Real time acquisition of depth and color images using structured light and its application to 3D face recognition

Filareti Tsakalakidou\textsuperscript{a,b,\ast}, Frank Forster\textsuperscript{c} Sotiris Malassiotis\textsuperscript{b}, Michael G. Strintzis\textsuperscript{a,b}

\textsuperscript{a}Information Processing Laboratory, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

\textsuperscript{b}Informatics and Telematics Institute, Centre for Research and Technology Hellas, 1st km Thermi-Panorama Road, Thermi 57001, P.O. Box 361, Thessaloniki, Greece

\textsuperscript{c}Siemens AG, CT PS 9, Otto-Hahn-Ring 6, 81337 Munich, Germany

Abstract

In this paper, a novel real time 3D and color sensor for the mid-distance range (0.1-3m) based on color-encoded structured light is presented. The sensor is integrated using low-cost off-the-shelf components and allows the combination of 2D and 3D image processing algorithms, since it provides a 2D color image of the scene in addition to the range data. Its design is focused on enabling the system to operate reliably in real-world scenarios, i.e. in uncontrolled environments and with arbitrary scenes. To that end, novel approaches for encoding and recognizing the projected light are used, which make the system practically independent of intrinsic object colors and minimize the influence of the ambient light conditions. The system was designed to assist and complement a face authentication system integrating both 2D and 3D images. Depth information 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. The performance and robustness of the proposed system is tested on a face database recorded in conditions similar to those encountered in real-world applications.

*Key words:* depth image acquisition, structured light, face recognition, pose, illumination

1 Introduction

Acquiring the 3D surface data of arbitrary scenes has been a primary objective of computer vision and related fields for many decades. While there has been considerable success with respect to specific tasks, e.g. capturing the 3D geometry of static scenes, the problem of acquiring high-resolution, high accuracy depth maps of rapidly moving objects in real time and at reasonable cost is still unsolved. This paper presents a novel low-cost 3D sensor capable of real-time acquisition of range images of moving objects in arbitrary scenes up to a few meters away. So far, it has been successfully used for several commercial applications such as medical implant modelling and wheel alignment in the automotive industry. In this paper, we exploit the availability of 3D data for robust face authentication.

During the last 10 years, state-of-the-art face recognition systems using 2D intensity images have advanced to be fairly accurate under controlled conditions. However, public face recognition benchmarks have shown that their performance degrades dramatically when pose or illumination variations are present in face im-

* Corresponding author.

Email addresses: filareti@iti.gr (Filareti Tsalakanidou), frank.forster@siemens.com (Frank Forster), malasiot@iti.gr (Sotiris Malassiotis), strintzi@eng.auth.gr (Michael G. Strintzis).
ages [1]. To improve performance under these conditions, 3D face recognition is proposed. The use of 3D images along with 2D images for personal identification is based on the fact that the shape of faces can be highly discriminatory and is not affected by changes in lighting or by facial pigment [2].

In the following section, we review the related work in the field of range image acquisition and 3D face recognition, highlighting the novelties of the proposed work. A detailed description of the proposed range image acquisition method is given in Section 3. A full face detection, normalization and authentication system using 2D and 3D facial data is outlined in Section 4, while experimental results evaluating the performance of the developed techniques in real-world conditions are presented in Section 5. Finally, conclusions are drawn in Section 6.

2 Previous work

2.1 Real-Time Acquisition of Range Images

In this paper, a 3D sensor based on an improved and extended version of the well-known Coded Light Approach (CLA) for 3D-data acquisition is proposed. The basic principle lying behind this technique is the projection of an encoded light pattern on the scene and measurement of its deformation on the objects surfaces. A detailed review of pattern codification methods in structured light systems can be found in [3, 4]. Despite the fact that numerous coded light systems have been proposed in the literature, only a few are aimed at real time operation; even fewer additionally permit unlimited movement of the scene. These systems are based on spatial coding with a single static projection pattern, where the projected light rays or planes are encoded by spatial markings, called subpatterns, within this pattern.
They are sometimes called one-shot systems, as they are able to extract the depth data from a single gray-level or color snapshot of the scene.

The first category of one-shot systems uses a black-and-white projection pattern [5,6]. The main advantage of a black-and-white approach is that it works well, even with strongly colored scenes. Moreover, it is straightforward to make a monochrome projection pattern unobtrusive to humans via an infrared-only illumination or to block off most background light by employing a narrow-band illumination in combination with a matching camera-side band-pass. The common downside of the black-and-white one-shot systems is their coarse lateral resolution, typically in the area of $64 \times 64$ depth values [5] or less [6], which is due to the low transmission capacity of the typically binary (composed of only two distinct gray levels) projection patterns. The latter aspect also prevents the use of error-detecting encoding that would make the systems more robust, i.e. that would help to avoid outliers in the depth data.

A color pattern on the other hand, offers three separate color channels, i.e. potentially the triple transmission capacity over a monochrome projection pattern. Thus, coded light systems with a colored projection pattern tend to achieve a much higher lateral resolution of the depth data, even the ones employing error-detecting codes. However, the colors reflected from a scene strongly depend on its intrinsic reflectivity, implying that the colors seen by the camera can differ substantially from the originally projected ones. Nevertheless, most one-shot color-based coded light systems simply equate the former with the latter. As a result, they are limited to scenes with neutral colors [6–8].

A technique that theoretically avoids this problem is the rainbow approach [9, 10], which is based on the projection of a pattern of monochromatic colors and its en-
coding using the wavelength. The key assumption is that such a pattern yields a reflection which is modified in its intensity, but not in its spectral composition. Given the continuum of distinct wavelengths, the rainbow approach is in principle suited for acquiring depth maps of very high lateral resolution. However, in practice, mutual reflection and ambient light tend to invalidate this assumption. Moreover, off-the-shelf color cameras are not suited for distinguishing subtle changes in the wavelength, thus making the rainbow approach very sensitive to projection and imaging noise. The fact that the rainbow approach does not employ error-detecting encoding, makes it even more problematic.

The above paragraphs discussed depth image acquisition in general. Clearly it is significantly simpler to acquire the 3D surface data of human faces, given the fact that the approximate 3D shape of a face is known a-priori and that human faces represent a single, mostly continuous and convex surface with rather benign reflection properties. Although, there are some depth sensors that are exclusively targeted at the acquisition of 3D images of the face [11], they are all commercial products, and thus no technical details are available in the literature.

The proposed technique, a previous version of which has been presented in [12,13], is a new approach to color-coded light via spatial encoding with a single colored projection pattern. As all such approaches, it computes a depth map from a single snapshot of the scene illuminated by the pattern; accordingly, it copes with scenes that move rapidly, the only limit being the motion blurring in the color image, and is in principle capable of a 3D frame rate equal to that of the camera used. Moreover, a simple slide projector suffices to generate the necessary un-varying illumination.
2.2 3D Face Recognition

The earliest approach adopted towards 3D face recognition is based on the extraction of 3D facial features by means of differential geometry techniques. Surface curvature is used for the localization of facial features such as the eyes, eyebrows, nose, etc [14] or for the construction of the Extended Gaussian Images (EGI) of the face [15]. More sophisticated feature-based techniques, such as Point Signatures, which describe the local structure of face points and are used to find correspondences between 3D faces, have also been proposed [16].

The most significant disadvantage of feature-based approaches is that they rely on accurate 3D maps of faces, usually extracted by expensive off-line 3D scanners. Their applicability on noisy data acquired by real-time sensors can be problematic, especially if computation of curvature is involved. Also, the computational cost is significantly high.

To avoid these problems, an appearance based approach is employed in this work. Appearance based methods like PCA or Fisherfaces, simplify 3D data processing by using 2D depth images, where pixel values represent the distance of corresponding points from the sensor. The main problem with such techniques is the requirement for accurate alignment between probe and gallery images. In this paper, a pose and illumination compensation scheme is proposed to achieve such geometric and photometric alignment. We also propose a combination of 2D and 3D images for reliable multimodal authentication.

The combination of 2D and 3D facial images has also been proposed in [17–19] and considerable performance improvement has been achieved. In [18], a database of significant size was used to produce comparative results of face identification using
eigenfaces for 2D, 3D and their combination. This test however considered only frontal images captured under frontal illumination. In [19], a database containing several variations (pose, illumination, expressions, glasses, several recording sessions) was used. To cope with intrapersonal variations due to viewpoint or lighting conditions, the face database was automatically enriched with artificially generated examples depicting variations in pose and illumination.

In this paper, we describe and evaluate a complete face authentication system using a combination of 2D color and 3D range images captured in real-time using a novel low-cost 3D and color sensor. Variations of facial images due to changes in pose or lighting conditions are compensated prior to classification. Also, a novel real-time face detection and localization method relying exclusively on depth data and prior knowledge of face geometry and symmetry is presented. The efficiency and robustness of the proposed system is demonstrated on a data set of significant size.

3 Range Image Acquisition Method

In this section, a novel approach towards real-time acquisition of range data is proposed based on color-coded light and spatial encoding with a single colored projection pattern.

The main disadvantage of color patterns is that while decoding requires recognizing the projected colors in the color image, the reflected colors cannot be relied upon for that task. For example, perceived red might be due to projected red, but also due to projected white and a red surface. The idea used to overcome this problem is derived as follows: an intrinsic assumption with spatial encoding - namely that most projected subpatterns are reflected rather integrally - corresponds to the assumption
that depth and reflectivity of the scene vary smoothly almost everywhere, i.e. to a continuity and a reflectivity smoothness constraint. This implies that if spatial encoding is used anyway, these two constraints can be exploited without introducing new restrictions. Then, given reflectivity smoothness, the scene exhibits only occasionally reflectivity edges of its own. Given the continuity constraint, there are only a few edges due to object boundaries or sharp changes in the scene’s surface orientation. This also implies that projected edges appear as edge segments in the pattern image and that spatial adjacency relations of the imaged segments will in most cases remain as projected. This argumentation leads to the idea of employing local color edge segment patterns as subpatterns.

3.1 The Projection Pattern

Clearly the projection pattern is the key to build a practically usable depth acquisition technique based on the above idea. Accordingly, a systematic approach is taken with its design. First, a set of required or looked-for pattern properties is derived from a theoretical model of a coded light system interpreted as a data transmission system. This results in several conclusions, some of which differ from the state of the art. For example, existing coded light systems focus exclusively on symbol errors and they strive for a large minimal distance of the code. However, it is unavoidable that some parts of the transmission are irreversibly lost. Equally inevitable are ghost symbols, i.e. code symbols received even though they have never been sent (e.g. a color edge part of the scene). So, in coded light sensors, synchronization errors occur potentially frequently and have to be taken into account during (de)coding. It is important to note that the minimal distance of a code tells nothing about its robustness with respect to them: the two codewords $c_1 = 0101$
and \(c_2 = 1010\) have a maximal Hamming distance; assume \(c_1\) is part of a longer message, e.g. 1111 0101 0111 \ldots and that its leading 0 is lost due to synchronization problems. Then the sequence 1111 1010 111 \ldots is received, containing \(c_2\) in place of the sent \(c_1\). In short, a single synchronization error can result in an undetectable decoding error, even if the two codewords involved are maximally distant according to their Hamming distance.

Given the set of requirements, a new type of projection pattern is developed to meet them, which guarantees that, for example, a single missing or a single ghost symbol per codeword is detected. Moreover, at least \(n/2\) symbol errors are detected per codeword, where \(n\) is the codeword length. In its simplest form, which is assumed throughout the rest of this paper, a pattern of the proposed type is composed of parallel color stripes such that \(n\) adjacent color edges form a codeword. The colors used are the eight corners of the RGB cube. Adjacent stripes are required to differ in at least two color channels, implying there are exactly 20 distinct types of projected edges. The number of stripes is chosen so that the projected edges can be expected to be about 4 pixels apart in the color image. The projected color-coded pattern can be seen in Fig. 1. A detailed description of the projection pattern can be found in [4].

### 3.2 Data Processing

In this section, we discuss the generation of depth images of arbitrary scenes based on color images depicting the scene illuminated with a stripe pattern. The task of demodulation, i.e. detecting the projected color edges and establishing their spatial relationship, appears to be straightforward: edge detection in images is one of the most classical and best understood problems of image processing. However, all
state-of-the-art edge detection approaches present certain disadvantages for the task at hand. The most important is that they detect all edges present, even though only certain types of edges, namely the projected ones, are of interest. For that reason, the proposed system uses a new approach to edge detection, by splitting up the task in two parts, namely edge pixel detection and edge segment or contour detection.

3.2.1 Edge Pixel Detection

After smoothing the image with a $3 \times 3$ Gaussian, edge pixel detection is performed separately in each monochromatic channel image $I_l$. The local orientation $v$ of pattern stripes may be established using two different approaches. The first is to compute the gradient direction $w$ and assume that $v$ is orthogonal to $w$. The local orientation is thus well-defined in the 2D image plane, as long as the gradient does not vanish. The second approach simply assumes that $v$ is the orientation of the projected light stripes, e.g. vertical for a pattern of vertical stripes. The orientation $w$ is then chosen accordingly. In that case, the algorithm might miss projected color edges whose imaged orientation deviates strongly from the projected one, i.e. edges projected on surfaces with certain position/orientation combinations. At the same time, the second approach is computationally more efficient and, more importantly, is unaffected by intrinsic edges whose orientation differs strongly from $v$. In any case, the following vectors and functions are well-defined almost everywhere in the image:

$$w_0 = \frac{w}{\|w\|}, \quad I_{lw}(t) = \frac{\partial I_l(x, y)}{\partial w_0}, \quad I_{lw}'(t) = \frac{\partial^2 I_l(x, y)}{\partial^2 w_0} (w \neq 0)$$

The derivative $I_{lw}(t)$ of $I_{lw}(x, y)$ in the direction of $w$ is computed and its local extrema, the transversal zero-crossings of $I_{lw}'(t)$, are established. The sub-pixel location of an extremum is chosen as the vertex of a parabola fitted to the ex-
tremum and its two unit-distance neighbors along w. Only these extrema are kept for further processing, a step known as non-extrema suppression. Performing it for each channel independently yields three images that contain only local extrema, called single-channel local extrema. Next, the three separate channel images are combined into a single image. To that end, single-channel extrema are grouped to multi-channel (local) extrema of the composite color signal: two or three single-channel local extrema, each from a distinct channel, are combined, if they share the same direction and are sufficiently close spatially. The sub-pixel position of such a multi-channel local extremum is computed using a weighted average of the positions of its components. Also, each multi-channel local extremum is classified into one of the 20 multi-channel color edge classes. Only multi-channel extrema are classified as edge pixels. Single-channel local extreme values that are not part of multi-channel extrema are ignored. In other words, a second non-extrema suppression is performed. Since projected edges are known to affect at least two color channels, and since the probability of a missed pattern related local extreme value is rather low (only a very small threshold is used), this step is very effective for filtering out noise without loosing any of the sought-after information.

3.2.2 Edge Segment Detection

Next, spatially adjacent edge pixels of the same class are traced to obtain edge segments. Classical tracing would filter out all remaining unwanted edge pixels except the ones that form segments themselves, e.g. the ones due to reflectivity edges. To cope with them, the identification problem is solved prior to tracing. The spatial adjacency relationship of edge pixels is established by determining sequences of n multi-channel extrema that share the same orientation w and lie along a line parallel to w. Next, we attempt to decode the resulting words, i.e. to check whether the
edge pixels, interpreted as code symbols according to their edge class, form a codeword when read from left to right along \( w \). Since most errors can be detected, if a codeword is found, it is very unlikely to represent an undetected error. Only edge pixels that are part of a valid codeword are hypothesized to reveal the location of a projected color edge orthogonal to \( w \) and only they are used as starting points of the tracing operation. A well-known problem with edge segment detection via tracing is streaking, the “breaking up of an edge contour caused by the operator output fluctuating above and below the threshold along the length of the contour” [20].

A common solution to overcome streaking is hysteresis with two thresholds [20]. Edge pixels above the higher threshold are accepted, the ones below the lower threshold are rejected, and the ones between the two are accepted only if their segment is connected to high-threshold edge pixels in both directions. Here, streaking is less a question of the threshold, and more a question of fluctuating between a valid and an invalid codeword. The algorithm uses an analogous approach, where edge pixels of the same edge class as the starting point pass the low threshold, and those that also represent the same letter of the same codeword pass the high threshold. A third threshold is used to pick up occasional edge pixels misclassified into a class close to the sought one according to a suitable signal space metric. The beginning and the end of a segment need to pass the high threshold, since often, edges of the same class originating from different projected edges join seamlessly, e.g. at the boundary between two objects of distinct height. Obviously two such edges cannot be distinguished by considering the segment by itself, this is why the high threshold is crucial for resolving such situations.

The proposed tracing is an error correction step, which is more reliable than replacing words that contain a detected error with the codeword that is closest according to some code space metric. Only edge pixels that are part of a ridge that exceeds a
small minimal length are used for further processing. This can be seen as another error detection step that detects errors that result in seemingly valid codewords, i.e. ones that are undetectable in the classical sense. Undetected errors are unlikely per se, and it is even more improbable that several of them form an edge segment. All in all, the algorithm determines color edges in the pattern image, yet not via a standard direct approach, but one designed to detect only color edge segments originating from the projected pattern. It does so by explicitly combining demodulation and decoding into one operation, rather than solving one after the other. This scheme was shown to be successful, even under noisy real-world conditions.

In Fig. 2, the various steps of the proposed detection algorithm of the projected color edges are depicted.

### 3.3 A Novel Range Sensor Based on the Method

Several sensors utilizing the presented method have been built, among them a commercial low-cost 3D sensor that is produced in larger quantities. It employs an inexpensive camera targeted at mobile phones as image sensor and a custom-made LED projector to generate the color pattern. A made-to-order interference filter is used as projection slide. For the experiments described below, i.e. the face recognition application, the version of the sensor shown in Fig. 3 has been used. In this case, the color images are acquired with a Basler 302fc single-chip Bayer-Pattern CCD RGB camera with resolution $780 \times 580$ pixels and a digital IEEE 1394a interface. A Panasonic LPT multimedia projector with a native resolution of $800 \times 600$ pixels projects the pattern on the scene. The length of the baseline between the camera and the projector is about $300\text{mm}$; the projector is rotated by a convergence angle of about $20^\circ$ towards the camera. The resulting working space is about
600mm x 400mm x 500mm (width x height x depth). Computed depth values are quantized into 16 bits. The software runs on an off-the-shelf PC.

Switching rapidly between a colored pattern and white light, a color image may be captured as well, approximately synchronized with the depth image. The acquired range images contain artifacts and missing points, mainly over areas that cannot be reached by the projected light and/or over highly refractive (e.g. eye-glasses) or low reflective surfaces (e.g. hair, beard).

4 3D Face Authentication System

In this section, we describe an application of the proposed real-time 3D sensor for robust user authentication based on 3D facial images. The system consists of several components. First, 3D data is exploited for the detection and localization of the face in the image. Then, we compensate for the pose of the face and the illumination of the scene, thus automatically generating neutral frontal images. These normalized images are subsequently used for authentication. In the following, the various steps of the face authentication algorithm are described in more detail.

4.1 Face Detection, Localization and Normalization

In the proposed system, face detection and localization is based solely on 3D data, thus exhibiting robustness under varying illumination and occlusion of facial features (e.g. beard, hair, etc). 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 incorporating a-priori constraints on the relative
dimensions of the body parts [21].

To localize the position of the face and estimate the pose, first the tip of the nose is accurately detected [21]. Then, 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 $2^\circ$. Face detection-localization runs near to real-time (about 10fps) and is used to provide the user with feedback regarding the quality of recorded depth images.

Specifically, a low-resolution sub-image from the center of the face is extracted using the acquired depth image and the center of the face estimated above. Then, this image is projected on a face-subspace [22], which allows to compute a measure of “face-ness”, i.e. how much does this sub-image resembles a human face. By providing the user with information about the quality of the image, we may facilitate frontal images and thus improve verification rates.

4.1.1 Pose Compensation

After 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, bringing the face in up-right orientation. The reference coordinate frame is defined during the training using gallery images, as will be described below. The transformation between the local and reference coordinate frames is further refined to pixel accuracy by apply-
ing the ICP surface registration algorithm [23] between the warped and a reference (gallery) depth image of the same person.

The rectified depth image contains missing pixels, some of which are filled exploiting face symmetry. Remaining missing pixel values are linearly interpolated from neighboring points. The interpolated depth map is subsequently used to rectify the associated color image (see Fig. 4) [21].

For the training phase, a 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, which define 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.

4.1.2 Illumination Compensation

Inspired by recent work on image-based scene relighting used for rendering realistic images [24], the proposed illumination compensation technique is based on the generation of a novel image relit from a frontal direction.

First, the scene illumination from a pair of color and depth images is recovered. 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 $L$ using a set of artificially generated bootstrap images [25].
Given the estimate of the light source direction $L$ and the frontal light direction $L_0$, the illumination compensated image $\tilde{I}_C$ is approximated by

$$\tilde{I}_C(u) = I_C(u) \frac{R(I_D, L, u)}{R(I_D, L_0, u)}$$

where $I_C, I_D$ are the input pose compensated color and depth images respectively, and $R$ is a rendering of the surface with constant albedo [25]. In other words, the illumination compensated image is given by multiplication of the input image with a ratio image. Fig. 5 illustrates the relighting of side illuminated images.

The same relighting procedure is 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.

### 4.2 Face Authentication

A multimodal classification scheme integrating both 2D and 3D images of the face is proposed in this paper. Two independent classifiers, one for color and one for depth images are used. Since classifiers based on depth data demonstrate robustness in handling cases where color-based classifiers show relative weakness (e.g. illumination variations, use of cosmetics), and vice versa (e.g. use of glasses, facial expressions), it is expected that their combination will exhibit enhanced authentication accuracy, compared to the use of each modality alone.

Among several state-of-the-art face recognition techniques examined, the Probabilistic Matching (PM) algorithm [22] gave the lowest error rates, while being com-
putationally efficient. The PM algorithm is applied to both depth and color data. In case of depth images, pixel values correspond to distance from the sensor rather than brightness.

The normalized color and depth images generated after pose and illumination compensation are the input to the color and depth classifier respectively. For practical reasons only, the red component of the color images is used. Pixel values are normalized to have zero mean and unit variance before classification. The scores (maximum-likelihood probabilities) returned by each classifier are subsequently normalized in the range $[0, 1]$ using the $\tanh$ normalization technique [26]. Fusion of color and depth data is achieved by simply adding the corresponding scores.

5 Experimental Results

Next, we evaluate the proposed 3D sensor and 3D face authentication system and provide experimental results measuring their performance in real world applications and experiments.

5.1 Evaluation of the proposed 3D acquisition system

5.1.1 3D Sensor Accuracy

This section describes experiments aimed at evaluating the accuracy of the sensor described in Section 3.3. It focuses on the depth error, here defined as the magnitude of the difference between the measured depth and the ground truth. The depth error is mainly due to the so-called localization error: light planes (the border between adjacent stripes) cannot be located with infinite accuracy in the image, but
only with a certain uncertainty. The statistical depth error of the system is quantified by acquiring several depth maps of a given static scene, i.e. by repeating the same measurement over and over, and analyzing the scatter of the depth measurements. Corresponding experiments yield a standard deviation of about $0.01\text{mm}$ to $0.04\text{mm}$. This small error is due to the fact that for a given set-up and static scene, the localization error has a large systematic component, when considering fixed pixel/scene patch combinations. In this case, only the imaging noise and other minor factors, such as the ambient illumination, represent a statistical component that causes the observed small scatter in the $z$-coordinate.

In a second experiment, the system acquires a depth map of a planar object - a glass plate with an imprinted black-and-white pattern - that is approximately parallel to the camera’s image plane, i.e. one located at an about constant depth. Subsequently, it performs a least-square fit of a 3D space plane to the acquired surface data. Table 1 lists the root-mean-square depth deviation of the measured points to this plane for several mean depth values. From these values, the quadratic relationship between depth and depth error, typical for a triangulation-based system, can be clearly seen.

In a third experiment, a planar calibration target is imaged. For each image, an external calibration is performed. Then, the target defines by definition the world-coordinate system’s $z_w = 0$ plane. This permits directly analyzing the scatter of the measured $z$ world coordinates. The difference to the previous experiment is that the measured values are not transformed, i.e. that absolute values are used. In particular, no plane fit is performed. Table 2 gives the corresponding statistical parameters of the measured depth values.

According to the above results, the depth accuracy of the 3D sensor is in the range of $0.1$ to $0.3\text{mm}$ standard deviation. In the course of the development of several
3D vision systems based on the sensor, this result has been confirmed as the accuracy that is practically achievable with the described geometric sensor set-up and components.

5.1.2 3D Sensor Data Rate

With the proposed technique, a sensor’s frame rate is not a fixed number, but it is influenced by several factors, the primary ones being the size of the scene in the color image and the exposure time of the color camera. Table 3 describes the data rate achieved for a set of representative image sequences on off-the-shelf 2.4 and 3.2 GHz PCs. Fig. 6 and 7 show exemplary intensity and depth images from these sequences, where the 3D data is visualized by mapping the depth value \( z \) of an image point, measured in millimeters, to a gray value \( g \) between 0 and 255 as follows: 

\[
g = (z \cdot 100.0) \mod (256)
\]

According to Table 3, the system misses real-time operation only by a narrow margin, if the scene is rather small as in the case of the gesture sequence. If the scene fills most of the image, as with the talking head sequence, a framerate of almost 18 fps is achieved. These results are consistent with the authors’ experience that the system delivers between 17 and 25 frames of dimension \( 780 \times 580 \) per second on a fast PC, with 20 fps being a representative, 17 fps a practical worst-case rate.

Empirically, the frame rate scales fairly linearly with the processor’s clock rate. This can be seen in Table 3, where it scales even super-linearly due to the faster front side bus of the 3.2 Ghz PC. True real-time operation will be possible in the near future. Alternatively, reducing the frame size turns the system into a real-time system at once. A corresponding experiment shows that the system consistently reaches a frame rate of more than 30 fps, if the image resolution is cropped to
5.2 Evaluation of the proposed 3D face authentication system

The focus of the experimental evaluation was to investigate the efficiency of the complete face authentication system in 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 two weeks 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 (±20°), images with glasses, and frontal images (see Fig. 8). In total, 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.

Table 4 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. In this case, rectification is performed by 2D affine warping of the images. As shown in Table 4, the proposed scheme results in authentication errors which are much lower compared to those achieved by manual image normalization, especially for images depicting pose variations. Moreover, a significant improvement is reported, when the illumination compensation algorithm is used. In Fig. 9, the receiver operator characteristic curve is shown. By combining color
and depth information, a correct recognition rate of 95% is obtained for a false acceptance rate of 0.5%.

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 4.3 GHz 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.

6 Conclusions

In conclusion, we have proposed a novel 3D acquisition system and its application on biometric user authentication. By exploiting 3D information, robust face authentication under heterogeneous conditions is achieved, using only low-cost devices.

One of the unique properties of the proposed 3D acquisition system is its ability for real-time acquisition of moving scenes. We are currently exploiting the availability of a stream of 3D images in several applications. One of the most promising is natural human-machine interaction, including head tracking and gesture recognition. Using range images we may easily detect and track the movement of the user even under background clutter and harsh illumination variations. Real-time un-obtrusive human motion capture and its applications in 3D graphics production and modelling is also one of the emerging applications of real-time 3D image acquisition, which we intend to address in future work.

Finally, a version of the 3D sensor that is based on near infrared illumination source is currently under investigation and is expected to overcome the obtrusiveness of the current approach using visible light.
7 Acknowledgement

This work has been supported by research project BioSec IST-2002-001766 (“Bio-
metrics and Security”, http://www.biosec.org) under Information Society Tech-
nologies (IST) priority of the 6th Framework Programme of the European Com-

References


Technische Universität München, Munich, Germany (2005).


322–327.


List of Figures

1 Color coded pattern. (a) Clipping from the projection pattern and (b) image captured by the color camera with color pattern projected on the face. 28

2 Detection of projected color edges: (a) color image, (b) image formed by combining the three single-channel local extrema images via logical AND, (c) traced ridges of correctly identified edge pixels. 29

3 A view of the 3D sensor. 30

4 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. 31

5 Illumination compensation examples. (a) Original image, (b) $R(I_D, L, u)$, (c) novel image relit by frontal light. 32

6 Test sequence “talking head with bright orange shirt”. The depth data is visualized by mapping the depth value of a given image point, measured in $100th$ of millimeter, to the gray value range of 0 to 255: $g = (zc \cdot 100.0)mod(256)$. One gray level period corresponds to 2.56mm. 33

7 Test sequence “rotating dark-green fan” (mapping of depth to gray levels as in Fig. 6). 34
Examples of face color images and corresponding depth maps depicting various expressions, pose, as well as illumination variations, during the recording of a single session. Bright pixels correspond to points closer to the camera and white pixels correspond to undetermined depth values.

Receiver operating characteristic curve showing false acceptance rate (FAR) versus correct recognition rate (1-FRR) for the two modalities and their combination.
Fig. 1. Color coded pattern. (a) Clipping from the projection pattern and (b) image captured by the color camera with color pattern projected on the face.
Fig. 2. Detection of projected color edges: (a) color image, (b) image formed by combining the three single-channel local extrema images via logical AND, (c) traced ridges of correctly identified edge pixels.
Fig. 3. A view of the 3D sensor.
Fig. 4. 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.
Fig. 5. Illumination compensation examples. (a) Original image, (b) $R(I_D, L, u)$, (c) novel image relit by frontal light.
Fig. 6. Test sequence “talking head with bright orange shirt”. The depth data is visualized by mapping the depth value of a given image point, measured in 100th of millimeter, to the gray value range of 0 to 255: \( g = (ze \cdot 100.0) \mod(256) \). One gray level period corresponds to 2.56\text{mm}.
Fig. 7. Test sequence “rotating dark-green fan” (mapping of depth to gray levels as in Fig. 6).
Fig. 8. Examples of face color images and corresponding depth maps depicting various expressions, pose, as well as illumination variations, during the recording of a single session. Bright pixels correspond to points closer to the camera and white pixels correspond to undetermined depth values.
Fig. 9. Receiver operating characteristic curve showing false acceptance rate (FAR) versus correct recognition rate (1-FRR) for the two modalities and their combination.
**List of Tables**

1. Depth error relative to fitted plane. 38
2. Depth error relative to $z_w = 0$ plane. All values are in $mm$. 39
3. Data rate of the 3D sensor. 40
4. 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). Even rows show results with the proposed pose (PC) and illumination compensation (IC) techniques, while odd rows are results with manual image rectification. 41
Table 1
Depth error relative to fitted plane.

<table>
<thead>
<tr>
<th>Mean depth in $mm$</th>
<th>647</th>
<th>778</th>
<th>871</th>
<th>936</th>
<th>984</th>
<th>1043</th>
<th>1079</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS in $mm$</td>
<td>0.09</td>
<td>0.12</td>
<td>0.16</td>
<td>0.20</td>
<td>0.23</td>
<td>0.28</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Table 2
Depth error relative to $z_{ref} = 0$ plane. All values are in mm.

<table>
<thead>
<tr>
<th>Mean camera $z$-coordinate</th>
<th>Mean world $z$-coordinate</th>
<th>RMS dev.</th>
<th>Min dev.</th>
<th>Max dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>871</td>
<td>-0.01</td>
<td>0.17</td>
<td>-0.84</td>
<td>+0.77</td>
</tr>
<tr>
<td>936</td>
<td>-0.02</td>
<td>0.23</td>
<td>-1.32</td>
<td>+2.30</td>
</tr>
</tbody>
</table>
Table 3
Data rate of the 3D sensor.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>FPS (2.4 Ghz)</th>
<th>FPS (3.2 Ghz)</th>
<th>Avg. number of range values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>12.3</td>
<td>18.3</td>
<td>192 000</td>
</tr>
<tr>
<td>Fan</td>
<td>14.5</td>
<td>21.3</td>
<td>142 000</td>
</tr>
<tr>
<td>Gesture</td>
<td>16.0</td>
<td>23.7</td>
<td>83 000</td>
</tr>
</tbody>
</table>
Table 4
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). Even rows show results with the proposed pose (PC) and illumination compensation (IC) techniques, while odd rows are results with manual image rectification.

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