Scale-Invariant Range Features for Time-of-Flight Camera Applications

Martin Haker, Martin Böhme, Thomas Martinetz, and Erhardt Barth
Institute for Neuro- and Bioinformatics, University of Lübeck
Ratzeburger Allee 160, 23538 Lübeck, Germany
http://www.inb.uni-luebeck.de

Abstract

We describe a technique for computing scale-invariant features on range maps produced by a range sensor, such as a time-of-flight camera. Scale invariance is achieved by computing the features on the reconstructed three-dimensional surface of the object. The technique is general and can be applied to a wide range of operators. Features are computed in the frequency domain; the transform from the irregularly sampled mesh to the frequency domain uses the Nonequispaced Fast Fourier Transform. We demonstrate the technique on a facial feature detection task. On a dataset containing faces at various distances from the camera, the equal error rate (EER) for the case of scale-invariant features is halved compared to features computed on the range map in the conventional way. When the scale-invariant range features are combined with intensity features, the error rate on the test set reduces to zero.

1. Introduction

The fact that the apparent size of an object in a camera image changes with the distance of the object from the camera leads to one of the fundamental problems in computer vision: Finding scale-invariant image features, i.e. features that, by their mathematical formulation, are unaffected by image scale (for an example of a recent approach, see [7]). Achieving scale invariance usually requires increased algorithmic complexity and additional computation. For example, the image can either be scanned for objects of different sizes, or it can be transformed into scale-space [5], where the feature extraction is computed individually at different levels of scaling. In both cases, the treatment of objects at different scales has to be made explicit within the algorithm. In this paper, we suggest a novel approach to the problem of scale-invariance: If we use a range sensor – such as a time-of-flight (TOF) camera [8] – to image the object, we can compute features directly on the reconstructed surface of the object in 3D space. In this way, the features become scale-invariant, because the 3D reconstruction – unlike the image of the object – does not undergo scale changes as the object moves towards or away from the camera.

We are particularly interested in the TOF camera as a basis for this type of approach because it provides a range map that is perfectly registered with an intensity image in a single device, making it easy to create detectors based on a combination of range and intensity features. The TOF camera uses an active illumination to produce a distance measurement at each pixel at 20 frames per second or more, depending on the integration time. The distance is computed from the phase shift of an emitted sinusoidally modulated infrared signal and the reflected signal, which depends on the scene. The intensity image corresponds to the amplitude of the reflected signal and is hence often referred to as the *amplitude image*. It depends on the objects' reflectivity and quantifies the confidence one has in the distance measurement, which can be readily inferred from the signal to noise ratio.

Naturally, one can compute image features on the regular image grid of both range and amplitude images directly [3]. Note, however, that interpreting the range map as an array of height values measured over a regular grid is equivalent to the *weak perspective* assumption, i.e. to assuming that the total depth variation within the object is small compared to the distance of the object from the camera. If this assumption is violated, the geometry of the surface reconstructed using weak perspective will differ markedly from the true geometry of the object. Furthermore, the size of an object in the image changes with its distance to the camera.

Alternatively, if the intrinsic camera parameters are known, we can apply the inverse of the camera projection to the range data, thus obtaining an actual sampling of the object's surface in three-dimensional Cartesian coordinates. Obviously, the measured object does not undergo any scaling in the 3D scene if it is moved towards or away from the camera; instead, only the spatial sampling frequency decreases as the distance to the camera is increased (see Fig. 1). It is important to note, however, that the sampling grid in this representation is no longer regular; the spacing between two samples depends on the distance of the relevant part of the

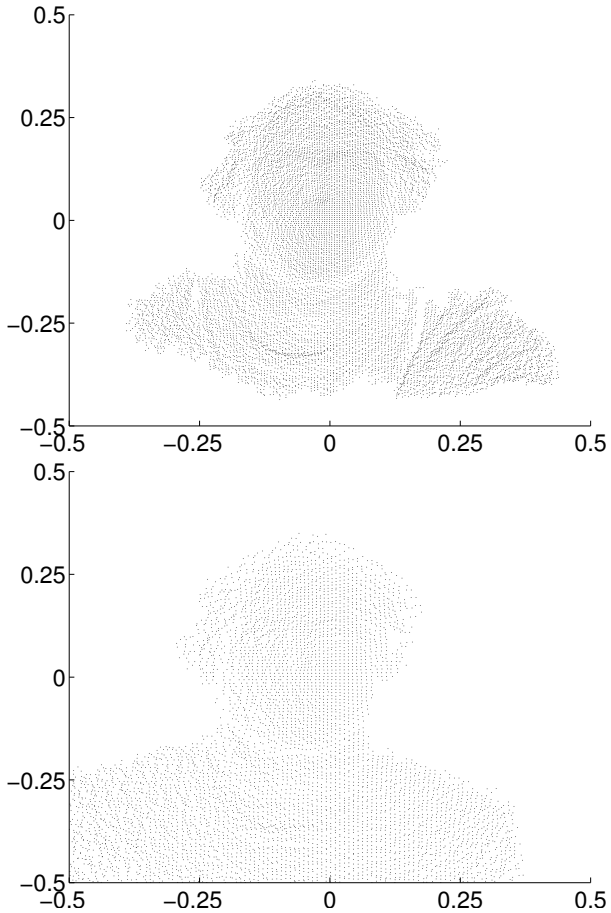

Figure 1. **Top:** Sampling of a face 35 cm from the camera. **Bottom:** Sampling of a face at 65 cm from the camera. Note that the spatial sampling frequency is significantly lower compared to the face at 35 cm, but the physical size of the face in 3D space is still the same.

object to the camera. Many techniques for extracting image features, such as convolution with a filter kernel, require a regular sampling and can thus no longer be used.

To overcome this problem, we compute the features not in the spatial domain, but in the frequency domain. We transform the sampled object shape to a frequency domain representation using the Nonequispaced Fast Fourier Transform (NFFT, see Sect. 2), an efficient algorithm for computing the Fourier transform of a signal sampled on an irregular grid. Thus, any feature computation that can be performed in the Fourier domain can now be evaluated efficiently. The main advantage of this approach is that any filter operation has, in theory, the same effect on an object independently of the distance of the object to the camera. The NFFT has been used for image processing tasks such as CT and MRI reconstruction (see e.g. [9]), and it has also been used to extract features from vector fields for visualization [11]. However, to our knowledge, the approach of using the NFFT to compute scale-invariant features for classification is novel.

We demonstrate this approach using a set of geometric features called *generalized eccentricities*, which are related to mean and Gaussian curvature (see Sect. 3). When evaluated on the image in the conventional way, these features are sensitive to scale changes; however, when evaluated on the object surface using the NFFT, the features are invariant to scale. We verify this using a synthetic test object in Sect. 5.

Finally, we use these features for a facial feature tracking problem. In previous work [2], we tackled this problem using features evaluated conventionally on the camera image; this solution had limited robustness towards scale variations. The new scale-invariant features yield greatly improved detection at varying distances from the camera, as we show in Sect. 5.

2. Nonequispaced Fast Fourier Transform (NFFT)

2.1. Definition

As noted in the introduction, we need to compute the Fourier transform of a function sampled on a nonequispaced grid. To do this, we use the NFFT [10], an algorithm for the fast evaluation of sums of the form

$$f(\mathbf{x}_j) = \sum_{\mathbf{k} \in I_{\mathbf{N}}} \hat{f}_{\mathbf{k}} \mathrm{e}^{-2\pi \mathrm{i} \mathbf{k} \mathbf{x}_j} \quad , \tag{1}$$

where the $\mathbf{x}_j \in [-\frac{1}{2}, \frac{1}{2})^d, j = 1, \ldots, M$ are arbitrary nodes in the spatial domain (of dimension d), the $\mathbf{k}$ are frequencies on an equispaced grid, and the $\hat{f}_{\mathbf{k}}$ are the corresponding Fourier coefficients. The equispaced frequency grid $I_{\mathbf{N}}$ is defined as

$$I_{\mathbf{N}} := \left\{ \mathbf{k} = (k_t)_{t=1,\ldots,d} \in \mathbb{Z}^d : -\frac{N_t}{2} \leq k_t < \frac{N_t}{2}, t = 1, \ldots, d \right\} , \tag{2}$$

where $\mathbf{N} = (N_1, \ldots, N_d)$ is the so-called *multibandlimit*, which specifies the band limit along each dimension. (Note that all N_t must be even.)

We refer to (1) as the Nonequispaced Discrete Fourier Transform (NDFT). The Discrete Fourier Transform (DFT) (with equispaced nodes in the spatial domain) can be obtained by setting the $\mathbf{x}_j$ to the nodes of the grid $\mathbf{x} = (\frac{k_t}{N_t})_{t=1,\ldots,d}$, $k_t \in \{-\frac{N_t}{2}, \ldots, \frac{N_t}{2} - 1\}$.

Equation (1) describes the transform from the frequency domain to the spatial domain. In the case of the equispaced DFT, because the matrix that describes the transform is unitary, the same algorithm can be used for the opposite transform (from the spatial to the frequency domain). This is not true in the nonequispaced case; here, to transform from the

spatial to the frequency domain, i.e. to find Fourier coefficients $\hat{f}_{\mathbf{k}}$ such that evaluation of (1) will yield certain given values $f(\mathbf{x}_j)$, we need a second algorithm (see [6]), which is based on a combination of the conjugate gradient method with the NFFT. Note that this algorithm (which transforms from the spatial to the frequency domain) is sometimes referred to as the "inverse NFFT", whereas the term "inverse" is otherwise usually applied to a transform from the frequency to the spatial domain.

2.2. Applying the NFFT to Range Data

We assume that the object surface, reconstructed from the range data by inverting the camera projection, is given in Cartesian coordinates x, y, z, where the x-y-plane is parallel to the image plane and the z-axis is parallel to the camera's optical axis. To apply the NFFT to this data, we interpret the z-coordinate as a function $z(x, y)$ of the x-y-coordinates; hence, the x-y-coordinates define the grid nodes for the NFFT. As noted in the previous section, these nodes need to lie in the interval $[-\frac{1}{2}, \frac{1}{2}) \times [-\frac{1}{2}, \frac{1}{2})$. Generally, this means that the x-y-coordinates of the surface points need to be scaled to this interval. We wish to use the same scaling for all images so that the interpretation of the Fourier domain remains the same.

To define this scaling, we will introduce the concept of an *equivalence range*; we want to choose such a scaling that, for an object at the equivalence range e, the effect of applying a particular transfer function to the FFT spectrum and to the NFFT spectrum is the same. The correct scaling is computed by intersecting the camera's field of view with a plane perpendicular to the view direction at distance e from the camera, yielding a rectangle; the x-y-plane is then scaled such that this rectangle fits exactly within the interval $[-\frac{1}{2}, \frac{1}{2}) \times [-\frac{1}{2}, \frac{1}{2})$.

Note that the x-y-coordinates of points beyond the equivalence range may lie outside the interval $[-\frac{1}{2}, \frac{1}{2}) \times [-\frac{1}{2}, \frac{1}{2})$; these points are discarded. The equivalence range thus needs to be chosen such that the resulting clipping volume is large enough to contain the objects of interest. The centroid of these objects should be shifted to $x = 0$, $y = 0$ to ensure they are not clipped.

Another point of note is that it is advisable, if possible, to segment the foreground object of interest and apply the NFFT only to the points belonging to that object. There are various reasons for doing this: (i) Steep edges between the foreground and background can lead to ringing artefacts. (ii) The grid nodes in the background region are spaced further apart; the greater the spacing between grid nodes, the lower the frequency where aliasing begins. (iii) Passing fewer points to the NFFT reduces the computational requirements.

Finally, note that the transform from the spatial domain to the frequency domain is often an underdetermined operation. In this case, the NFFT computes the solution with minimal energy, meaning that the background region, where there are no grid nodes, is implicitly set to zero. To avoid steep edges between the foreground and the background, we subtract a constant offset from the z values so that the maximum z value becomes zero.

3. Geometric Features

The features we employ for demonstrating our approach to scale invariance are related to the Gaussian curvature and are referred to as *generalized eccentricities* [1]. These features were already applied to TOF images in [3, 2] for the task of nose detection and tracking, using the weak perspective assumption as described in Sect. 1. In the following, we will extend this work to the perspective camera model to obtain scale invariant features.

For the definition of the invariant geometric features, we interpret the range data as a particular type of surface, the *Monge patch* or the *2-1/2-D* image, defined as a function $f : \mathbb{R}^2 \to \mathbb{R}^3, (x, y) \mapsto (x, y, f(x, y))$. Note that this definition holds for both the weak perspective and the true perspective projection model. In the first case, x and y specify a position on the image sensor and $f(x, y)$ is the corresponding range value. In the latter case, $(x, y, f(x, y))$ are the Cartesian coordinates of a point on the object surface (i.e. $f(x, y)$ is the z coordinate).

On this data model, the generalized eccentricities [1] are defined as

$$\epsilon_n^2 = \left(c_n(x, y) * f(x, y)\right)^2 + \left(s_n(x, y) * f(x, y)\right)^2 \quad , \quad (3)$$

for $n = 0, 1, 2, \ldots$. Here, $c_n(x, y)$ and $s_n(x, y)$ are convolution kernels corresponding to the transfer functions

$$C_n = \mathrm{i}^n A(\rho) \cos(n\theta) \quad , \quad S_n = \mathrm{i}^n A(\rho) \sin(n\theta) \qquad (4)$$

(defined in terms of polar coordinates ρ and θ), where $A(\rho)$ is a radial filter tuning function. The radial filter tuning function can be combined with a low-pass filter for noise reduction where the noise filter, e.g. a Gaussian low-pass filter, should be adapted to the distribution of noise inherent in the data.

As discussed in more detail in [3], the generalized eccentricities provide basic and reliable alternatives to the Gaussian curvature K and the mean curvature H for the purpose of surface classification.

In particular, the measures ϵ_n for $n = 0$ and $n = 2$ can be used to distinguish between the six well-known surface types in the feature space spanned by ϵ_0 and ϵ_2. Figure 2 shows where the different surface types lie in feature space. For example, because the nose is a local minimum in the range data, we would expect the corresponding pixels to lie in the region labeled *pit*. Conversely, since the nose tends to be a local maximum in the intensity data, we would expect to find the corresponding pixels in the region labeled *peak*.

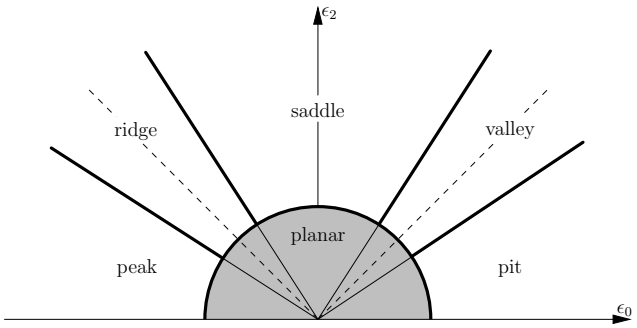

Figure 2. Discrimination of the six surface types *pit*, *peak*, *saddle*, *valley*, *ridge*, and *planar* within the feature space spanned by ϵ_0 and ϵ_2.

4. Nose Detector

In previous work, we have used the geometric features for the task of nose detection and tracking, and we briefly review the algorithm described in [3]. For each pixel in an image we compute the generalized eccentricities ϵ_0 and ϵ_2 and, thus, map it to the feature space given in Fig. 2. In feature space, noses are characterized by a bounding box which is learned from a set of labeled training data. During classification, a given pixel is said to belong to the tip of a nose iff it is mapped into this bounding box in feature space.
We computed feature values only on the foreground object of interest. For the FFT-based approach, the background was set to a constant value (the maximum value occurring in the foreground); for the NFFT-based approach, background pixels were simply discarded, as described in Sect. 2.2.
To segment the foreground object, we first applied an adaptive threshold to the amplitude image (see [3]); then, we applied a range threshold (at the top 20th percentile of pixels segmented in the first step) to ensure a smooth transition between foreground and background all along the border.

5. Experimental Results

The algorithms were implemented in Matlab; the NFFT 3.0 library [4] (implemented in C) was used to compute the NFFT.

5.1. Synthetic Data

To begin, we will examine the feature values computed on a synthetic test object using both the classical scale-dependent FFT-based approach and the scale-independent NFFT-based approach. We synthesized range images of a sphere at various distances from the virtual camera and computed the generalized eccentricity ϵ_0 using the FFT- and NFFT-based approaches.
Figure 3 shows the value of ϵ_0 at the apex of the sphere as a function of distance. (ϵ_2 is not shown because it is identically zero for objects that, like the sphere, exhibit the same curvature in all directions.) It is apparent that while the feature value changes noticeably with distance for the FFT-based approach, it remains essentially constant for the NFFT-based approach. Note also that at the equivalence range of $e = 0.5\,\mathrm{m}$, both approaches compute the same feature value.

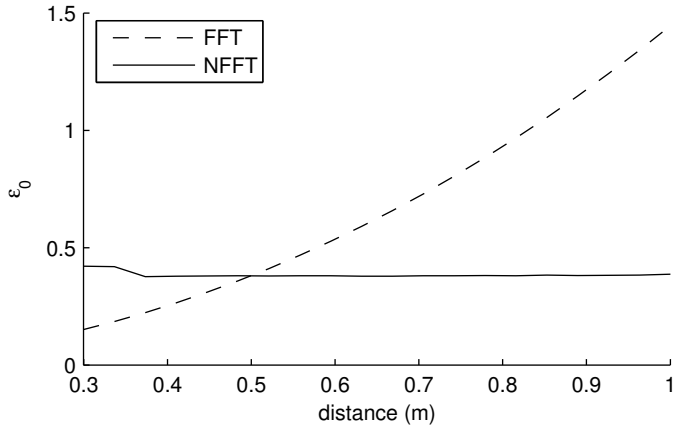

Figure 3. Generalized eccentricity ϵ_0, computed on a synthetic test image of a sphere, as a function of distance from the camera.

5.2. Real-World Data

To evaluate the performance of the features on a real-world problem, we compare the detection rates of the nose detector [3] on images of an SR3000 TOF camera using the NFFT-based algorithm and the original FFT-based version, respectively. In both cases, the training was done on a database of three subjects who were imaged at a fixed distance of 60 cm from the camera. During evaluation, the detector had to generalize to a dataset of 87 face images showing a different subject. The test images were taken at distances ranging from 35 to 70 cm. Figure 4 shows two examples of range maps from the test set along with the scale-invariant features.
In the case where only the range data of the TOF images is considered, the results are given in Fig. 5. Here, the NFFT-based algorithm achieves an EER of 20% in comparison to 39% in case of the FFT-based version and, thus, clearly yields a significant improvement in detection performance.
We have shown the detection results on features extracted from the range data alone to make the effect more clearly visible; for optimal detection performance, we would additionally use the same type of features computed on the intensity data. As we will discuss in Sect. 6, we still compute the intensity features in the conventional way using the FFT, and they are thus not scale-invariant. Nevertheless, when they are combined with the FFT-based range features, we obtain an EER of 4%; when NFFT-based range features

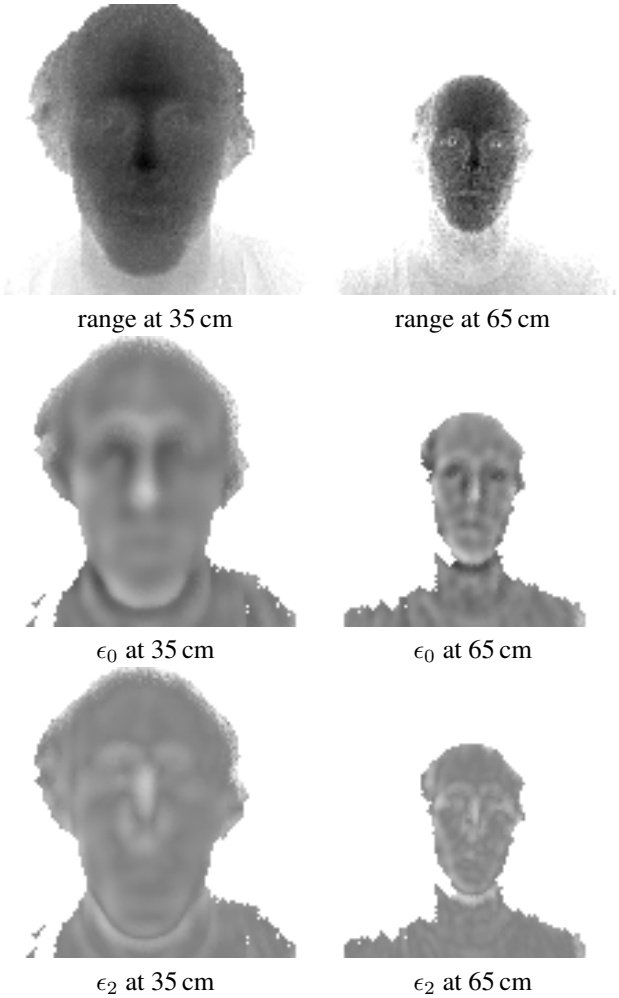

Figure 4. Two different range image samples from the test set, taken at 35 cm (left column) and 65 cm distance (right column). The corresponding features ϵ_0 and ϵ_2 computed at foreground pixels via the NFFT are shown, respectively.

are used, the EER drops to 0%, i.e. there are no errors on the test set – a larger test set would be required to measure a more meaningful EER. (As a point of note, the EER on the intensity features was 78%; it is only the combination of range and intensity features that yields low error rates.)

It should be mentioned that, while the NFFT has the same asymptotic running time as the FFT, it is slower by a relatively large constant factor. Our Matlab implementation, running on a 2.66 GHz E6750 Intel CPU, requires 0.1 s to compute the FFT-based features, versus 5 s for the NFFT-based features. A C implementation of the FFT-based detector runs at camera frame rates [2]; the NFFT-based detector is currently too slow for this type of application. However, we believe there are ways of achieving the same effect at lower computational cost; in the meantime, we see NFFT-based scale-invariant features as an attractive technique for non-interactive applications.

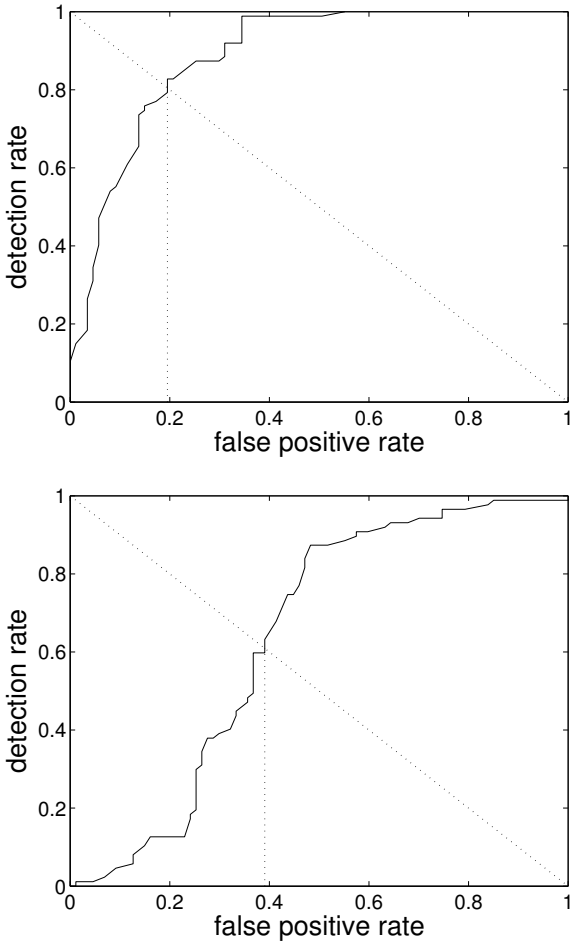

Figure 5. **Left:** ROC curve showing detection rate vs. false positive rate for the nose detection task using the NFFT. **Right:** The ROC curve obtained on the same database using the FFT. The detection rate indicates the percentage of images in which the nose has been identified correctly, whereas the false positive rate denotes the percentage of images where at least one non-nose pixel has been misclassified. Thus, strictly speaking, the curves do not represent ROC curves in the standard format, but they convey exactly the information one is interested in for this application, that is, the accuracy with which the detector gives the correct response per image.

6. Discussion

Features computed directly on the three-dimensional geometry of the object are, by their nature, scale-invariant. As we have shown, this allows for a more robust classification than when the same features are computed directly on the range map, where they are sensitive to scale variations. We have demonstrated this using a specific set of features, the generalized eccentricities, but the method itself is very general and can be applied to a wide range of operators.

We have used our technique to implement a facial feature

detector using a time-of-flight camera. The detector generalizes from faces presented at a fixed training distance to a test set containing different faces at different distances. It achieves good detection performance, making no errors on the test set. Two important factors that help us achieve this result are the scale-invariant features and the fact that the time-of-flight camera provides a perfectly registered intensity image in addition to the range map. The combination of range and intensity data yields substantially better classification results than either type of data alone.

Currently, we still compute the intensity features on the image, where they are sensitive to scale variations. Ideally, we would like to compute these features on the object surface, too. This is, however, a slightly more complicated problem, because intensity is a function of the position on a two-dimensional sub-manifold (the object surface) in three-dimensional space; the geometry of this sub-manifold must be taken into account when computing the intensity features. This is an avenue for future work.

On a general level, our key point is this: The perspective transformation that is inherent in the image formation process causes scale variations, which present additional difficulties in many computer vision tasks. This is why, in our view, the time-of-flight camera is an attractive tool for computer vision: In effect, it can act as a digital orthographic camera, thereby simply eliminating the problem of scale variations.

Acknowledgment

This work was developed within the ARTTS project (www.artts.eu), which is funded by the European Commission (contract no. IST-34107) within the Information Society Technologies (IST) priority of the 6th Framework Programme. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

[1] Erhardt Barth, Terry Caelli, and Christoph Zetzsche. Image encoding, labeling, and reconstruction from differential geometry. *CVGIP: Graphical Models and Image Processing*, 55(6):428–46, November 1993.

[2] Martin Böhme, Martin Haker, Thomas Martinetz, and Erhardt Barth. A facial feature tracker for human-computer interaction based on 3D TOF cameras. In *Dynamic 3D Imaging – Workshop in Conjunction with DAGM*, 2007. (in print).

[3] Martin Haker, Martin Böhme, Thomas Martinetz, and Erhardt Barth. Geometric invariants for facial feature tracking with 3D TOF cameras. In *Proceedings of the IEEE International Symposium on Signals, Circuits & Systems (ISSCS)*, volume 1, pages 109–112, Iasi, Romania, 2007.

[4] Jens Keiner, Stefan Kunis, and Daniel Potts. NFFT 3.0, C subroutine library. `http://www.tu-chemnitz.de/~potts/nfft`, 2006.

[5] Jan J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–70, 1984.

[6] Stefan Kunis and Daniel Potts. Stability results for scattered data interpolation by trigonometric polynomials. SIAM *Journal on Scientific Computing*, 29:1403–1419, 2007.

[7] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the IEEE International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999.

[8] Thierry Oggier, Bernhard Büttgen, Felix Lustenberger, Guido Becker, Björn Rüegg, and Agathe Hodac. SwissRanger^{TM} SR3000 and first experiences based on miniaturized 3D-TOF cameras. In Kahlmann Ingensand, editor, *Proc.* 1^{st} *Range Imaging Research Day*, pages 97–108, Zurich, 2005.

[9] Daniel Potts and Gabriele Steidl. A new linogram algorithm for computerized tomography. *IMA Journal of Numerical Analysis*, 21(3):769–782, 2001.

[10] Daniel Potts, Gabriele Steidl, and Manfred Tasche. Fast Fourier transforms for nonequispaced data: A tutorial. In John J. Benedetto and Paulo J. S. G. Ferreira, editors, *Modern Sampling Theory: Mathematics and Applications*, pages 247–270. Birkhäuser, Boston, 2001.

[11] Michael Schlemmer, Ingrid Hotz, Vijay Natarajan, Bernd Hamann, and Hans Hagen. Fast Clifford Fourier transformation for unstructured vector field data. In *Proceedings of the Ninth International Conference on Numerical Grid Generation in Computational Field Simulations*, pages 101–110, 2005.