SURVEILLANCE AND ACTIVITY RECOGNITION WITH DEPTH INFORMATION

Frank Wallhoff, Martin Ruß, Gerhard Rigoll

Technische Universität München
Human-Machine Communication
Theresienstr. 90, 80333 Munich

Johann Göbel, Hermann Diehl

EADS Corporate Research Center Germany
Depearment LG-ME
81663 München

ABSTRACT

In the present treatise an image sensor acquiring additional depth information is applied to extend regular computer vision algorithms. The so called Photonic Mixer Device (PMD) basing on the time-of-flight principle can measure the distance between a smart pixel on the image sensor and the object being recorded. Since the resolution of such a sensor is rather low, they are not intended to replace existing image acquisition technologies, i.e. CMOS or CCD cameras, but can rather be employed to assist and speed up several tasks, especially image segmentation or object detection tasks.

Index Terms— Computer Vision, 3D Surveillance, Photonic Mixer Device, Segmentation

1. INTRODUCTION

Computer vision with its numerous number of application fields, such as surveillance, face recognition or object classification, has consolidated itself as a very important sector within the signal processing domain. Herein many elaborated and reliable algorithms have been introduced especially over the last decade. However, most of these algorithms are based on and restricted to the fundamental fact, that the real 3D environment is represented by 2D images. Although several hard- and software approaches exist to bridge this gap, it has turned out that most of them are either computationally too expensive, impractical due to their hardware constraints or both [1]. Furthermore most of them have lacking real-time capabilities, i.e. the acquisition of approx. 25 frames per second.

As a consequence, we have concentrated on a hybrid way to incorporate the good results from well-known 2D algorithms with a rather coarse but fast 3D representation of the observed scenery in a sophisticated way. To achieve this goal the outputs from an arbitrary image sensor and a spatially low resolution pixel range scanner are combined and interpreted. Since the herein employed Photonic Mixer Device (PMD) is capable to scan images with a frame rate of up to 25 Hertz by working in an autarkic manner, it can be foreseen and integrated even into applications with hard real-time constraints. Together with additional depth information an image segmentation task can become trivial, for example to isolate an image's foreground from its background.

To demonstrate the functionality of the presented approach, the rest of the paper is outlined as follows: after a brief theoretical introduction of the obeyed PMD technology, the setup for acquiring the desired image pairs is presented. Then a calibration procedure for overlaying images with depth information is introduced. After the presentation of three integrated applications the paper closes with some conclusions and an outlook.

2. PMD IMAGE ACQUISITION PRINCIPLE

The non-invasive optical image acquisition principle of the Photonic Mixer Device (PMD) is based on the run-time difference of a light impulse directly send to the detector and the reflected light from the surface of objects in the environment. In Figure 1 the simplified so-called time-of-flight measurement principle for one smart pixel is depicted.

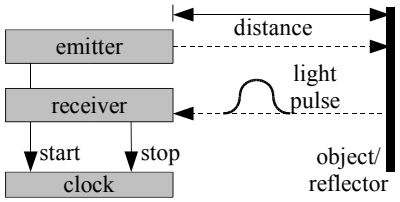

Fig. 1. Time-of-flight measurement principle.

With utmost precise counters, emitters and receivers the distance between the camera pixel and the object can be approximated by $d = \frac{t}{2} \cdot C$, where t represents the measured turnaround time between the start of a light impulse and its return to the receiver. The variable C represents the speed of light. The measurement of the flight-time is carried out using the phase shift of modulated infrared light pulses [2]. By combining several cells in a two dimensional structure a sensor with fully parallel operating smart pixels arises, allowing the 3D surface reconstruction of the scene. Since this measurement paradigm is directly implemented in the detector's hardware there is no additional computational effort, such as that arising from stereo cameras [3]. The refresh rate for one measurement loop allows between 5 and 50 frames/second.

However, to overcome the problem of background illumination which superposes the running pulse, various further techniques, such as optical filters and active circuits are implemented on the chip's cite. The sensor's usage of the suppression of background illumination makes it even possible to overcome the effects of bright ambient light [4] thus this measurement becomes independent from existing lighting conditions. The emitted infrared light has a wavelength of 870nm. By integrating the received light impulses over a certain interval the PMD camera could further serve as a NIR infrared camera.

3. EXPERIMENTAL SETUP

To enable image overlaying algorithms for the introduced combination of depth maps with high resolution images, two image sensors are equipped with the same Cosmicar/Pentax lens having a focal length of 16mm resulting in a field-of-view of $\approx 40^\circ$. For all further

experiments the same hardware components and test conditions are satisfied. The deployed image sensors are:

- A PMD[vision] model 3k-s 3D video range sensor [5] with 64×48 pixels running at 25 frames per second. The resolution in z-direction is specified by $z > 6$mm. The camera is connected to the PC by FireWire.
- A 3CCD Sony XC003P with a resolution of 752×582 (PAL) in full frame mode. Images from the camera are captured using a Cinergy USB 200 device.

By mounting the CCD image sensor piggyback on the PMD device as shown in Figure 2 a parallel offset of the camera axis arises. To overcome this problem instead of the follwing calibration also an semipermeable mirror together with a tilted mirror could be added into the optical path.

Aiming at having pixel respectively block-wise overlapping results the camera setup as well as intrinsic camera parameters have to be calibrated as introduced in the next chapter.

Fig. 2. Piggyback assembly of PMD and color CCD.

4. CAMERA CALIBRATION AND OVERLAY

As introduced, the fundamental idea bases on overlapping a higher resolution image with a coarse depth map. This will enable a variety of different applications such as face recognition, where a few pixels are enough to find a face first but are not enough to recognize its identity later.

As a result of the physical setup both images can be brought together by an coordinate translation followed by an expansion, see Figure 3. The variable $I(x, y)$ denotes a RGB color triple of the CCD camera in the x/y coordinate system, $d(u, v)$ is the distance matrix of the PMD in the displaced and scaled u/v system measured in meters. $\mathcal{T}$ symbolizes the transformation of the u/v system to the reference system x/y.

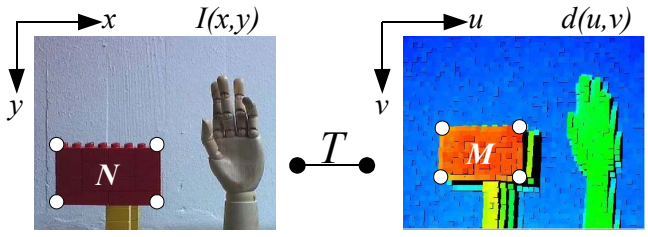

Fig. 3. Calibration body from both sensors' perspectives: CCD (left) and PMD (right), where distances are color coded.

The scaling parameters are denoted by s_x and s_y, the adjustment by a_x and a_y. Ideally the scaling is given by the fraction of both resolutions and a_x should be zero. However, due to the coarse resolution of the PMD and other external impacts all parameters have to be estimated automatically by finding characteristic landmarks in both images. In principle the presented calibration object may have an arbitrary shape and color as long as it can be distinguished from the scenery's background, here the nearest closed shape with distance $d_{\text{min}} = min(d(u, v))$. In our examples the calibration corpus consists of a LEGO model with a flat red facing and an depth of 15cm. Its surface map is denoted with M.

$$M(u,v) = \begin{cases} 1 & \text{if } d(u,v) \leq d_{\text{min}} + 15\text{cm} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

From M four calibration landmarks are derived by a bounding box surrounding this shape. In the high resolution image the corresponding four landmarks are found by a fitting box around the red object N, where its HSV triple falls into the following domain:

$$N(x,y) = \begin{cases} 1 & \text{if } ((H \geq 319)|(H \leq 31))\&(S \geq 5)\&(S \leq 16)\& \\ & (V \geq 0)\&(V \leq 16) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The scale factors then become:

$$s_x = \frac{max_x(N) - min_x(N)}{max_u(M) - min_u(M)} \text{ and } s_y = \frac{max_y(N) - min_y(N)}{max_v(M) - min_v(M)} \quad (3)$$

After expanding $d(u, v)$ by s_x and s_y, a equally scaled binary map $D(x, y)$ gained by bicubic interpolation arises. This has to be displaced by a_x and a_y so that the color and depth box around the calibration body overlap. Thus an area results where the color image and the depth map are defined simultaneously.

$$a_x = min_x(N) - min_x(D) \text{ and } a_y = min_y(N) - min_y(D) \quad (4)$$

5. APPLICATIONS AND USE CASES

5.1. Person Counting

In a first approach it is demonstrated how the personnel flow through a door can be measured using a segmentation gained by depth information. Through the derivation of shape information it becomes possible to segment a head and distinguish a person passing the scene from other items. Due to the PMD's measurement principle there are no restrictions regarding external lighting conditions.

Human heads are assumed to bear some resemblance to egg shaped structures when seen from above. A typical image recorded with the PMD above the door is depicted in Figure 4.

By cutting the observed depth map (Fig. 4, upper left) into slices with a height of $50cm$ a head inside this subspace will appear as a connected 3D object (Fig. 4, upper right).

After a connected components analysis and a thresholding operation this can be transferred to a flat area (Fig. 4, lower left). Thus the outer curve or edge function of a head will appear as an ellipsis, which can be detected by a classical correlation with predefined head patterns or invariant moments [6]. To enable a robust detection of heads from persons with different body heights, this sampling procedure starts at the top of the door ($\approx 2m$) and ends at the height of a children ($\approx 0.8m$). In order to seek for heads in all heights, several overlapping slices with a step width of $25cm$, the typical skull

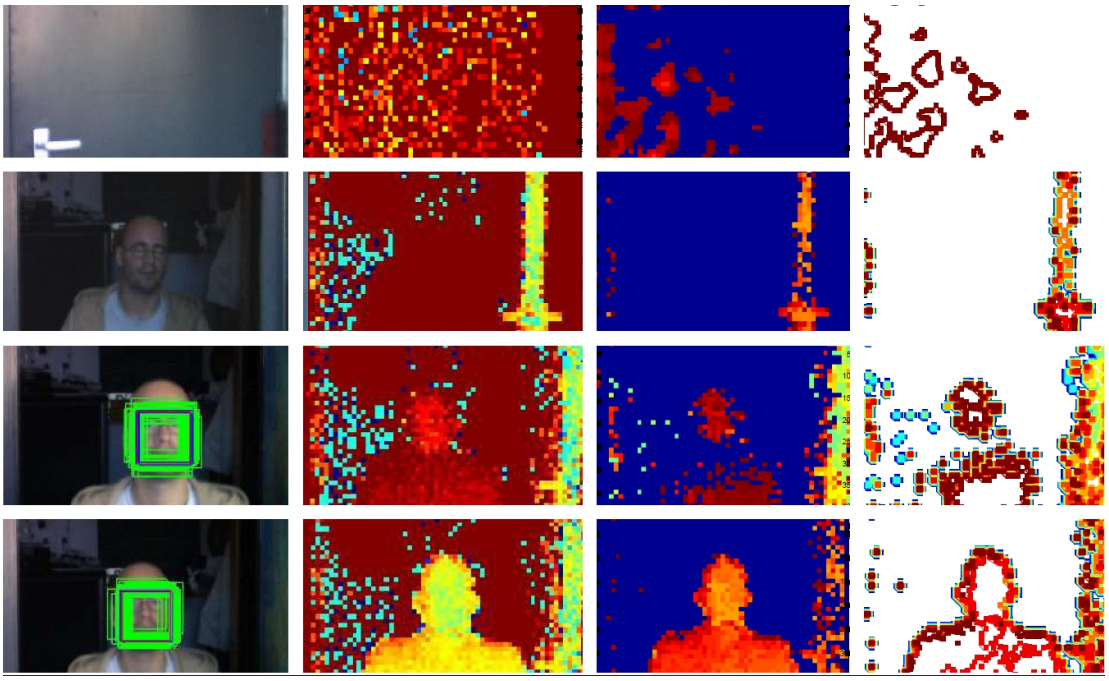

Fig. 5. Face detection examples: CCD images with detected face (left), color coded raw range map (2nd column), relevant range interval (3rd) and edge map (right).

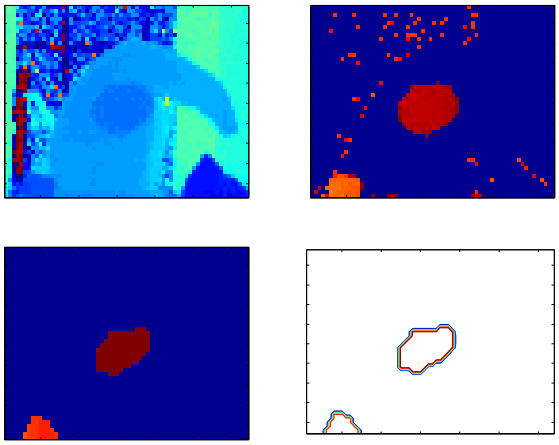

Fig. 4. Original color coded top-down depth map (top left), isolated depth information of one slice (top right), preprocessed shape matrix (bottom left) and final edge map (bottom right).

height, are processed. Several overlapping head detection results are clustered by their mean values.

A person can finally be counted if its trajectory can be tracked within a pre-defined region-of-interest. The deployed tracking algorithm is introduced in the next section.

5.2. Face Detection

Face detection has a long research tradition and various 2D algorithms have been introduced [7]. However, most of them either use example based search strategies or skin color. Although well studied in many treatises, skin color recognition tends to be unstable for scenarios with varying lighting conditions, such as a regular room with windows [8]. Furthermore the discrimination of real skin and skin alike background pixels increases this difficulty, for example hand before a salmon colored shirt. Example based algorithms on the other hand usually have to seek for faces at all possible positions and scales, which means high computational efforts.

In order to track faces in a robust manner, the above mentioned detection paradigms can be merged using a so-called Condensation algorithm [9] or particle filter. After an initialization phase, i.e. the exhaustive search for a face, the new location is randomly sampled around the old face with N particles and some heuristics until the face disappears. The i-th particle f represents a rectangular area at time k and consists of the following values: the upper left edge coordinate x and y, the box size s, a face likelihood l, the particle's weight π together with their predecessors at $k-1$: $p_k^i = [x_k, y_k, s_k, l_k, \pi_k, x_{k-1}, y_{k-1}, s_{k-1}]$.

The face likelihood of one particle is estimated by a neural network and the relative amount of skin pixels [10]. Now, this existing algorithm is extended with additional depth information aiming at improving accuracy and speed. By seeking for a face-like area in a defined interval in the range map, the novel algorithm is initialized when a face appears in it. Furthermore, the formerly to a face posterior and skin color restricted face-likelihood is expanded now by a 3D head similarity, which is computed analogous to the above introduced elliptical edge correlation (but now from the front).

Typical tracking results from a test sequence with a person entering the room are demonstrated in Figure 5. The tracking is initialized as soon as an ellipsis in the door plane is detected. By the use of the depth information the false triggering rate has vanished to zero.

5.3. Gesture Recognition

The last use case deals with the improvement of an action and gesture recognition system constituted on difference image based global motion features [11]. A seven dimensional vector consisting of the mass-center, the deviation and its change (each in x- and y direction) as well as the intensity of the motion is extracted for every time step. An unknown sequence is dynamically classified by Hidden Markov Models. However, self occlusions, changing lighting conditions and compression artefacts cause a high noise level within the feature stream, which could not be significantly removed even with a Kalman filter [12].

Again, the image segmentation gained by depth information can easily subsist this approach. Therefore the region of interest is masked by the object in the foreground. Due to the calibration, this mask can be overlapped pixel-wise with the CCD image, as shown in Figure 6. The advantage becomes obvious, since there is no interference with the objects on the whiteboard even if they are close to skin color.

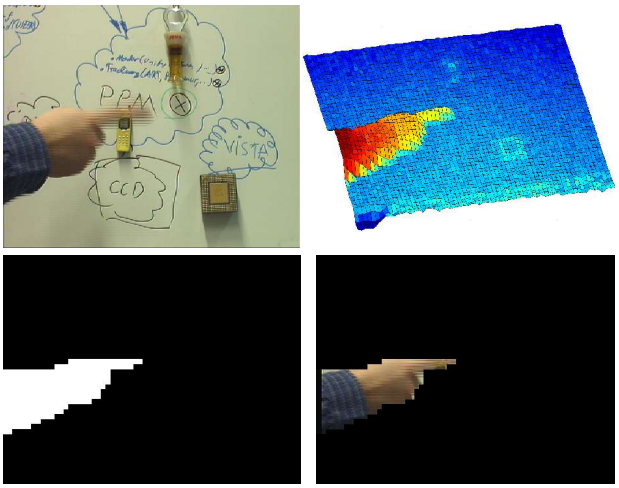

Fig. 6. Pointing gesture: CCD image (upper left), depth map (upper right), foreground mask (lower left) and overlapped image (lower right).

6. SUMMARY AND CONCLUSIONS

A reliable image segmentation approach has been introduced and integrated with very low computational efforts by making use of depth information from a PMD range sensor.

It has been shown, that existing algorithms can be extended efficiently, which enables possibilities to new image processing algorithms. However, there are still some drawbacks that the employed measurement technique is suffering from. Range measurement problems occur in conjunction with highly reflective surfaces that are too close to the sensor. By the mirroring effect of the infrared diodes on the material's face the pixel distances become too large. Other, less problematic surfaces are light adsorbent materials. On these facings the depth values are very noisy.

Three uses cases have been presented, which show very promising qualitative results. Therefore it is planned to measure the quantitative improvements that can be gained by integrating additional depth information into regular approaches in the future. However, to do so commonly available databases would have to be acquired first. Furthermore it is planned to concentrate on 3D surface reconstruction tasks for augmented reality applications as well as environmental exploration for autonomous navigation systems.

7. ACKNOWLEDGEMENT

The authors would like to express their thanks to Bianca Hagebeuker from PMDTechnologies [13] for her support and the short term provisioning of a sensor device.

8. REFERENCES

[1] Frank Forster, *Real-Time Range Imaging for Human-Machine Interfaces*, Ph.D. thesis, Technische Universität München, Lehrstuhl für Mensch-Maschine-Kommunikation, 2004.

[2] T. Kahlmann, F. Remondino, and H. Ingensand, "Calibration for increased accuracy of the range imaging camera swissranger," in *Proceedings of the ISPRS Commission V Symposium Image Engineering and Vision Metrology*, D. Schneider Editors: H.-G. Maas, Ed., Dresden, Germany, 25-27 September 2006, vol. XXXVI, pp. 136–141.

[3] Z. Xu, R. Schwarte, H. Heinol, B. Buxbaum, and T. Ringbeck, "Smart pixel - photonic mixer device (pmd)," *Proc. M2VIP '98 - International Conference on Mechatronics and Machine Vision in Practice, Nanjing*, pp. 259–264, 1998.

[4] T. Möller, H. Kraft, J. Frey, M. Albrecht, and R. Lange, "Robust 3d measurement with pmd sensors," in *In: Proceedings of the 1st Range Imaging Research Day at ETH Zurich, Zurich, Switzerland, pp. "upplement to the Proceedings"*, 2005.

[5] PMDTechnologies, "Data Sheet PMD(vision) 3k-s," Online document *http://www.pmdtec.com/inhalt/download/documents/PMDvision 3k-S_000.pdf*.

[6] A. Chalechale, F. Safaei, F. Naghdy, and P. Premaratne, "Hand posture analysis for visual-based human-machine interface," in *WDIC 2005 APRS Workshop on Digital Image Computing*, In B. Lovell & A. Meader (Eds.), Ed. Queensland: The Australian Pattern Recognition Society, 2005, pp. (pp. CD Rom 91–96).

[7] M-H. Yang, D.J Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE transactions PAMI*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[8] Moritz Störring, *Computer Vision And Human Skin Colour*, Ph.D. thesis, Faculty of Engineering and Science, Aalborg University, 2004.

[9] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision (IJCV)*, vol. 29, no. 1, pp. 5–28, 1998.

[10] F. Wallhoff, M. Zobl, G. Rigoll, and I. Potucek, "Face tracking in meeting room scenarios using omnidirectional views," *Proceedings Intern. Conference on Pattern Recognition (ICPR)*, Aug. 2004.

[11] F. Wallhoff, M. Zobl, and G. Rigoll, "Action segmentation and recognition in meeting room scenarios," *Proc., IEEE Int. Conf. on Image Processing (ICIP)*, Oct. 2004.

[12] M. Zobl, A. Laika, F. Wallhoff, and G. Rigoll, "Recognition of partly occluded person actions in meeting scenarios," *Proc., IEEE Int. Conf. on Image Processing (ICIP)*, Oct. 2004.

[13] PMDTechnologies, "Homepage," *http://www.pmdtec.com/-e_index.htm*, 2007.