Touchless Detailed 3D Scan of Human Hand Anatomy Using Time-of-Flight Cameras

Jochen PENNE*, Martin PROFITTLICH, Thorsten RINGBECK, Bernd BUXBAUM
PMDTechnologies GmbH, Siegen, Germany;

Abstract

Time-of-Flight (ToF) cameras provide an easy, robust and touchless way to accomplish a three-dimensional surface scan at real-time framerates of ≥25Hz and with lateral resolutions of up to 204x204 pixels. This paper describes a processing pipeline to detect anatomical details the hand like finger tips, root finger joints and center of palm. The processing pipeline robustly analyzes hand in the field of view even if they are covered by gloves or sleeves. Additionally, the available depth data leads to a robust segmentation and validation of the anatomical structures.

Keywords: Time-of-Flight cameras, ToF cameras, hand detection

1. Introduction

Time-of-Flight (ToF) cameras provide an easy, robust and touchless way to accomplish a three-dimensional surface scan at real-time framerates of ≥25Hz and with significant lateral resolution. The ToF measurement principle estimates the distance of a point in the scene by measuring the phase delay of an actively emitted intensity-modulated infrared signal which is reflected in the scene [1]. This measurement principle can be implemented in standard CMOS technology making it commercially feasible to address the consumer market with this sensor technology.
The illumination units of ToF cameras emit intensity modulated light, which is usually in the near infrared range (NIR). A commonly used modulation frequency is *f* = 20MHz. Light emitting is triggered by an internal reference signal *s*. The detected optical signal *r* is received with an phase offset *τ*, since the emitted signal is reflected at the surface of objects at varying distances.
Each pixel determines the correlation between the detected optical signal and the reference signal:

$$c(\tau) = r \otimes s = \lim_{T\to\infty} \int_{T/2}^{T/2} r(t) \cdot s(t+\tau) dt \qquad (1)$$

If sinusoidal signals are assumed, that means

$$s(t) = \cos(\omega t) \qquad (2)$$

and

$$r(t) = b + a\cos(\omega t - \varphi)\,, \qquad (3)$$

where $\omega = 2\pi f$ is the modulation frequency, the above Eq. *(1)* yields

$$c(\tau) = \frac{a}{2}\cos(\omega\tau + \varphi). \qquad (4)$$

Eq. *(3)* intrinsically models some relevant aspects of the signal spread and acquisition: *b* is the amount of background light, while *a* models the received amplitude of the active illumination signal.
The correlation function *(4)* is sampled four times in each pixel: Four measurements $I_i = c(\tau_i)$ with $i = 0..3$ are acquired, where each measurement is done at a different internal phase shift $\tau_i = i \cdot \frac{\pi}{2\omega}$. Using these four samples of the correlation function, amplitude *a*, phase-shift φ and incident light intensity *b* can be computed by

$$a = \frac{1}{2}\sqrt{(I_3 - I_1)^2 + (I_2 - I_0)^2}\,, \qquad (5)$$

$$\varphi = \arctan 2(I_3 - I_1, I_2 - I_0) \qquad (6)$$

and

$$b = \frac{1}{4}\sum_0^3 I_i\,. \qquad (7)$$

* j.penne@pmdtec.com; +49 172 6699027; www.pmdtec.com

Finally, the distance *d* of a point observed by a pixel can be computed by

$$d = \frac{c}{4\pi \cdot f}\varphi, \tag{8}$$

where $c \approx 3 \cdot 10^8 \frac{m}{s}$ is the speed of light. Since the modulation is periodic it becomes obvious from equation (8) that valid distance measurements can be accomplished in an unambiguous distance range, which depends of the modulation frequency *f*. For *f*=20MHz this range is 7.5m, which is half the modulation wavelength $\lambda_{\mathrm{mod}} = 15m$. Fig. 1 depicts the measurement principle and Fig. 2 depicts available ToF cameras and example data.
The best signal-to-noise ratio (SNR) is achieved when each pixels only receives actively modulated light and no background light. Since the non-presence of background light can not be assumed, pixels in ToF cameras from PMDTechnologies GmbH provide an integrated SBI (suppression of background illumination) circuit: By removing background light components of the received light directly in the pixel the SNR is significantly improved. This feature is a key step towards valid distance measurements according to the ToF principle in uncontrolled illumination conditions.
Cameras from PMDTechnologies GmbH provide a lateral resolution of up to 204x204 pixels: This is the highest resolution currently available.

It is worth noting that ToF cameras in general have been intensively investigated regarding statistical and systematic error sources [2,3].
Statistical noise: As for every sensor the delivered data may be corrupted by measurement noise. In the context of ToF data temporal averaging and bilateral filtering have been reported to provide effective noise reduction with a controllable effect of feature preservation.
Systematic wiggling error: Since the perfect sinusoidal shape of the modulation is in praxis not met due to hardware and cost limitations the reconstruction of the correlation signal using four samples consequently leads to a systematic distance measurement error. Fortunately, the behavior of this error can be modeled well and appropriate calibration routines are at hand.
Flying pixel: At the border of objects so-called *flying pixels* can occur due to the rather large solid angle covered by one pixel of a ToF camera. These pixels receive a superimposed signal of light reflected from the background and the foreground. Since the distance computation scheme described above assumed one consistently reflected signal, the distance value computed at *flying pixels* can only be unvalid and has to neglected for future processing steps. A similar effect can be caused by secondary (or multiple) reflections.

In general, the acquisition of the data necessary for wiggling error compensation can be time-consuming and complex, since there is the need for ground truth distance data. Thus, in praxis it has to be judged regarding the application whether there is a valid trade-off between the required accuracy and the effort of an wiggling error compensation.

For many applications a range map of the observed scene is not sufficient: Instead, Cartesian coordinates of the observed scene are required. The computation of these coordinates involves knowledge of the projective properties of the receiving optic. A widely used camera model is the pinhole camera model with extensions to model radial and tangential lens distortions. Appropriate calibration routines are at hand [4]. Having determined the focal length $f_{foc,}$ and the coordinates of the principal point (c_x,c_y), the 3D coordinates $(x_{i,j},y_{i,j},z_{i,j})$ of a point observed at distance $d_{i,j}$ at pixel (i,j) in the sensor matrix can be computed by

$$x_{i,j} = \frac{(i - c_x)p_w}{d_{i,j}'} d_{i,j} \tag{9}$$

$$y_{i,j} = \frac{(j - c_y)p_h}{d_{i,j}'} d_{i,j} \tag{10}$$

$$z_{i,j} = \frac{f_{foc}}{d_{i,j}'} d_{i,j} \tag{11}$$

where

$$d_{i,j} = \sqrt{(i - c_x)^2 + (j - c_y)^2 + f_{foc}{}^2} \tag{12}$$

and p_w and p_h denote the physical width and height of a pixel.

If the calibration routine revealed significant lens distortions, an image undistortion should be applied to the depth map before computing the 3D coordinates, since it is necessary to use the undistorted distance value of each pixel for the computation.

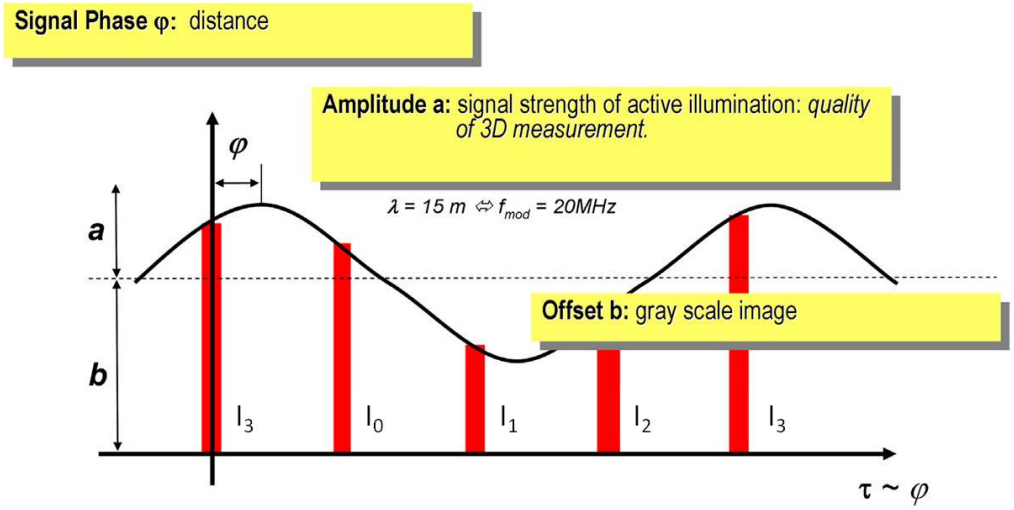

Figure 1: ToF measurement principle; Amplitude, phase delay and incident light intensity.

Figure 2: Same technology, two different cameras. PMDTechnologies' CamBoard (left) and CamCube (middle). Both cameras provide 204x204 px lateral resolution and an unambiguous measurement range of 7.5m. While the webcam-sized CamBoard was designed as a completely USB-powered solution for the consumer market, the CamCube was designed as a versatile ToF camera with exchangeable receiving optic and illumination unit for manifold applications. To the right a color coded depth 3D visualization and gray value amplitude data is displayed.

2. Proposed Method For Hand Detection And Analysis

The following notation is used: $a_{i,j}$, $d_{i,j}$, $x_{i,j}$, $y_{i,j}$ and $z_{i,j}$ denote the amplitude-, distance-,x-,y- respectively z-value at the pixel at position *(i,j)* of the sensor matrix (0≤i,j≤203). A,D,X,Y and Z denote the set of all amplitude, distance, x-,y- respectively z-values. Distance values are given in cm, amplitude values have no unit; x-, y- and z-values are given in cm.

2.1. Rough Segmentation

First, a rough segmentation is computed. The pixel positions $P' = \{(i,j) \mid a_{i,j} \leq a_{\min} \wedge d_{\min} \leq d_{i,j} \leq d_{\max}\}$ of pixels whose amplitude is greater than a_{min} and whose distance value is inside a certain distance range bounded by d_{min} and d_{max} are identified and the rough segmentation is given by the corresponding sets of amplitude and values as well as the corresponding x-, y- and z-coordinates: $A' = \{a_{i,j} \in A \mid (i,j) \in P'\}$, $D' = \{d_{i,j} \in D \mid (i,j) \in P'\}$, $X' = \{x_{i,j} \in X \mid (i,j) \in P'\}$, $Y' = \{y_{i,j} \in Y \mid (i,j) \in P'\}$ and $Z' = \{z_{i,j} \in Z \mid (i,j) \in P'\}$.

2.2. Fine Segmentation

The segmentation is refined by the following steps. Let d'_{min} denote the average of the five smallest distance values given in D'. The pixel positions $P'' = \{(i,j) \in P' \left| d'_{\min} \le d_{i,j} \le d'_{\min} + depth_{hand} \right.\}$, which correspond to pixels who are maximum $depth_{hand}$ cm behind the pixel closest to the camera, are identified: The assumption is, that the hand is the object closest to the camera and that d'_{min} reflects this distance; consequently, $depth_{hand}$ defines a depth region in which the hand is assumed to be. The fine segmentation is completed by the application of a 3x3 eroding mask to slightly shrink the segmentation: The reason is that, as mentioned above, *flying pixels* can occur at the border of objects. The resulting pixel positions are denoted P'''.

2.3. Palm Contour And Center Segmentation

A binary mask $B = \{b_{i,j} \left| 0 \le i,j \le 203, b_{i,j} = 0 \Leftrightarrow (i,j) \in P''', b_{i,j} = 1 \; otherwise \right.\}$ is computed and a distance transformation $T = \{t_{i,j} \in R^+ \left| 0 \le i,j \le 203 \right.\}$ is computed, where $t_{i,j}$ is the Euclidean distance of the pixel at position (i,j) to the nearest pixel position with $b_{i',j'}=0$. Spoken informally, in our case $t_{i,j}$ is the distance of a pixel to the object border, since B represents the result of the fine segmentation. The contour of the palm is than computed as pixel positions $C = \{(i,j) \in P''' \left| t_{i,j} = palm_{border} \right.\}$; that means pixel positions are identified, whose value in the distance transformation is equal to $palm_{border}$, consequently they have a distance of $palm_{border}$ to the object border.
The center of the palm is identified as the average pixel position of pixels having the maximum value in the distance transformation.

2.4. Finger Tip Identification: Analysis Of Convexity Defects

First, closed contours in P''' are computed. There may be more than one contour; the biggest contour with a length of more than 100 pixels is utilized furthermore. For this contour the convex hull and convexity defects are computed: Convexity defects with a defect depth of more than $convex_{depth}$ and an Euclidean pixel distance of less than $convex_{length}$ are considered as indicators for extremities of the hand (fingers, thumb). The starting and end pixel of each convexity defect are considered as n potential extremities of the hand. These pixel positions are denoted $P^{ex} = \{ p_k^{ex} = (i,j) \left| \right. , 0 \le k \le n-1 \}$. Naturally, the corresponding measured distances $D^{ex} = \{ d_k^{ex} = d_{p_k^{ex}} \left| p_k^{ex} \in P^{ex}, 0 \le k \le n-1 \right.\}$ are available. The values D^{ex} are analyzed starting with the point closest to the camera, which means the point with the smallest measured distance value is considered first. This point and all points are considered in increasing distance from the camera and a point is considered as a refined potential extremity if the distance difference between the current point and the last considered point is less than $depth_{diff}$. The remaining n' refined potential extremities are denoted $P^{ex'}$.

2.5. Root Finger Joint Identification And Validation

Assuming that the potential extremities $P^{ex'}$ are the tips of fingers the corresponding root finger joints $R^{ex'}$ are computed as the pixel positions in the palm contour C which are closest to the assumed finger tips. Formally: $R^{ex'} = \{ r_{k'}^{ex'} = (i',j') \in C \left\| r_{k'}^{ex'} - p_k^{ex'} \right\| \to \min \}$. Finally, the computed finger tips $P^{ex'}$ and root finger joints $R^{ex'}$ are validated by computing the 3D Euclidean length between finger tips and corresponding root finger joints: $l_k = \left\| (x_{i_k',j_k'}, y_{i_k',j_k'}, z_{i_k',j_k'}) - (x_{i_k,j_k}, y_{i_k,j_k}, z_{i_k,j_k}) \right\|$, where $(i_k', j_k') = r_k^{ex'} \in R^{ex'}$, $(i_k, j_k) = p_k^{ex'} \in P^{ex'}$ and $0 \le k \le n'-1$. For each value of l_k absolute thresholds $length_{min}$ and $length_{max}$ are applied to verify whether the identified finger tip and root finger joint are belonging to a structure with a reasonable dimension to be a finger. It may be noted that the 3D Euclidean length of a finger is invariant with respect to the orientation of the finger towards the camera, since 3D point coordinates are available from ToF cameras.

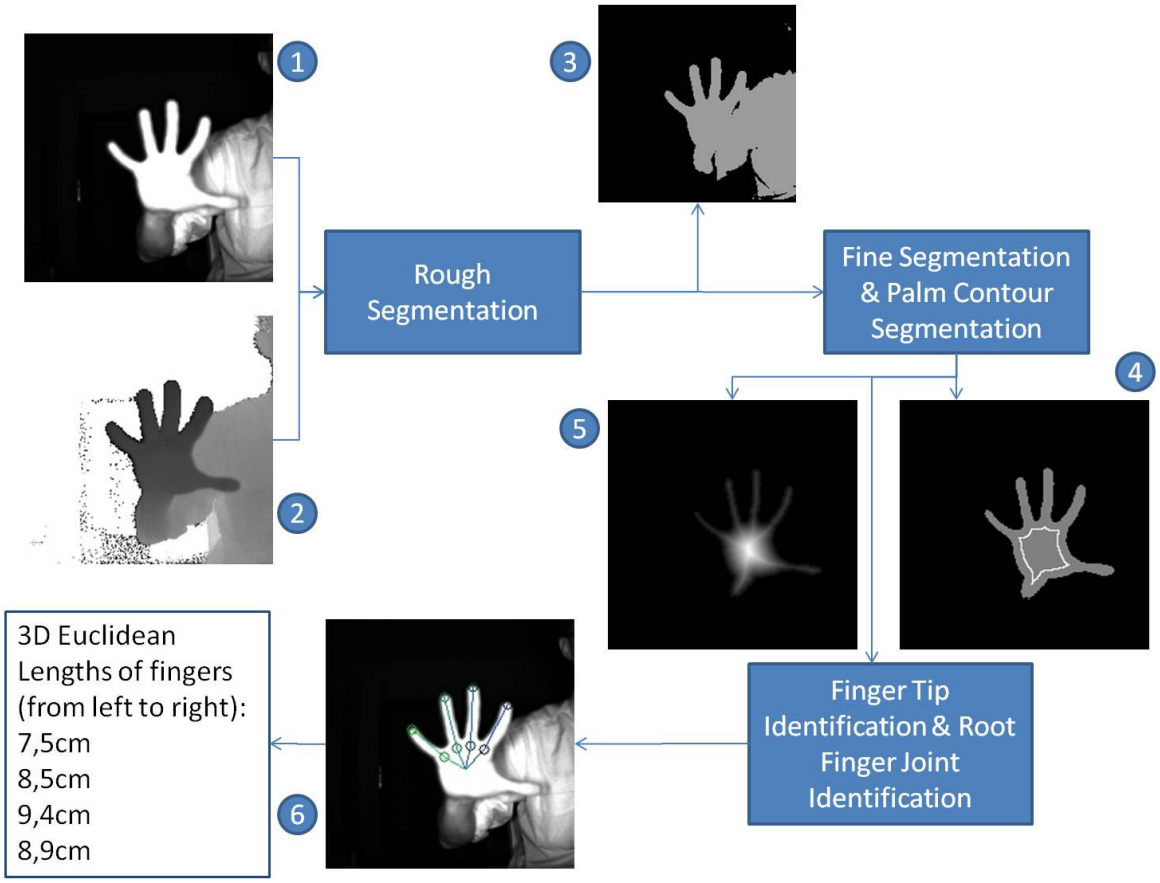

Figure 3: Overview of the processing pipeline. 1: Original amplitude data; 2: Original depth map; 3: Rough segmentation; 4: Fine segmentation (gray), palm contour (white) computed from distance transformation; 5: Distance transformation of the fine segmentation; 6: Identified finger tips, root finger joints and center of palm.

2.6. Comments On The processing Pipeline

Figure 3 gives an overview of the previously described pipeline. The utilized values for the various mentioned parameters and thresholds were:

$$a_{\min} = 3000,\ d_{\min} = 0cm,\ d_{\max} 100cm,\ depth_{hand} = 20cm,\ palm_{border} = 7,\ convex_{depth} = 5,$$

$$convex_{length} = 40,\ depth_{diff} = 5cm, length_{\min} = 5cm, length_{\max} = 12cm.$$

The parameterization was chosen such that a relatively large working area was available in front of the camera (determined by d_{min} and d_{max}). Additionally, $depth_{hand}$ enables the fine segmentation of the hand even if the hand is rotated rather strong towards the camera, that means the palm is not oriented parallel towards the image plane. The chosen values for the parameters $convex_{length}$, $convex_{length}$ and $depth_{diff}$ enforce that extremities respectively fingers to be detected are relatively close to each other and that the fingers are straight and not bend strongly (which would lead to a large distance difference between the finger tips).

Summarizing, the pipeline was a) designed to detect straight fingers and b) designed to not be easily tricked by obstacles.

3. Experiments

Various experiments were conducted to verify the feasibility of the previously described processing pipeline. The results are depicted in Fig. 4-6. Several capabilities of the proposed processing pipeline are emphasized: First, the texture invariance of the proposed approach: Since ToF cameras use active illumination and depth information is incorporated into the processing pipeline the detection of a hand and its fingers is possible for uncovered hand and hands covered by gloves or arms covered by sleeves. Second, obstacles like other hands do not disturb the hand and finger detection, since the depth information enables a valid fine segmentation of the hand closest to the camera (this hand is then subject to the further analysis). Third, the processing pipeline reliably detects straight fingers (since this was the intention) but also does reliably neglect bend fingers.

4. Conclusion And Outlook

A processing pipeline for the identification of finger tips, root finger joints and palm centers in ToF data was described. The results verify that a detailed analysis of hand anatomy using ToF cameras benefits from the additional depth information available. Regarding the low lateral resolution of ToF cameras compared to standard consumer 2D cameras, the proposed pipeline is a proof of concept for the fact that it is possible to take the step from identifying a hand as a whole object in ToF data towards a detailed analysis of anatomical structures of a human hand.

Especially the texture and illumination independence of the proposed pipeline opens up the field for various applications: Using a successful detection of four straight fingers as a trigger for virtually pressing a mouse button provides the starting point for a touchless human-machine interface based on opening and closing the hand in front of a ToF camera. Mouse cursor positioning may be accomplished by relating the mouse cursor movement to the movement of the detected center of the palm. Application scenarios targeting such a usage of ToF cameras have been described [5]. The technological foundation for such systems is at hand considering the ToF cameras from PMDTechnologies GmbH.

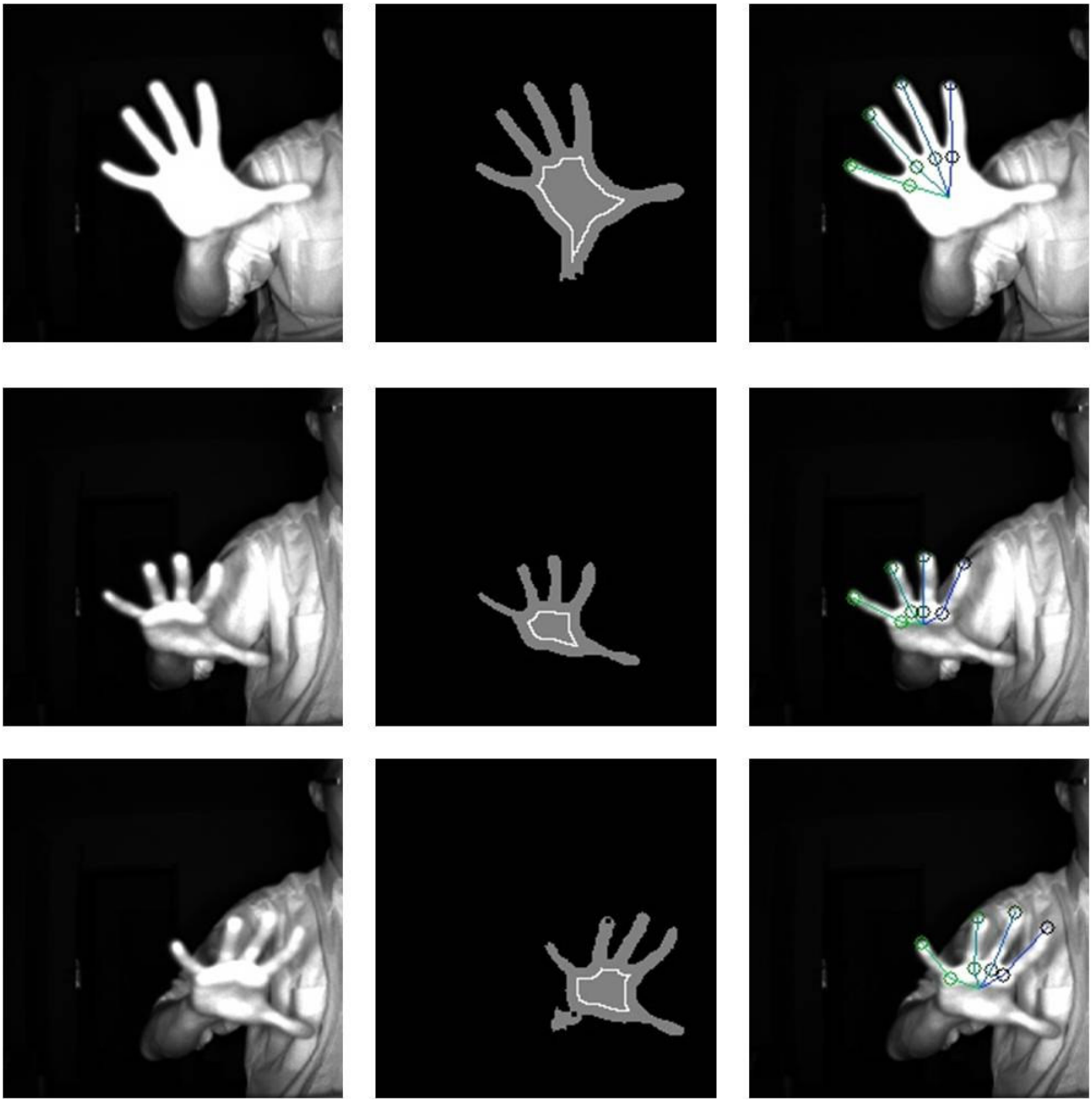

Figure 4: Experimental results. Hand detection of an uncovered hand with no sleeves. From left to right each row displays the original amplitude data, the result of the fine segmentation (gray: segmentation; white: palm contour) and the detected finger tips, root finger joints and center of palm. Note that in row two and three the hand is not oriented aligned with the camera plane and still the detection of fingers works.

Figure 5: Experimental results. Hand detection of a hand covered by a glove and with sleeves. From left to right each row displays the original amplitude data, the result of the fine segmentation (gray: segmentation; white: palm contour) and the detected finger tips, root finger joints and center of palm. Note that a second hand in the field of view does not disturb the hand analysis, since the processing pipeline was designed to analyze only the hand which is closest to the ToF camera.

Figure 6: Experimental results. The first row shows the detection capabilities for a hand covered by a glove and an arm covered by sleeves (left: original amplitude data; middle: fine segmentation (gray) and palm contour (white); right: identified fingers, root finger joints and palm center). The second row shows the sensitivity of the processing pipeline: Since the fingers are strongly bent, no fingers are detected as the processing pipeline was designed to detect straight fingers (left: original amplitude image; middle: depth map; right: detection result with no marked fingers).

References

1. T. Möller and H. Kraft and J. Frey and M. Albrecht and R. Lange, (2005): "Robust 3D Measurement with PMD Sensors", Proceedings of the 1st Range Imaging Research Day at ETH, pp
2. M. Lindner and A. Kolb (2007): "Calibration of the intensity-related distance error of the PMD TOF-camera", SPIE: Intelligent Robots and Computer Vision; pp. 6724-35
3. A. Kolb, E. Barth, R. Koch, R. Larsen (2009): „Time-of-Flight Sensors in Computer Graphics", Proceedings of Eurographics, pp. 119-134
4. Zhang, Z. (2000): "A Flexible New Technique For Camera Calibration". IEEE Transactions on Pattern Analysis and Machine Intelligence 22(11), pp. 1330–1334
5. J. Penne; S. Soutschek; M. Stürmer; C. Schaller; S. Placht; J. Kornhuber; J. Hornegger (2009): "Touchscreen without Touch - Touchless 3D Gesture Interaction for the Operation Room", i-com - Zeitschrift für interaktive und kooperative Medien 1 / 2009 (2009) No. 8 pp. 19-23