computer Vision and Pattern Recognition, 2000, pp. I: 254–261.

[22] D. W. Jacobs, P. N. Belhumeur, R. Basri, Comparing images under variable illumination, in: Proc. Conf. Computer Vision and Pattern Recognition, 1998, pp. 610–614.

[23] M. Savvides, B. V. Kumar, Efficient design of advanced correlation filters for robust distortion-tolerant face recognition, in: Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS'03), 2003, pp. I: 254–261.

[24] W. Zhao, R. Chellappa, Illumination-insensitive face recognition using symmetric shape-from-shading, in: Proc. Conf. Computer Vision and Pattern Recognition, 2000, pp. IV: 297–301.

[25] A. Shashua, T. R. Raviv, The quotient image: Class based re-rendering and recognition with varying illuminations, IEEE Trans. Pattern Anal. and Mach. Intell. 23 (2) (2001) 129–139.

[26] P. N. Belhumeur, D. J. Kriegman, What is the set of images of an object under all possible illumination conditions?, International Journal of Computer Vision 28 (3) (1998) 1–16.

[27] A. S. Georghiades, P. N. Belhumeur, D. J. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. and Mach. Intell. 23 (6) (2001) 643–660.

[28] R. Ishiyama, S. Sakamoto, Geodesic illumination basis: compensating for illumination variations in any pose for face recognition, in: Proc. of Int. Conference on Pattern Recognition, 2002, pp. IV: 297–301.

[29] R.Basri, D.Jacobs, Lambertian reflectance and linear subspaces, in: Proc. IEEE Int. Conf. Computer Vision, 2001, pp. II:384–390.

[30] L. Zhang, D. Samaras, Face recognition under variable lighting using harmonic image exemplars, in: Proc. Conf. Computer Vision and Pattern Recognition, 2003, pp. I:19–25.

[31] S. R. Marschner, Inverse rendering for computer graphics, Ph.D. thesis, Cornell University (1998).

[32] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, M. Sagar, Acquiring the reflectance field of a human face, in: Proc. of SIGGRAPH 2000, Computer Graphics Proceedings, Annual Conference Series, 2000, pp. 145–156.

[33] Z. Wen, Z. Liu, T. S. Huang, Face relighting with radiance environment maps, in: Proc. Conf. Computer Vision and Pattern Recognition, 2003, pp. II:158–162.

[34] S. Malassiotis, M. G. Strintzis, Real-time head tracking and 3D pose estimation from range data, in: Proc. Int. Conf. Image Processing, Barcelona, Spain, 2003.

[35] S. Malassiotis, M. G. Strintzis, Robust real-time 3D head pose estimation from range data, Pattern Recognition (submitted).

[36] P. J. Besl, N. D. McKay, A method for registration of 3D shapes, IEEE Trans. Pattern Anal. and Mach. Intell. 14 (2) (1992) 239–256.

[37] I. Sato, Y. Sato, K. Ikeuchi, Acquiring a radiance distribution to superimpose virtual objects onto a real scene, IEEE Transactions on Visualization and Computer Graphics 5 (1) (1999) 1–12.

[38] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE Neural Networks 12 (2) (2001) 181–201.

[39] F. Samaria, Face recognition using hidden markov models, Ph.D. thesis, University of Cambridge (1984).

[40] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Trans. Pattern Anal. and Mach. Intell. 19 (7) (1997) 696–710.

[41] L. Wiskott, J.-M. Fellous, N. Kruger, C. Malsburg, Face recognition by elastic bunch graph matching, IEEE Trans. Pattern Anal. and Mach. Intell. 19 (7) (1997) 775–779.

[42] J. Kittler, M. Ballette, J. Czyz, F. Roli, L. Vandendorpe, Enhancing the performance of personal identity authentication systems by fusion of face verification experts, in: Proc. IEEE Int. Conf. on Multimedia and Expo, Vol. 2, 2002, pp. 581–584.

List of Figures

(a) (b)

Fig. 1. (a) Image captured by the color camera with color pattern projected. (b) Computed range image. Brighter pixels correspond to points closer to the camera and white pixels correspond to undetermined depth values)

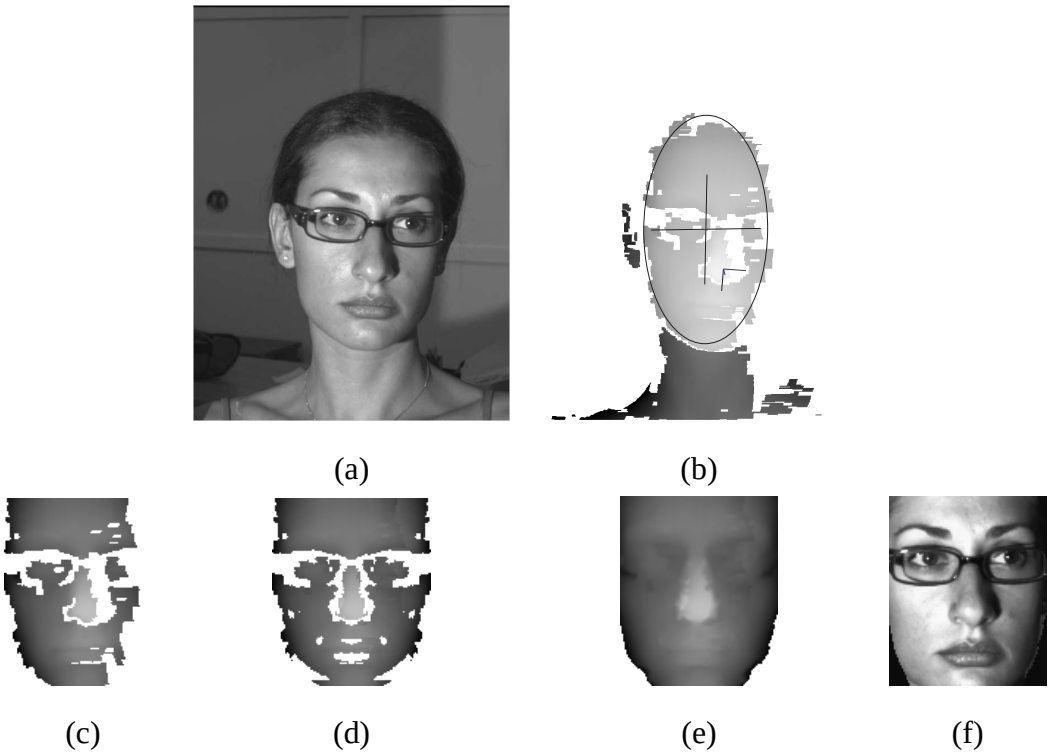
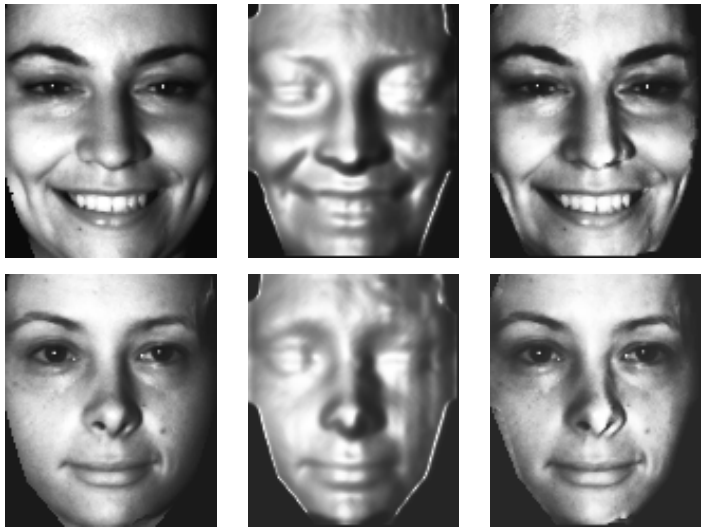


Fig. 2. Pose compensation example. (a) Original color image, (b) original depth image showing detected head blob and estimated local coordinate system fixed on the nose, (c) rectified depth image, (d) symmetry-based interpolation, (e) final linear interpolated depth image, (f) rectified color image.



(a) (b) (c)

Fig. 3. Illumination compensation example. (a) Original image, (b) $R(I_D, \mathbf{L}, \mathbf{u})$, (c) frontally illuminated image

List of Tables

	All	Frontal	Pose	Glasses	Expressions
C	85.3	93.3	60.2	95.0	100.0
D	83.8	81.7	80.1	80.2	93.4
C+D	95.6	93.3	90.4	95.0	100.0
1st Database					
C	90.7	100.0	80.3	69.7	96.7
D	96.4	95.5	90.4	97.7	95.3
C+D	99.0	100.0	96.4	97.7	99.4
2nd Database					

Table 1
Recognition rates after pose compensation for different image variations (color images: C, depth images: D, color + depth: C+D).

	1st Database			2nd Database		
	C	D	C+D	C	D	C+D
No compensation	69.0	27.1	69.7	67.0	58.3	76.5
Database enrichment	76.1	37.2	79.6	70.1	62.0	77.3
Affine warping	81.9	81.5	94.6	88.8	96.8	98.9
3D pose compensation	85.3	83.8	95.6	90.7	96.4	99.0
3D warping	85.5	84.5	96.1	90.7	97.8	99.2

Table 2
Recognition rates for different pose compensation schemes (Color images: C, depth images: D, color + depth: C+D).

	All	Frontal	Illumination	Pose
C	85.4	94.1	78.3	52.4
C+D	95.6	94.3	91.7	89.2
1st Database				
C	93.8	100.0	92.9	90.4
C+D	99.0	100.0	99.5	97.2
2nd Database				

Table 3
Recognition rates after pose and illumination compensation for different image variations (color : C, color + depth: C+D).

	1st Database		2nd Database	
	C	C+D	C	C+D
No compensation	85.3	95.6	90.7	99.0
Relighting	85.4	95.6	93.8	99.1
Robust similarity [21]	67.8	85.5	87.8	96.3
Quotient image [25]	79.3	90.4	89.3	97.1
Symmetric SFS [24]	81.3	92.2	92.8	98.7

Table 4
Recognition rates for different illumination compensation algorithms (color : C, color + depth: C+D).