Geometric Invariants for Facial Feature Tracking with 3D TOF Cameras

Martin Haker, Martin Böhme, Thomas Martinetz, and Erhardt Barth
Insitute for Neuro- and Bioinformatics
University of Lübeck, Germany
Email: {haker, boehme, martinetz, barth}@inb.uni-luebeck.de

Abstract— **This paper presents a very simple feature-based nose detector in combined range and amplitude data obtained by a 3D time-of-flight camera. The robust localization of image attributes, such as the nose, can be used for accurate object tracking. We use geometric features that are related to the intrinsic dimensionality of surfaces. To find a nose in the image, the features are computed per pixel; pixels whose feature values lie inside a certain bounding box in feature space are classified as nose pixels, and all other pixels are classified as non-nose pixels. The extent of the bounding box is learned on a labeled training set. Despite its simplicity this procedure generalizes well, that is, a bounding box determined for one group of subjects accurately detects noses of other subjects. The performance of the detector is demonstrated by robustly identifying the nose of a person in a wide range of head orientations. An important result is that the combination of both range and amplitude data dramatically improves the accuracy in comparison to the use of a single type of data. This is reflected in the equal error rates (EER) obtained on a database of head poses. Using only the range data, we detect noses with an EER of 0.66. Results on the amplitude data are slightly better with an EER of 0.42. The combination of both types of data yields a substantially improved EER of 0.03.**

I. Introduction

In this paper we focus on object tracking, more precisely, on the task of nose and head tracking. The most common forms of digital images that are utilized in computer vision to solve such tasks are intensity and range images. The former type is by far the most popular, which is mainly due to the low cost of the corresponding image sensors.

However, within the last decade a novel type of image sensor – the 3D time-of-flight (TOF) camera – has been developed that fuses the acquisition of both intensity and range data into a single device at a relatively low cost. The future pricing of such cameras is expected to be comparable to a standard webcam. In contrast to webcams, the 3D TOF camera simplifies the determination of geometrical properties of the 3D scene significantly, thus it is worth investigating methods that make explicit use of the available data.

We will discuss geometrically invariant measures that are suitable for identifying facial features in a 3D TOF camera image. Based on these features, we construct a simple nose detector and test its performance on range and intensity data individually, as well as on the combination of these two types of data. An important result is that the performance of the detector on the combined range and intensity data is substantially better than on either type of data alone. This underlines the potential of 3D TOF cameras for machine vision applications.

Previous work has already identified the nose as an important facial feature for tracking e.g. in [1] and [2]. In the former approach the location of the nose is determined by template matching, under the assumption that the surface around the tip of the nose is a spherical Lambertian surface of constant albedo. This approach gives very robust results under fixed lighting conditions and at a fixed distance of the user from the camera. The latter approach is based on a geometrical model of the nose that is fitted to the image data.

We also consider the nose as being very well suited for head tracking, because the nose is obviously a distinctive characteristic of the human face. In terms of differential geometry, the tip of the nose is the point of maximal curvature on the object surface of a face. A benefit of analyzing the 3D surface in terms of differential geometry is that a major portion of differential geometry is concerned with the description of invariant properties of rigid objects. Although curved surface patches have been shown to be unique [3], [4], Gaussian curvature is rarely used as a feature because its computation is based on first and second order derivatives, which are sensitive to noise. We propose alternative features that can be related to generalized differential operators. These features, which are computed per pixel of the input image, are used to decide for each input pixel if it corresponds to the tip of the nose based on the simple application of thresholds learned from a set of labeled training data.

We will first review and motivate the geometric features and evaluate them with respect to their suitability for the specific task of nose detection. Then, we will discuss the robustness of the method by presenting results on a database of head pose images acquired using a MESA SR3000 TOF camera [5].

II. Geometric Invariants

For the definition of invariant geometric features we will restrict ourselves to a special type of surface known as the *Monge* patch or the *2-1/2-D* image. Such surfaces are defined as a function $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ in the following manner:

$$(x, y) \mapsto (x, y, f(x, y)). \tag{1}$$

The ARTTS project 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.

This is the natural definition for digital intensity images, as each pixel value is bound to a fixed position on the image sensor without explicit reference to a coordinate representation in the 3D world. In the case of range data, however, each pixel is associated with explicit 3D coordinates via the geometry of the optical system of the camera. Nevertheless, we will assume a weak-perspective camera model for both range and amplitude data of the 3D TOF camera, because we do not expect a great difference in range for the pixels of interest. (Within a frontal face we do not expect any range differences greater than 5 cm, which would roughly correspond to a relative shift of only 5% in the coordinates x and y at a distance of 1 meter from the camera if we assumed a perspective camera model.) Thus, we can treat both types of data as Monge patches, which results in a simplified mathematical formulation and comparable results for range and amplitude data.

The features derived for the above data model are mainly due to a conceptual framework for image analysis which was introduced in [6] and [7]. Within this framework, image regions are associated hierarchically, with respect to their information content, with 0D (planar), 1D (parabolic), and 2D (elliptic/hyperbolic) regions of a Monge patch. Naturally, the concept of curvature is fundamental to this representation. Within this framework, the authors proposed a set of measures that provide basic and reliable alternatives to the Gaussian curvature K and the mean curvature H for the purpose of surface classification.

Let us first recall the definition of Gaussian curvature for a Monge patch:

$$K = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2}. \tag{2}$$

In case only the sign of the curvature is relevant, one can rely on the DET-operator D, which can be formulated in terms of the determinant of the Hessian

$$(h_{ij}) = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}. \tag{3}$$

This amounts to the numerator of (2). Thus the DET-operator takes the following form:

$$D = f_{xx} f_{yy} - f_{xy}^2 = \det(h_{ij}) = d_1 d_2. \tag{4}$$

Here, d_1 and d_2 denote the eigenvalues of the Hessian. Rearranging the first part of the formula [4] yields

$$\begin{aligned} D &= \tfrac{1}{4}(f_{xx} + f_{yy})^2 - \tfrac{1}{4}(f_{xx} - f_{yy})^2 - f_{xy}^2 \\ &= (\Delta f)^2 - \epsilon^2, \end{aligned} \tag{5}$$

where Δf denotes the Laplacian and ϵ is referred to as the eccentricity, which is defined as

$$\epsilon^2 = \frac{1}{4}(f_{xx} - f_{yy})^2 + f_{xy}^2. \tag{6}$$

The above formulation yields a relationship of the curvature to the Laplacian and the eccentricity. A generalized representation of the operators Δf and ϵ can be achieved in the Fourier domain by defining the generalized eccentricity ϵ_n via

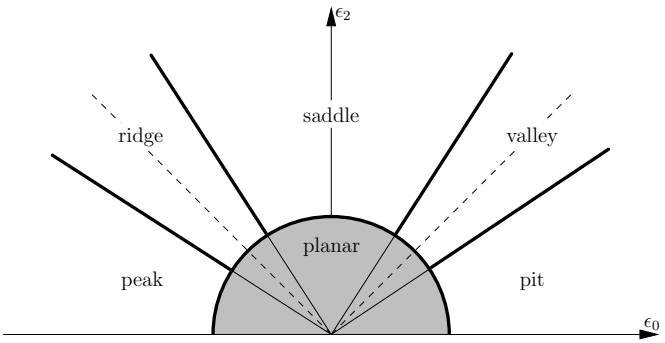


Fig. 1. Discrimination of the six surface types *pit*, *peak*, *saddle*, *valley*, *ridge*, and *planar* within the feature space spanned by ϵ_0 (Δf) and ϵ_2 (ϵ).

the following filter functions in polar coordinates ρ and θ, where $A(\rho)$ represents the radial filter tuning function:

$$\begin{aligned} C_n &= \mathrm{i}^n A(\rho) \cos(n\theta), \\ S_n &= \mathrm{i}^n A(\rho) \sin(n\theta). \end{aligned} \tag{7}$$

Recall that the transfer functions of partial derivatives are of the form $(\mathrm{i}f_x)^n$ and $(\mathrm{i}f_y)^n$, where f_x and f_y represent the spatial frequencies and n denotes the order of differentiation. Even-order partial derivatives correspond to real transfer functions, whereas odd-order partial derivatives correspond to imaginary transfer functions.

The transfer functions in (7) correspond to convolution kernels $c_n(x, y)$ and $s_n(x, y)$ in the image domain. Using these, we obtain the generalized eccentricity

$$\epsilon_n^2 = \left(c_n(x, y) * l(x, y)\right)^2 + \left(s_n(x, y) * l(x, y)\right)^2 \tag{8}$$

for $n = 0, 1, 2, \ldots$, which corresponds to $|\Delta f|$ for $n = 0$ and to the eccentricity ϵ for $n = 2$, when $A(\rho) = (2\pi\rho)^2$. The gradient is defined by ϵ_n for $n = 1$ and $A(\rho) = 2\pi\rho$. In a purely geometrical interpretation, all measures ϵ_n are positive, and as a result one cannot distinguish between convex and concave curvature using ϵ_0 and ϵ_2. An extension to this formulation in [4] justifies the use of ϵ_0 with the sign of Δf, i.e. $\epsilon_0 = -c_0 * l$.

For practical applications, the radial filter tuning function $A(\rho)$ can be combined with a low-pass filter, e.g. Gaussian blurring of the form $G(\rho, \sigma) = \exp(-\pi\rho^2/4\sigma^2)$. Ideally, the low-pass filter should be adapted to the distribution of noise inherent in the data.

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. Fig. 1 shows where the different surface types lie in feature space. 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*.

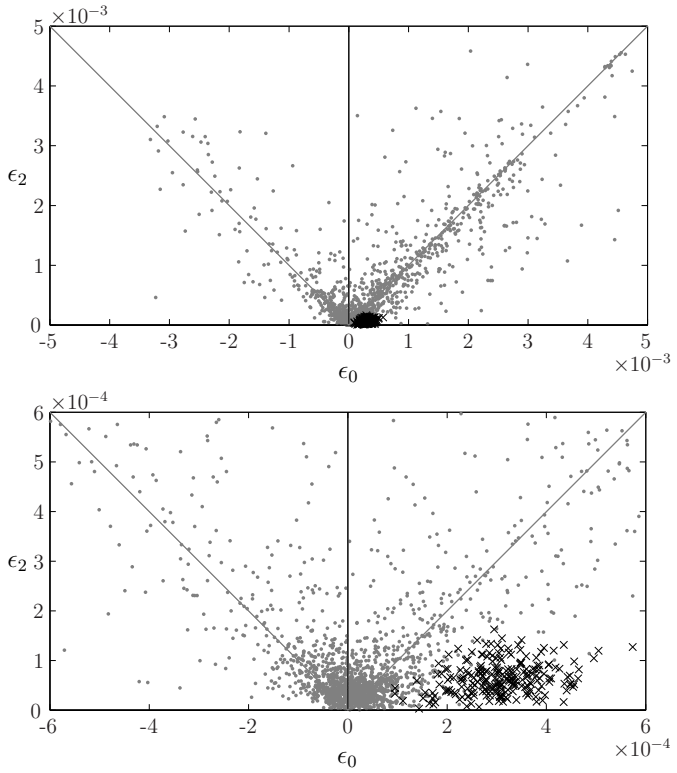


Fig. 2. **(Top)** Distribution of feature points for pixels taken from range data of the SR3000 camera projected into the 2D feature space spanned by ϵ_0 (Δf) and ϵ_2 (ϵ). **(Bottom)** The feature space around the origin at a higher resolution. The black crosses represent feature points corresponding to the nose tip of various subjects and clearly cluster in the region associated with the surface type *pit* as expected. The grey dots represent randomly chosen non-nose pixels.

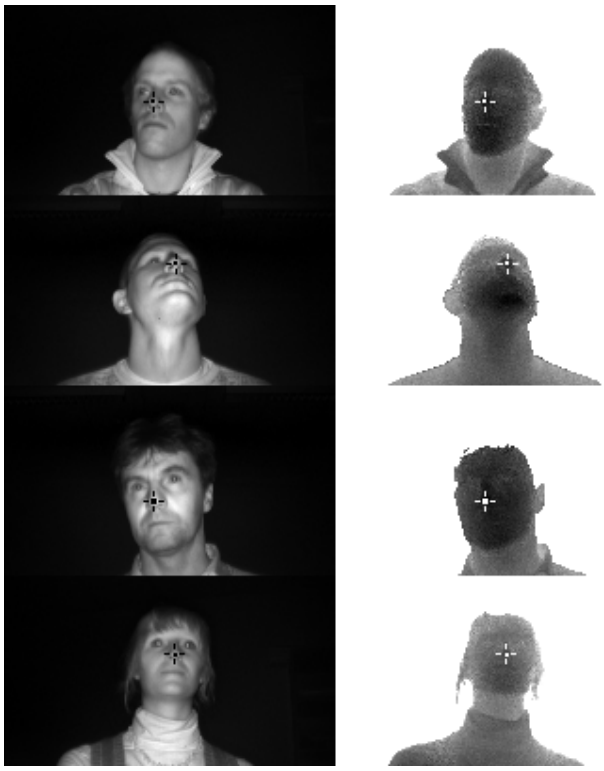


Fig. 3. Sample images from the database of head poses. The amplitude data (left column) and the range data (right column) are given for four subjects. All pixels identified as nose pixels by our detector are marked in each image, the cross simply highlighting the locations.

III. Feature Selection

The interpretation of Fig. 1 not only demonstrates how the features ϵ_n can be interpreted intuitively, but it also shows how the interpretation can be used to select meaningful features for the task at hand.

The top plot in Fig. 2 displays feature points for pixels computed on range data in the feature space spanned by ϵ_0 and ϵ_2. Only pixels with high amplitude, i.e. pixels belonging to an object that is close to the camera, were considered. First, it is noticeable that the frequency of occurrence is ordered with respect to 0D structures around the origin, 1D structures along the diagonals, and 2D structures. Thus, only a small portion of pixels belongs to curved regions. However, in the bottom plot in Fig. 2, one can observe that the feature points associated with the tip of the nose cluster nicely and correspond to the surface type *pit* as one would expect. Qualitatively similar results can be observed on the amplitude data, the major difference being that nose pixels cluster in the region associated with the surface type *peak*.

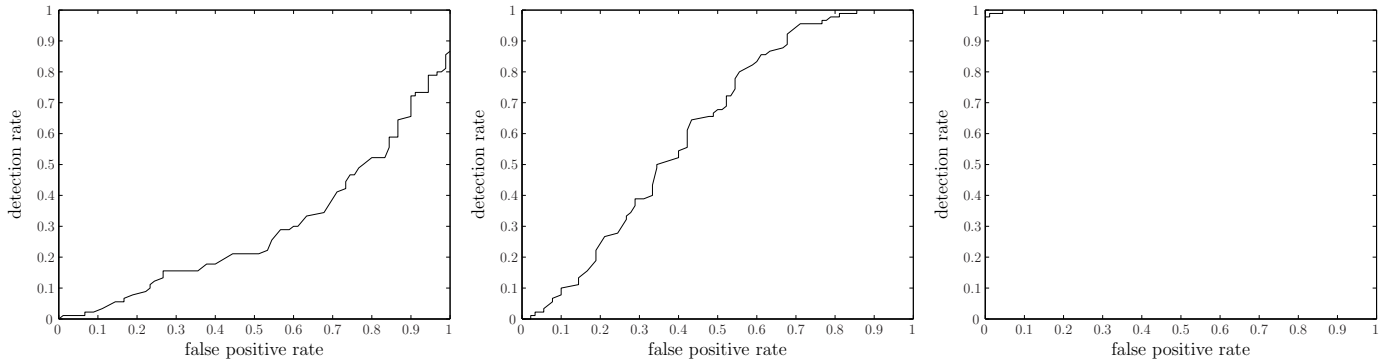
IV. Nose Detector

We use the geometric invariants ϵ_0 and ϵ_2, as introduced in Section II, to construct a nose detector. It decides per pixel of the input image if it corresponds to the tip of a nose or not. In other words, each pixel of the input image is mapped to a d-dimensional feature space, where d is the number of features considered. Within this feature space, we estimate a bounding box that encloses all pixels associated with the tip of a nose. It is important to mention that the bounding box should be estimated in polar coordinates due to the interpretation of the feature space (see Fig. 1).

The estimation is done in a supervised fashion based on a set of feature points computed for pixels that were hand-labeled as belonging to the tip of the nose. The extent of the bounding box within each dimension of the feature space is simply taken as the minimum and the maximum value of the corresponding feature with respect to all training samples. A softness parameter can be introduced to control the sensitivity of the detector.

To detect a nose within an image, the features are computed for each pixel, and a pixel is classified as the tip of a nose if its feature point lies within the estimated bounding box in the feature space. Despite the simplicity of this approach, we obtain very accurate and robust results, as we will show below.

The input for each feature computation was either the range or the amplitude data of an image from the SR3000 camera. The raw camera data was preprocessed by scaling it to the interval $[0, 1]$. Then, a threshold computed using Otsu's method [8] was applied to the amplitude data to separate the foreground from the background. The background was



(a) Range data. (b) Amplitude data. (c) Range and amplitude data.

Fig. 4. ROC curve of detection rate vs. false positive rate on range data (a), amplitude data (b), and the combination of both (c). The detection rate gives 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.

set to a fixed value in both range and amplitude data. This was mainly done to avoid unwanted spatial filter responses on the range data due to high levels of noise in regions with low confidence. The radial filter tuning function was set to $A(\rho) = (2\pi\rho)^2 \cdot \exp(-\pi\rho^2/4\sigma^2)$ with $\sigma = 0.3$ for all feature computations. We expect that filter optimization will further improve the results.

V. Results

The procedure was evaluated on a database of images taken of people at different head poses. A sample of such images is shown in Fig. 3, where both amplitude and range data are given for four subjects. Our database consists of a total of 13 subjects; for each subject, nine images were taken at roughly the same distance from the camera for different orientations of the head. The extent of the bounding box was estimated on a training set of three subjects, and the method was evaluated on the remaining ten subjects. The results presented in the following show that the method generalizes very well when using the combination of range and amplitude data.

Fig. 4 shows the ROC curves for different combinations of input data. For Fig. 4(a) only the range data was used from each image, whereas Fig. 4(b) shows the results for the amplitude data. The features ϵ_0 and ϵ_2 were used in both cases. The method achieves an equal error rate (EER) of 0.64 on the range data and 0.42 on the amplitude data. Even though the range data seems to be better suited for the proposed geometric features the amplitude data gives slightly better results. We cannot give a final explanation for this effect, but we assume that it is due to a higher level of noise in the range data. Although the EER is not satisfying in both cases, the results are quite good considering the simplicity of the classifier.

We were able to improve the performance dramatically by using a combination of features on range and amplitude data. We used the two features ϵ_0 and ϵ_2 for both types of data, which amounts to a total of four features. The corresponding ROC curve is shown in Fig. 4(c), and we can report an EER of 0.03 for this method. We believe that a very robust tracking of the nose can be achieved using this feature based approach.

VI. Conclusion

In this paper, we presented a very simple detector for the human nose based on the 3D TOF camera SR3000. The method yields very accurate and robust detection rates. However, we can point out three aspects that have potential to increase the performance the detector: (1) The use of a more sophisticated classifier, (2) an adaptation of the low-pass filter to the noise distribution, and (3) the use of additional features.

Also, we point out that when the detector is used to track the nose over several frames, as opposed to performing detection on individual frames, robustness can be improved by exploiting the fact that, in general, the position of the nose does not change much from frame to frame.

References

[1] D. Gorodnichy, "On importance of nose for face tracking," in *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition (FG'2002)*, Washington, D.C., May 2002.

[2] L. Yin and A. Basu, "Nose shape estimation and tracking for model-based coding," in *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, May 2001, pp. 1477–1480.

[3] C. Mota and E. Barth, "On the uniqueness of curvature features," in *Dynamische Perzeption*, ser. Proceedings in Artificial Intelligence, G. Baratoff and H. Neumann, Eds., vol. 9. Köln: Infix Verlag, 2000, pp. 175–178.

[4] E. Barth, T. Caelli, and C. Zetzsche, "Image encoding, labeling, and reconstruction from differential geometry," *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 6, pp. 428–46, November 1993.

[5] T. Oggier, B. Büttgen, F. Lustenberger, G. Becker, B. Rüegg, and A. Hodac, "SwissRanger™ SR3000 and first experiences based on miniaturized 3D-TOF cameras," 2006.

[6] C. Zetzsche and E. Barth, "Fundamental limits of linear filters in the visual processing of two-dimensional signals," *Vision Research*, vol. 30, pp. 1111–1117, 1990.

[7] ——, "Image surface predicates and the neural encoding of two-dimensional signal variation," in *Human Vision and Electronic Imaging: Models, Methods, and Applications*, B. E. Rogowitz, Ed., vol. SPIE 1249, 1990, pp. 160–177.

[8] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.