4D Photogeometric Face Recognition with Time-of-Flight Sensors

Sebastian Bauer, Jakob Wasza, Kerstin Müller and Joachim Hornegger
Pattern Recognition Lab, Department of Computer Science
University Erlangen-Nuremberg, Germany
sebastian.bauer@informatik.uni-erlangen.de

Abstract

Methods for 2D/3D face recognition typically combine results obtained independently from the 2D and 3D data, respectively. There has not been much emphasis on data fusion at an early stage, even though it is at least potentially more powerful to exploit possible synergies between the two modalities. In this paper, we propose photogeometric features that interpret both the photometric texture and geometric shape information of 2D manifolds in a consistent manner. The 4D features encode the spatial distribution of gradients that are derived generically for any scalar field on arbitrary organized surface meshes. We apply the descriptor for biometric face recognition with a time-of-flight sensor. The method consists of three stages: (i) facial landmark localization with a HOG/SVM sliding window framework, (ii) extraction of photogeometric feature descriptors from time-of-flight data, using the inherent grayscale intensity information of the sensor as the 2D manifold's scalar field, (iii) probe matching against the gallery. Recognition based on the photogeometric features achieved 97.5% rank-1 identification rate on a comprehensive time-of-flight dataset (26 subjects, 364 facial images).

1. Introduction

The human face has emerged as one of the most promising biometrics. Facial recognition systems have the potential to become a key component in a variety of applications like identity authentication, access control, surveillance and security, or law enforcement [3, 6, 15]. Compared to conventional biometric technologies, such as fingerprint and iris imaging, face recognition is non-intrusive and requires a minimal extent of cooperation from the user. However, face recognition systems do not achieve accuracies of conventional technologies yet. In the past decade, the predominance of face recognition on 2D images has decreased in favor of 3D or multi-modal 2D/3D approaches [4, 27]. The incorporation of 3D data, providing the facial surface topology, has several benefits: On the one hand, it can eliminate illumination- and viewpoint-related issues and improve performance under these conditions. On the other hand, depth information simplifies face localization and segmentation and delivers additional information about physical dimensions. It is generally expected that the combination of complementary 2D and 3D information can lead the way to the demanding requirements of real-world applications. The promotion of the extensive and challenging face recognition grand challenge database [22] including 3D data (FRGC v2) by leading U.S. agencies, and experimental results from recent comparative studies like [4] indicate that multi-modal approaches seem more promising. However, to date, most efforts in the field of 2D/3D face recognition use a fairly simplistic fusion of results that are obtained independently from the 2D and 3D data, respectively.

In terms of surface imaging modalities, face recognition systems typically use passive stereo, structured light, and hybrid combinations of both technologies. However, each modality implies open issues for real-world application: Although significant progress has been made in stereo vision, systems require precise calibration and the recovery of depth from textureless regions or repetitive patterns is still an open research topic. The obtrusive nature and potential eye safety issues lower the appeal of structured light systems. Recent advances in active time-of-flight (ToF) surface imaging technology have opened new perspectives in its application on face recognition. The device is compact, portable, easy to integrate and delivers complementary 4D data (metric 3D coordinates + grayscale intensities) in real-time with a single sensor. In particular, the resolution (25k-40k points), framerate (25-40 Hz), eye-safe illumination (eye safety class 1 LEDs), field of view ($40^o \times 40^o$) and the flexible depth of field (up to 7 m) of recent sensors have potential for biometric applications.

Related Work ToF imaging has been proposed for face detection [2, 12], tracking [1, 9] and recognition [11, 19]. Typically, the use of methods from conventional 2D computer vision has been proposed, interpreting the range information as a 2D image. Böhme et al. applied the Vi-

ola and Jones face detector [25] on the range and intensity data and showed that a detector on combined data has a higher detection rate than detectors trained on either type of data alone [2]. In general, present ToF approaches take advantage of both the range and intensity information delivered by the sensor. However, in analogy with existing 2D/3D face recognition systems, results are produced for both the 2D and 3D information independently and subsequently combined into a final decision.

Face recognition with ToF sensors has been addressed rarely. This might result from the fact that sensors with decent resolutions have been introduced only recently. Meers and Ward [19] presented a system for face recognition that detects the nose tip as central keypoint and encodes the facial surface topology with spherical intersection profiles, similar to the descriptors proposed in [21]. Ding et al. [11] proposed a facial identification method based on histogram features that analyze the orientation of the mesh surface normals, c.f. [13]. Both papers lack a comprehensive quantitative evaluation, Ding et al. state a mean rank-1 recognition rate of 79.0% on a dataset from three subjects.

In this paper, we present a biometric face recognition system using ToF surface imaging. After facial landmark detection, both the photometric texture and geometric shape characteristics of the lower forehead and eye area are encoded in a complimentary – we call it *photogeometric* – descriptor. The contribution of the paper is twofold. First, we address the encoding of multi-modal data in a synergetic descriptor, merging shape and texture information at an early stage, rather than making a decision using each mode independently and combining decisions. In this context, we introduce a gradient operator based on a circular uniform surface sampling technique that is derived generically for any scalar field on 2D manifolds and applicable for arbitrarily (non-uniformly) organized mesh representations. For ToF applications, we propose the use of the intensity information as scalar field. Second, we provide a comprehensive quantitative evaluation of the proposed face recognition system on real ToF sensor data (754 facial images, 26 subjects).

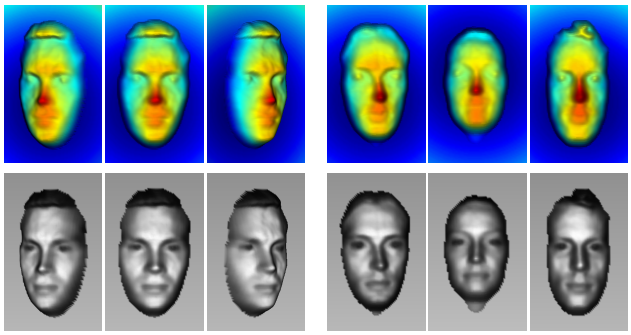

Figure 1. 3D visualization of preprocessed data from the ToF face recognition dataset. Upper row: color-coded range information (warm tones denote closeness to the sensor). Lower row: additional grayscale intensity information.

2. Method

The proposed recognition scheme is composed of three stages: First, we locate the eye positions as facial landmarks with a 2D histograms of oriented gradients (HOG) / support vector machine (SVM) sliding window framework (section 2.1) that is leading edge in terms of classification performance in the field of 2D object classification. Second, we extract photogeometric feature descriptors from ToF data, using the inherent grayscale intensity information of the sensor as the 2D manifold's scalar field (section 2.2). The descriptors encode the characteristics of the approximately rigid portion of the face around a landmark at the lower forehead, derived from the previously located eye positions. Third, the probes are matched against the gallery.

ToF Imaging Time-of-flight imaging directly acquires 3D surface information with a single sensor based on the phase shift ϕ between an actively emitted and the reflected optical signal [14]. Based on ϕ, the radial distance r from the sensor element to the object can be computed straightforward as: $r = \frac{c}{2f_{mod}} \cdot \frac{\phi}{2\pi}$, where f_{mod} denotes the modulation frequency, c the speed of light. The measurements of the ToF sensor with a resolution of $w \times h$ can be represented as a set of points or vertices $\boldsymbol{v}_i \in \mathbb{R}^3$,

$$\mathcal{V} = \{\boldsymbol{v}_i\}, \quad i \in \{1, \ldots, w \cdot h\} \tag{1}$$

In addition, as mentioned before, ToF sensors provide a grayscale intensity information for each vertex. Apart from the representation as a textured 3D point cloud, the range and intensity data from the sensor matrix can be interpreted as conventional 2D images.

Denoising An experimental study revealed that the performance rate of 3D face recognition systems decreases prominently at a depth resolution above 3 mm [5]. With regard to the trade-off between data denoising and preservation of topological structure, we perform ToF data preprocessing in a way that gives priority to the smoothness of the facial surface. Insufficient filtering results in topological artifacts that have a strong influence on the recognition performance. We suppose that this also applies for the results of previous work on face recognition [11, 19] where preprocessing was limited to median filtering. Our preprocessing pipeline consists of two edge-preserving bilateral filters [24], one operating on the range and one on the grayscale intensity data. In addition, we perform frame averaging within a temporal interval of 0.75 s. This interval is acceptable for authentication and recognition type

scenarios, where subjects are assumed to be cooperative. An additional benefit of the temporal averaging is the reduction of eye blink artifacts. Finally, based on a Delaunay triangulation, a surface mesh is generated from the denoised data (Fig. 1).

Head segmentation As stated before, metric 3D scene information simplifies face localization and segmentation. Depth thresholding is applied for an initial segmentation of the subject, yielding a binary foreground mask. Based on this mask, we detect the top of the head and incorporate a priori anthropometric knowledge to reduce the region of interest to the maximal dimensions of the face.

2.1. Eye Detection

In order to select a unique region that is subsequently analyzed with the photogeometric descriptor, facial landmarks are localized with a 2D HOG/SVM sliding window framework. In contrast to the common localization of the nose tip [12, 19], we detect the eyes being a distinctive feature in ToF grayscale intensity data (see Fig. 3). For face recognition (section 2.2), a central landmark on the lower forehead is then determined from the eye positions, and photogeometric descriptors are computed for the local neighborhood of this landmark.

The basic idea of the HOG descriptor is that local object appearance and shape is characterized by the distribution of intensity gradient directions. As the descriptor is well described in literature, we summarize here the structure of our implementation which closely follows the original detector [8]. The descriptor operates on the 2D intensity images, and evaluates rectangular patches in a sliding window fashion. First, the gradient directions and magnitudes are computed for each pixel of the image patch. Then, in order to measure local distributions of gradient values, the window is divided into 2×2 cells covering one quarter of the patch each. For each cell, the pixels are discretized into an orientation histogram according to its gradient direction. The contribution depends on the gradient magnitude at the respective pixel. Finally, the cell histograms are concatenated to the HOG descriptor. Contrast normalization is performed by scaling the feature vector to unit length.

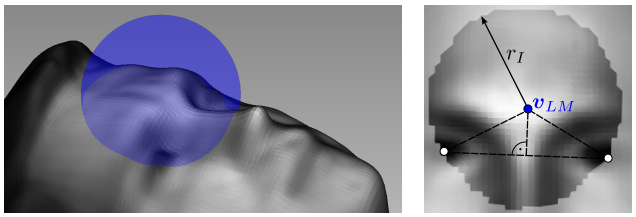

Figure 2. Left: Volume of interest (blue sphere). Right: ToF grayscale intensity scalar field, facial landmark $\boldsymbol{v}_{LM}$.

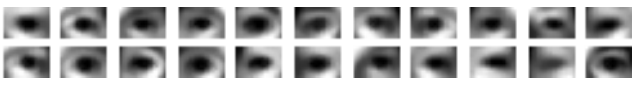

Figure 3. HOG eye samples, from ToF grayscale intensity images.

Based on the descriptor, a kernel SVM learns the implicit representation of the eye from examples and categorizes unseen image patches into one of two predefined classes: eye or non-eye. Part of the appeal for kernel SVMs is that nonlinear decision boundaries can be learnt by performing a linear separation in a high-dimensional feature space. We use a 2-norm soft margin kernel SVM with classification function $f(\boldsymbol{x})$,

$$f(\boldsymbol{x}) = sgn\left(\sum_{i=1}^{n_S} \alpha_i y_i K(\boldsymbol{s_i}, \boldsymbol{x}) + b_0\right) \qquad (2)$$

where $\alpha_i \geq 0$ denote the positive Lagrange multipliers, b_0 the bias, y_i the class label, $\boldsymbol{s_i}$ the n_S support vectors, $\boldsymbol{x}$ a HOG instance and $K(\boldsymbol{s}, \boldsymbol{x}) = e^{(-\gamma \|\boldsymbol{s}-\boldsymbol{x}\|^2)}$ the Gaussian kernel function. The kernel parameter γ and the weighting factor of slack variables are determined by an exhaustive grid search. A description of the general procedure is given e.g. in [7].

Based on the detected eye locations, we select a unique facial landmark $\boldsymbol{v}_{LM}$ on the lower forehead. Given the left and right eye positions, an isosceles triangle is induced, with the third vertex being located upwards in direction of the forehead (see Fig. 2). The height of the triangle is set to 25% of the inter-eye distance. The choice of this landmark ensures that only the approximately rigid portion of the face is evaluated.

2.2. Photogeometric Face Recognition

In this section, first, we introduce a gradient operator that computes a numerical gradient of a scalar function defined on a 2D manifold. The resulting gradient vectors hold both the photometric and geometric information in a consistent manner. Then, subsequently, we compute photogeometric 4D HOG descriptors that encode the spatial distribution of this gradient vector field in the neighborhood of the landmark $\boldsymbol{v}_{LM}$.

CUSS gradient Conventionally, in a 2D image, gradients are computed by differentiating the scalar function in two orthogonal directions. For 2D manifolds, we propose a gradient operator that is based on a circular uniform surface sampling (CUSS) technique. In contrast to the work of Zaharescu et al. [26], the operator is derived generically for any scalar field defined on a 2D manifold that can be represented by an arbitrary, possibly non-uniform mesh. Typically, the scalar field holds complementary information to

the surface data, e.g. any kind of photometric or texture information. For ToF applications, we propose the use of the grayscale intensity data. Benefits of the operator are:

- Invariance to mesh organization/representation
- Invariance to mesh density/resolution
- Direct applicability to parametric surfaces

Below, the proposed gradient operator $\nabla f(\boldsymbol{v}_i)$ is derived. Given is a scalar function $f(\boldsymbol{v}_i)$ that is defined for every mesh vertex $\boldsymbol{v}_i \in \mathcal{V}$. In case of ToF data, $f(\boldsymbol{v}_i)$ corresponds to the grayscale intensity information at the respective vertex. In a first step, the tangent plane T_i being defined by the corresponding normal $\boldsymbol{n}_i$ is determined for the vertex $\boldsymbol{v}_i$. In the next step, an arbitrary reference vector $\boldsymbol{a}_i \in \mathbb{R}^3$ is selected, $\boldsymbol{a}_i \in T_i$, $\|\boldsymbol{a}_i\|_2 = 1$. Then, a circular uniform sampling of the tangent plane T_i is performed via rotating $\boldsymbol{a}_i$ around $\boldsymbol{n}_i$ by the angle ϕ_s, yielding $\boldsymbol{R}_{\phi_s}\boldsymbol{a}_i$. Scaling the vectors $\boldsymbol{R}_{\phi_s}\boldsymbol{a}_i$ with an application-specific sampling radius r_s provides a set $\mathcal{P}$ of points $\boldsymbol{p}_s \in T_i$,

$$\mathcal{P} = \{\boldsymbol{p}_s | \boldsymbol{p}_s = \boldsymbol{v}_i + r_s \cdot \boldsymbol{R}_{\phi_s}\boldsymbol{a}_i\} \tag{3}$$

where $\boldsymbol{R}_{\phi_s}$ denotes the 3×3 rotation matrix for the angle $\phi_s = s \cdot \frac{2\pi}{N_s}$, $s \in \{1, \ldots, N_s\}$. N_s denotes the circular sampling density, $|\mathcal{P}| = N_s$. Finally, the surface sampling is performed by intersecting the mesh with rays that emerge from the points $\boldsymbol{p}_s$ and are directed parallel to $\boldsymbol{n}_i$ (see Fig. 4). The intersection points are denoted $\boldsymbol{m}_s$, the scalar function value $f(\boldsymbol{m}_s)$ is interpolated w.r.t. adjacent vertices. The CUSS gradient $\nabla f(\boldsymbol{v}_i)$ at the vertex $\boldsymbol{v}_i$ can then be expressed as:

$$\nabla f(\boldsymbol{v}_i) = \frac{1}{N_s} \sum_{s=1}^{N_s} \frac{f(\boldsymbol{m}_s) - f(\boldsymbol{v}_i)}{\|\boldsymbol{m}_s - \boldsymbol{v}_i\|_2} \cdot \boldsymbol{R}_{\phi_s}\boldsymbol{a}_i \tag{4}$$

The gradient vector field for the lower forehead and eye area is illustrated in Fig. 5.

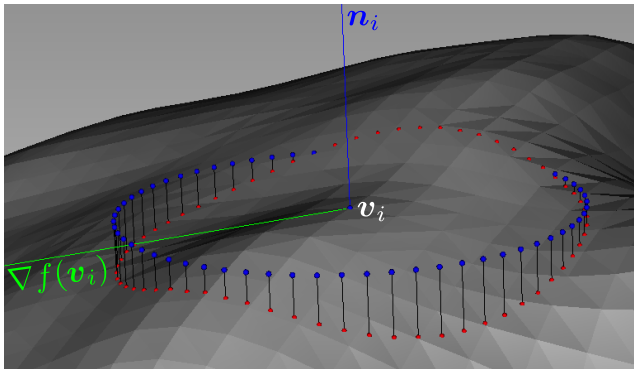

Figure 4. Illustration of the circular uniform surface sampling technique, computing the CUSS gradient $\nabla f(\boldsymbol{v}_i)$ for a vertex $\boldsymbol{v}_i$.

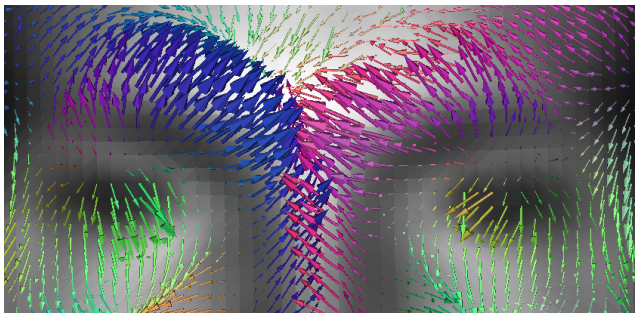

Figure 5. CUSS gradient vector field. The length of the arrows denotes the gradient magnitude $\|\nabla f(\boldsymbol{v}_i)\|_2$ at the respective vertex $\boldsymbol{v}_i$, the gradient orientation is additionally color-coded.

4D HOG Descriptor The 4D HOG descriptor encodes the spatial distribution of the CUSS gradient orientation within a spherical volume of interest (VOI), see Fig. 2. The VOI is defined within the radius r_I around a certain landmark vertex, and the intersection of the VOI with the object surface, denoted $\hat{\mathcal{V}} \subset \mathcal{V}$, is extracted. $\hat{\mathcal{V}}$ holds the set of vertices $\hat{\boldsymbol{v}}_j$ that reside within the VOI. For face recognition, we used the forehead landmark $\boldsymbol{v}_{LM}$, introduced in section 2.1. In a first step, the CUSS gradient vectors are projected on the three planes of a local coordinate system. Second, the projected vectors are binned in polar histograms, as applied in [26]. The proposed HOG descriptor is invariant to translation and rotation and can be interpreted as an extension of HOG to the case of scalar fields on 2D manifolds. The descriptor is not invariant to scale, as we incorporate the metric scale of the surface topology as an important characteristic. Below, the establishment of a unique local coordinate system and the computation of the 4D HOG descriptor are described.

Local coordinate system Establishing a unique local coordinate system is essential for rotation invariance of the descriptor. In addition to the surface normal $\boldsymbol{n}_{LM}$ at the landmark $\boldsymbol{v}_{LM}$, we define a second axis $\boldsymbol{m}_{LM}$ orthogonal to $\boldsymbol{n}_{LM}$ according to the following scheme: Each vertex $\hat{\boldsymbol{v}}_j \in \hat{\mathcal{V}}$ is projected onto the tangent plane T_{LM} of the landmark vertex, yielding $\boldsymbol{t}_j \in \mathbb{R}^3, \boldsymbol{t}_j \in T_{LM}$. Then, the weighted average over the vectors connecting $\boldsymbol{t}_j$ to $\boldsymbol{v}_{LM}$ is computed,

$$\boldsymbol{m}_{LM} = \frac{1}{w} \sum_{j=1}^{|\hat{\mathcal{V}}|} w_j \cdot (\boldsymbol{t}_j - \boldsymbol{v}_{LM}), \quad w = \sum_{j=1}^{|\hat{\mathcal{V}}|} w_j \tag{5}$$

where $w_j = \mathcal{G}(\|\boldsymbol{t}_j - \boldsymbol{v}_{LM}\|_2) \cdot \|\nabla f(\hat{\boldsymbol{v}}_j)\|_2$ denotes the gradient magnitude at vertex $\hat{\boldsymbol{v}}_j$, weighted by a Gaussian function $\mathcal{G}$ of the Euclidean distance between $\boldsymbol{t}_j$ and $\boldsymbol{v}_{LM}$. The local coordinate system is spanned by $\boldsymbol{n}_{LM}$, $\boldsymbol{m}_{LM}$ and $\boldsymbol{n}_{LM} \times \boldsymbol{m}_{LM}$.

Descriptor Based on this local coordinate system, we apply a projection scheme proposed by Zaharescu et al. [26]. The CUSS gradient vectors $\nabla f(\hat{\boldsymbol{v}}_j)$ are projected onto the three tangent planes of $\boldsymbol{n}_{LM}$, $\boldsymbol{m}_{LM}$ and $\boldsymbol{n}_{LM} \times \boldsymbol{m}_{LM}$. The projection planes are divided into n_c equally-sized circular segments. For each plane and circular segment, a polar histogram (with n_o bins) encodes the orientation distribution of the projected gradient vectors. The 4D HOG descriptor $\boldsymbol{h}$ is then made up by a concatenation of these polar histograms, $\boldsymbol{h} \in \mathbb{R}^d$, where $d = 3 \cdot n_c \cdot n_o$.

Matching Using the 4D HOG feature representation, each face corresponds to a point in the feature space $\mathbb{R}^d$. One-to-many matching from probe to gallery is performed by nearest neighbor matching. In section 3.1, we compare the performance of Euclidean, Pearson correlation and Jensen-Shannon divergence similarity metrics, respectively. The Jensen-Shannon divergence [16] is a symmetric version of the Kullback-Leibler divergence and a popular method of measuring the similarity between two probability distributions such as the proposed 4D HOG features, being a concatenation of histograms.

3. Experiments

Qualitative and quantitative evaluation is performed on real ToF data. Below, we introduce a comprehensive ToF face recognition dataset. Then, we present quantitative results in terms of (i) receiver operating characteristics (ROC) for the 2D HOG/SVM eye detection and (ii) cumulative match characteristics (CMC, identification scenario) for the 4D HOG descriptor matching. Last, we comment on runtime complexity.

Due to the unique characteristics of ToF sensor data, we have acquired a specific face recognition dataset. It covers facial data from 26 subjects, male and female, standing at a distance of about 60 cm in front of the camera. For each subject, we captured three sequences of neutral expressions and three sequences of non-neutral expressions (smiling, laughing, looking angry). A sequence consists of 150 frames. For reasons of pose variation, the subject was asked to leave the field of view of the camera and re-position in front of the device in between each of the three neutral expression captures. Data were captured using a CamCube 3.0[1] ToF camera with a resolution of 200×200 pixels, a framerate of 40 Hz, a modulation frequency of 20 MHz, an infrared wavelength centered at 870 nm, an integration time of 250 μs, and a lens with $40^o \times 40^o$ field of view. At the distance of 60 cm, the noise level of the range measurements is in the scale of $\sigma \approx 5$ mm, the average face resolution about 60×80 pixels.

Dataset	Dimension (images)
Gallery	26
Probe, neutral expressions	364
Probe, non-neutral expressions	390

Table 1. ToF face recognition dataset.

For the experiments, all sequences were preprocessed with the denoising pipeline described in section 2. Frames are averaged over a temporal interval of 0.75 s. The bilateral filter parameters were chosen empirically such that the remaining average standard deviation of the filtered range measurements did not exceed 1 mm. From each preprocessed sequence, we selected 5 images, at an interval of 20 frames. For each subject, the first image of the neutral expression sequence was set as gallery image, the remaining 14 images with neutral and 15 images with non-neutral expression as probes. The entire dataset includes 26 gallery images, 364 probes with neutral and 390 probes with non-neutral expression, see Table 1. In addition, for each image, we manually labeled the eye positions as ground truth annotation.

In the following section, the proposed methods for eye detection and face recognition are evaluated on this ToF dataset. The dataset is available from the authors for non-commercial research purposes and can be used for quantitative evaluation of face recognition approaches. To the best of our knowledge, this is the first comprehensive dataset of mid-resolution facial ToF data. The face detection dataset of the EU project ARTTS[2] contains 1300 facial and 3600 non-facial images, at a resolution of 24×24 pixels. This single-frame low-resolution dataset is not suitable for face recognition.

3.1. Recognition Performance

Eye detection Based on the annotated eye locations, positive samples were extracted from the 2D grayscale intensity images. In order to make maximum use of these samples, patches were horizontally mirrored, giving a total number of $N_{pos} = 26 \cdot 30 \cdot 2 \cdot 2 = 3120$ positive samples. For the negative set, we randomly sampled non-eye regions of the dataset, producing $N_{neg} = 26 \cdot 30 \cdot 10 = 7200$ negative patches. The patch dimension (18×12 pixels) was empirically set w.r.t. to the average eye size at the subject-camera distance of 60 cm. The orientation histogram is divided into six evenly spaced angular bins. For quantitative evaluation, we performed 2-fold cross validation. Fig. 6 shows the ROC curve of the kernel SVM, plotting detection rate over false alarm rate. At a false alarm rate of 0.01, the classifier achieved a detection rate of $>$99.5%.

[1] http://www.pmdtec.com/

[2] http://www.artts.eu/publications/3d_tof_db/

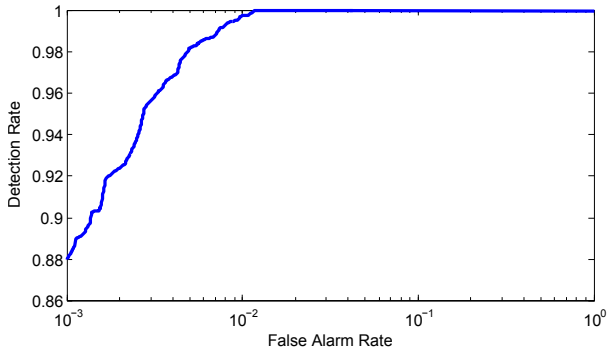

Figure 6. ROC curve (semi-log plot) of the 2D HOG eye detection, from 2-fold cross validation.

Photogeometric face recognition For the following face recognition experiments, the VOI radius was set to $r_I = 50$ mm. In general, a greater VOI results in an improved recognition performance. However, we empirically selected r_I with the provision that the VOI did not include any strand of hair for all subjects within the dataset. Experience shows that such artifacts can lead to spuriously increased recognition rates when the hair artifacts remain constant for all acquisitions of one individual. This results from the fact that the transition between skin and hair holds strong CUSS gradient information, hence the descriptor would be more discriminative than for the bare forehead. In practice, however, daily variations in hairstyle would have a strong impact on the descriptor, downgrading the system performance.

In order to determine appropriate descriptor parameters for our application - having a strong influence on the dimensionality of the feature vector - we systematically studied the effects of the CUSS gradient sampling radius r_s, the number of circular segments n_c and the number of orientation bins n_o for the polar histograms by performing a parameter grid search. The resulting rank-1 identification rates w.r.t. n_c and n_o are shown in Fig. 7, for a fixed CUSS gradient radius of r_s = 12.5 mm. Significantly smaller or larger radii r_s resulted in a performance deterioration. The figure illustrates that both fine orientation coding and fine spatial segmentation turn out to be essential for good performance. Increasing the number of orientation bins improves performance significantly up to about 6 bins, but makes little difference beyond this. Throughout the evaluation below, we refer results to our default descriptor that is selected according to the result of the grid search. It has the following parameter properties: $r_I = 50$ mm, $r_s = 12.5$ mm, $n_c = 6$, $n_o = 6$.

First, we demonstrate the robustness of the orientation of the local coordinate system w.r.t. a spatial tolerance of the landmark position. For the set of vertices adjacent to $\boldsymbol{v}_{LM}$, the respective local coordinate systems were determined and their axes compared to the reference $\boldsymbol{n}_{LM}$, $\boldsymbol{m}_{LM}$ and $\boldsymbol{n}_{LM} \times \boldsymbol{m}_{LM}$. For the set of neutral expressions, the angular mean deviation was $\Delta(\phi) = 6.8 \pm 5.1^o$. This deviation is small compared to the angular range of the circular segments (60^o, $n_c = 6$). Box plots for the individual axes are shown in Fig. 8.

For quantitative analysis, we evaluate the proposed face recognition method on the identification task, being a standard procedure for the face recognition vendor test (FRVT) [10, 23]. Fig. 9 shows the CMC curve for the closed-set identification scenario, for different similarity metrics. The 4D HOG descriptor achieved rank-1 identification rates of 97.5% and 90.5% for the probe sets with neutral and non-neutral expressions, respectively. A drop of identification rate on data with non-neutral expressions was expected,

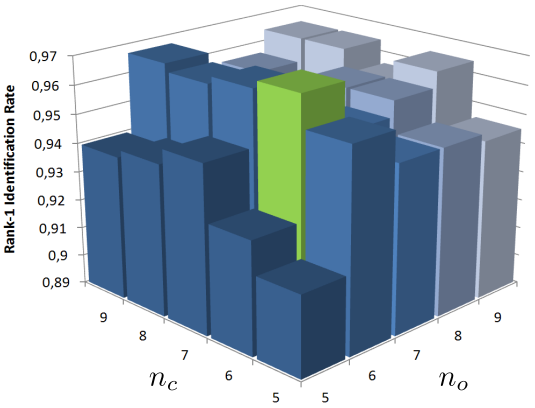

Figure 7. Rank-1 identification rates for the parameter grid search over n_c, n_o (Euclidean similarity metric). Fixed parameters: $r_I = 50$ mm, $r_s = 12.5$ mm. The parameter combination of the default descriptor used in the experiments is labeled in green.

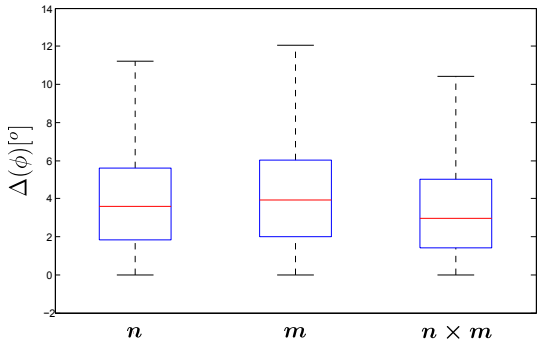

Figure 8. Box plot of the angular variation $\Delta(\phi)$ of the local coordinate system axes. From left to right: $\Delta(\phi)_{\boldsymbol{n}}$, $\Delta(\phi)_{\boldsymbol{m}}$ and $\Delta(\phi)_{\boldsymbol{n}\times\boldsymbol{m}}$. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers (within 1.5 times the interquartile range).

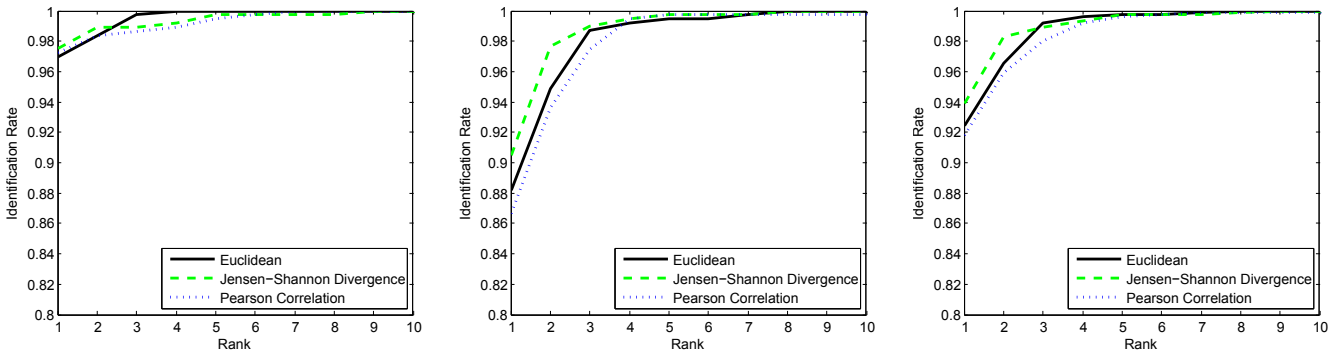

Figure 9. CMC curves, face identification. From left to right: gallery vs. neutral, gallery vs. non-neutral, gallery vs. all probes. Similarity metrics: Euclidean (black, continuous), Jensen-Shannon divergence (green, dashed), Pearson correlation (blue, dotted).

since 3D face recognition systems are more sensitive to expressions compared to 2D approaches. However, the decrease is small compared to previous work [17, 20] due to the fact that only the approximately rigid portion of the face from just below the nose up to the forehead is used in our approach. Table 2 summarizes the recognition results for different evaluation settings. The best performance was achieved using the Jensen-Shannon divergence, as proposed in section 2.2. Challenging results on ToF face recognition do not exist in literature yet. However, the recognition rates indicate a performance in the scale of state-of-the-art 2D/3D results [17, 18, 20] on data acquired with sensors from the Minolta Vivid series[3] that deliver highly accurate and dense 3D information compared to ToF cameras.

3.2. Computational Complexity

Giving recognition performance top priority, we use descriptors that are rather expensive from a computational point of view. The calculation of a 4D HOG descriptor, including the computation of the CUSS gradients, takes about 650 ms (default descriptor, $N_s = 64$) on an Intel Core2 Duo CPU @ 2.80 GHz, 4.0 GB RAM. The extraction of descriptors for all vertices within a local neighborhood (10 mm, approx. 50 descriptors) only slightly increases the runtime (by 50 ms) since the gradient vector field is already computed. We expect acceleration potential from the inherent parallelism of the CUSS sampling and gradient projections.

[3] http://www.konicaminolta.com/

Evaluation setting	Rank-1 Identification [%]
Gallery vs. Neutral	97.5
Gallery vs. Non-Neutral	90.5
Gallery vs. All	93.9

Table 2. Overview of the rank-1 identification rates (Jensen-Shannon divergence similarity metric).

Limitations As stated by Bowyer et al. [4], face recognition systems are generally susceptible to variations in illumination, occlusions and facial makeup. These issues apply for the proposed method, too. As for all surface imaging modalities, measuring 3D shape with ToF sensors is not completely illumination independent. On the one hand, specular reflections are likely to cause saturation effects, resulting in corrupted intensity and range information. On the other hand, the measurement reliability can decrease in regions with poor reflective properties.

4. Conclusion

In this paper, we have presented a ToF system for biometric face recognition, relying on a *photogeometric* descriptor that encodes both the photometric texture and topological shape information of a 2D manifold in a common representation. The system can be considered as a 2D/3D approach, where multi-modal information is provided by a single sensor. We have introduced a gradient operator based on a circular uniform surface sampling technique that is defined for any scalar fields on 2D manifolds. The gradient operator can be used with arbitrary organized surface mesh representations and is directly applicable to parametric surfaces. Experimental results show the robustness of the CUSS gradient and the photogeometric 4D HOG descriptor for face recognition w.r.t. noise and low-resolution range data. The rank-1 identification rates of 97.5% and 90.5% on data with neutral and non-neutral expressions, respectively, indicate the feasibility of face authentication and recognition with ToF sensors. In future research, we will investigate the benefit of using non-spherical or multiple local volumes of interest, and further explore photogeometric features as a generic object descriptor for different modalities and applications in the domain of computer vision and pattern recognition beyond face recognition.

References

[1] M. Böhme, M. Haker, T. Martinetz, and E. Barth. Head tracking with combined face and nose detection. In *Proceedings of the IEEE International Symposium on Signals, Circuits & Systems*, pages 1–4, 2009.

[2] M. Böhme, M. Haker, K. Riemer, T. Martinetz, and E. Barth. Face detection using a time-of-flight camera. In *Proceedings of the DAGM Workshop on Dynamic 3D Imaging*, pages 167–176. Springer, 2009.

[3] K. W. Bowyer. Face recognition technology: Security versus privacy. *IEEE Technology and Society*, pages 9–20, 2004.

[4] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

[5] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Face recognition using 2d and 3d facial data. In *Multimodal User Authentication Workshop*, pages 25–32, 2003.

[6] R. Chellappa, P. Sinha, and P. Phillips. Face recognition by computers and humans. *Computer*, 43(2):46–55, 2010.

[7] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines and other kernel based learning methods*. Cambridge University Press, 2006.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[9] S. B. Göktürk and C. Tomasi. 3d head tracking based on recognition and interpolation using a time-of-flight depth sensor. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 211–217, 2004.

[10] P. Grother, R. J. Micheals, and P. J. Phillips. Face recognition vendor test 2002 performance metrics. In *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication*, pages 937–945. Springer, 2003.

[11] A. S. H. Ding, F. Moutarde. 3d object recognition and facial identification using time-averaged single-views from time-of-flight 3d depth-camera. In *Eurographics Workshop on 3D Object Retrieval*, pages 39–46, 2010.

[12] M. Haker, M. Böhme, T. Martinetz, and E. Barth. Scale-invariant range features for time-of-flight camera applications. In *Proceedings of the CVPR Workshop on Time-of-Flight-based Computer Vision*, 2008.

[13] G. Hetzel, B. Leibe, P. Levi, and B. Schiele. 3d object recognition from range images using local feature histograms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, page 394, 2001.

[14] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. *Computer Graphics Forum*, 29:141–159, 2010.

[15] S. Z. Li, A. K. Jain, T. Huang, Z. Xiong, and Z. Zhang. Face recognition applications. In *Handbook of Face Recognition*, pages 371–390. Springer, 2005.

[16] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[17] X. Lu, A. K. Jain, and D. Colbry. Matching 2.5d face scans to 3d models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):31–43, 2006.

[18] M. H. Mahoor and M. Abdel-Mottaleb. Face recognition based on 3d ridge images obtained from range data. *Pattern Recognition*, 42(3):445–451, 2009.

[19] S. Meers and K. Ward. Face recognition using a time-of-flight camera. In *Proceedings of the International Conference on Computer Graphics, Imaging and Visualization*, pages 377–382, 2009.

[20] A. S. Mian, M. Bennamoun, and R. Owens. Keypoint detection and local feature matching for textured 3d face recognition. *International Journal of Computer Vision*, 79(1):1–12, 2008.

[21] N. Pears, T. Heseltine, and M. Romero. From 3d point clouds to pose-normalised depth maps. *International Journal of Computer Vision*, 89(2-3):152–176, 2010.

[22] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947–954, 2005.

[23] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:831–846, 2010.

[24] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 839–846, 1998.

[25] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[26] A. Zaharescu, E. Boyer, K. Varanasi, and R. P. Horaud. Surface feature detection and description with applications to mesh matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[27] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.