COMBINING 2D AND 3D FACE IMAGES FOR RELIABLE IDENTITY VERIFICATION

Sotiris Malassiotis and Michael G. Strintzis

Informatics and Telematics Institute, Thessaloniki, Greece
e-mail: malasiot@iti.gr

ABSTRACT

The paper describes a complete face authentication system using a combination of color and depth images. Depth information acquired by a novel 3D and color sensor is used for robust face detection, localization and 3D pose estimation. To cope with illumination and pose variations 3D information is used for the normalization of the input images. Illumination compensation exploits depth data to recover the illumination of the scene and relight the image under frontal lighting. When normalized images, depicting upright orientation and frontal lighting, are used for authentication significantly low error rates are achieved, as demonstrated on a face database with more than 3000 images.

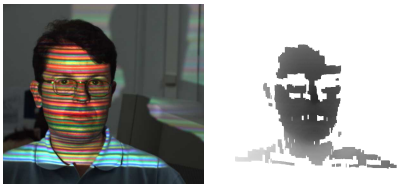
1. INTRODUCTION

Recent public face recognition tests demonstrated that the accuracy of state-of-the-art algorithms degrades significantly for images exhibiting pose and illumination variations. Current research efforts strive to achieve insensitivity to such variations.

The paper describes and evaluates a complete face authentication system using a combination of 2D color and 3D range images captured in real-time. The key innovation of the system lies on 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 compared with semi-automatic rectification.

Although the 3D structure of the human face conveys important discriminatory information only a few techniques have been proposed employing range images. This is mainly due to the high cost of available 3D digitizers and the fact that they 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 novel low cost 3D of low-cost sensors that are capable of real-time 3D acquisition [1]. A common approach adopted towards 3D face recognition is based on the extraction of 3D facial features by means of differential geometry techniques [2–4]. A few techniques [5, 6] also employ grayscale images but mainly for augmenting the detection of features such as the eyes that are harder to detect on the range image. Although feature-based techniques are robust to pose variations they rely on accurate 3D maps of faces, usually extracted by expensive off-line 3D scanners. Thus their applicability to real-world applications with highly noisy data is questionable. The recognition rates claimed by the above techniques were estimated using databases of limited size and without significant variations of the faces. Only recently [7] conducted an experiment with a database of significant size (275 persons) containing both grayscale and range images, and produced comparative results of face identification using eigenfaces for 2D, 3D and their combination and for varying image quality. This test however considered only frontal images captured under constant illumination conditions. For this work we have recorded a face database containing several appearance variations. These variations are compensated before reaching the classifier, thus leading to low error rates.

This work is funded by research project BioSec IST-2002-001766 (Biometrics and Security, http://www.biosec.org), under Information Society Technologies (IST) priority of the 6th Framework Programme of the European Community.



(a) (b)

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

2. ACQUISITION OF 3D DATA

The proposed system is based on real-time quasi-synchronous color and 3D image acquisition based on the color structrured-light approach [1] (fig. 1). The sensor is based on low cost devices, an off-the-shelf CCTV-color camera and a standard slide projector. The average depth accuracy of the system optimized for an access control application is about $0.5mm$. The spatial resolution of the range images is approximately equal to the color camera resolution.

3. POSE COMPENSATION

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

The proposed technique uses the range image only for face detection and pose estimation and therefore is robust especially under varying pose and illumination conditions, as demonstrated by the experimental results.

The detection of the face in the image is the first step of the algorithm. 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 incorporation of a-priori constraints on the relative dimensions of the body parts, described in detail in [8].

The estimation of 3D head pose, performed next is based on the detection of the nose [8]. 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. This 3D line defines two of the three degrees of freedom of the face orientation. The third degree of freedom, that is the rotation angle around the nose axis, is then estimated by finding the 3D plane that cuts the face into two bilateral symmetric parts. The error of the above pose estimation algorithm tested on more than 2000 images is 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, which is defined during the training faces using the gallery images, bringing the face in up-right orientation. The transformation between the local and reference coordinate frames is further refined to pixel accuracy by applying the ICP [9] surface registration algorithm between the warped and a reference (gallery) depth image corresponding to claimed person ID.

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. Remaining missing pixel values are linearly interpolated from neighboring points. The interpolated depth map is subsequently used to rectify the associated color image also using 3D warping (fig. 2).

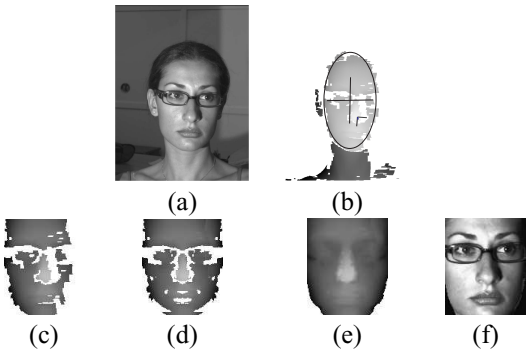


Figure 2: Pose compensation example. (a) Original color image, (b) original depth image showing detected head blob and estimated local coordinate system fixed on the nose, (c) rectified depth image, (d) symmetry-based interpolation, (e) final linear interpolated depth image, (f) rectified color image.

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 algorithm described above. For one of the pose compensated depth images of each person a simplified version of the automatic pose estimation algorithm above is applied thus estimating a reference coordinate frame. This last step is important since the slant of the nose differs from person to person.

4. ILLUMINATION COMPENSATION

In this section an algorithm is described that compensates illumination by generating from the input image a novel image relight from a frontal direction. Our approach is inspired by recent work on image-based scene relighting used for rendering realistic images. Image relighting relies on inverting the rendering equation, i.e. the equation that relates the image brightness with the object material and geometry and the illumination of the scene. Given several images of the scene under different conditions this equation may be solved (although an ill-posed problem) to recover the illumination distribution and then use this to re-render the scene under novel illumination.

The first step is therefore to recover the scene illumination from a pair of color and depth images. Assuming that the scene is illuminated by a single light source a technique is adopted that learns the non-linear relationship between the image brightness and light source direction $\mathbf{L}$ using a set of artificially generated bootstrap images.

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. 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 may be efficiently solved using Support Vector Machines (SVM) [10]. 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.

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,

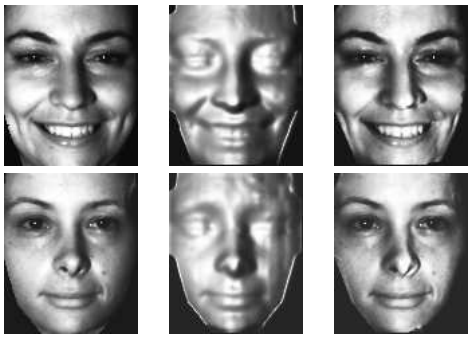
$$I_C(\mathbf{u}) = A(\mathbf{u})R(I_D,\mathbf{L},\mathbf{u}),\ \tilde{I}_C(\mathbf{u}) = A(\mathbf{u})R(I_D,\mathbf{L}_0,\mathbf{u}) \quad (1)$$

where A is the unknown face albedo or texture function (geometry independent component) and R is a rendering of the

surface with constant albedo. Equation 1 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 multiplication of the input image with a ratio image.



(a) (b) (c)

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

Figure 3 illustrates the relighting of a side illuminated image.

The same relighting procedure is also applied on training images. Then it is expected that illumination compensated probe and gallery images of the same person will only 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 (that makes the factor additive instead of multiplicative) and subsequently subtracting the mean value.

Although the description of the above relighting technique considers a single channel image, color images may be handled equally well by applying the same procedure (illuminant estimation and relighting) separately for each channel.

An important advantage of the previously described algorithm is the flexibility in coping with complex illumination conditions by adaptation of the rendering function R above. For example, accounting for attached shadows may be simply achieved by activating shadowing in the rendering engine. On the other hand, from our experience with different rendering models, good results may be also obtained with relatively simple renderings.

5. FACE CLASSIFICATION

Normalized color and depth images are subsequently provided as input to a color classifier and a depth classifier respectively. For practical reason only the red component of the color images was used. Pixel values are normalized to have zero mean and unit variance before classification. We have experimented with several state of the art techniques including Embedded Hidden Markov Models (EHMM) [11], Probabilistic Matching (PM) [12] and Elastic Graph Matching (EGM) [13]. The two latter techniques were shown to give the lowest error rates. We have finally selected PM algorithm since its computational complexity was one order of magnitude smaller than that of the EGM algorithm for similar error rates. For depth images the same algorithms were tested where in this case pixel values correspond to distance from the sensor rather than brightness. In addition we have conducted experiments with a simplified version of the Point Signature algorithm that has been proposed for surface matching [14]. The results using this feature-based approach were not as good as appearance-based techniques possibly due to the sensitivity of this algorithm to noise and artifacts (e.g. holes) in the depth images captured by the 3D sensor. Finally PM was adopted for both color and depth image classification. The scores (maximum-likelihood probabilities) returned by each classifier are subsequently normalized. This consists of first estimating for each enrolled user i the mean m_i and variance $_i$ of the genuine transaction score distribution (by means of robust estimation techniques) using a set of scores extracted using a bootstrap set of images. Then we apply the *tanh* normalization technique to map the scores in the interval $[0, \ldots, 1]$.

$$s' = \frac{1}{2}\{tanh(0.05(\frac{(s - \ _i)}{\ _i})) + 1\}$$

where s and s' are the initial and normalized scores respectively (see [15] for more details). Once the scores are normalized fusion of color and depth modalities is achieved by simply adding the corresponding scores.

6. EXPERIMENTAL RESULTS

The focus of the experimental evaluation was to investigate the efficiency of the complete face authentication chain on conditions that are similar to those encountered in real-world applications. Therefore a database containing several facial appearance variations was recorded. 73 volunteers participated in two recording sessions. The second recording session was 10 days after the first session. For each subject several images depicting different appearance variations were acquired: facial expressions (smiling, laughing), illumination (side spot light), pose variations ($\pm$20 degrees), images with glasses, and frontal images (up to 50 images per person). Totally, more than 3000 image pairs were recorded.

The PM algorithm was trained using 4 images per person from the first recording session, including one frontal image, two images depicting facial expressions and one image with glasses. Testing was performed using the remaining images.

	A	F	E	P	I	G
C	7.6	2.1	6.3	9.3	5.2	3.2
C (PC+IC)	4.7	1.2	5.1	6.9	3.1	3.8
D	5.7	1.2	4.6	7.1	2.2	4.8
D (PC)	4.3	1.2	4.7	5.1	2.3	4.8
C+D	5.2	1.0	3.9	8.6	2.7	3.2
C+D (PC+IC)	2.8	0.6	3.5	3.6	1.4	3.4

Table 1: Equal error rates (%) for different image variations (A: All variations, F: Frontal, E: Expressions, P: Pose, I: Illumination, G: Glasses) and modalities (C: Color, D: Depth, C+D: color + depth: C+D). Even rows show results with the proposed pose (PC) and illumination compensation (IC) techniques, while odd rows are results with manual image rectification

Table 1 demonstrates the equal error rates achieved with the proposed compensation scheme. This is compared with

the case of manual pose normalization i.e. three points over the eyes and mouth were selected by a human operator and used to rectify the images. Rectification in this case is performed by 2D affine warping of the images. As shown in table 1 the proposed scheme results in authentication errors which are much better to the accuracy achieved by manual image normalization especially for images depicting pose variations). Also there is a significant improvement after using the illumination compensation algorithm. In figure 4 the receiver operator characteristic curve is also shown. By combining color and depth information for a false acceptance rate of 0.5% a correct recognition rate of 95% is obtained.

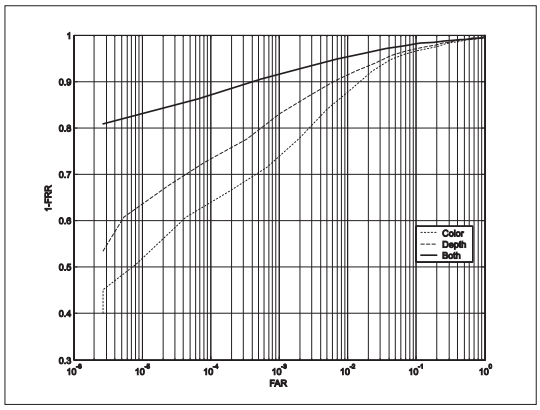


Figure 4: Receiver operating characteristic curve showing false acceptance rate (FAR) versus correct recognition rate (1-FRR) for the two modalities and their combination.

The running time of the algorithm from the acquisition of color and 3D images to verification of the identity is about 3 sec on a Pentium 4, 3Ghz processor. The most time consuming part of the algorithm is the 3D image warping step. Since this is a common 3D graphics operation it is possible to achieve significant reduction of the running time (less than 1 sec) by exploiting off-the-shelf graphics hardware.

7. CONCLUSIONS

In summary we have proposed a new approach for 2D + 3D face authentication based on automatic image normalization algorithms exploiting the availability of 3D information. Significant improvements in face classification accuracy were obtained using this scheme.

The system was shown to be more sensitive to pose variations and facial expressions. Although the proposed pose compensation procedure significantly reduces the error rates for images depicting pose variations it can lead to some distortion on the side of the face where occluded parts of the face appear after rotation. We plan to deal with this issue by investigation of novel classification schemes able to disregard missing pixels. Coping with facial expressions is also a challenging problem. We are currently working on a database enrichment scheme where synthetic images depicting facial expressions are automatically generated from frontal images using a 3D animated face model.

REFERENCES

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

[2] G.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.

[3] H. T. Tanaka, M. Ikeda, and 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.

[4] C.-S. Chua, F. Han, and 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.

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

[6] S. Tsutsumi, S. Kikuchi, and 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.

[7] K. Chang, K. Bowyer, and P. Flynn, "Face recognition using 2d and 3d facial data," in *Proc. Multimodal User Authentication Workshop*, Santa Barbara, December 2003, to appear.

[8] S. Malassiotis and M. G. Strintzis, "Real-time head tracking and 3d pose estimation from range data," in *Proc. Int. Conf. Image Processing*, Barcelona, Spain, September 2003.

[9] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 14, no. 2, pp. 239–256, February 1992.

[10] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Neural Networks*, vol. 12, no. 2, pp. 181–201, May 2001.

[11] F. Samaria, *Face Recognition Using Hidden Markov Models*, Ph.D. thesis, University of Cambridge, 1994.

[12] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenfaces: Probabilistic matching for face recognition," in *Proc. of Int'l Conf. on Automatic Face and Gesture Recognition (FG'98)*, Nara, Japan, April 1998, pp. 30–35.

[13] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 19, no. 3, pp. 775–779, July 1997.

[14] C. S. Chua and R. Jarvis, "Point signatures: A new representation for 3d object recognition," *International Journal of Computer Vision*, vol. 25, no. 1, pp. 63–85, October 1997.

[15] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*.