Manuskript

From E. Kollorz et. al
Human Machine Interface for Elderly People, pp. 383-386
1. Deutscher Kongress Ambient Assisted Living

Manuskript

Human Machine Interface for Elderly People

Dipl.-Inf., Eva Kollorz, Institute of Pattern Recognition, University of Erlangen-Nuremberg, Germany
Dipl. Med.-Inf., Jochen Penne, Institute of Pattern Recognition, University of Erlangen-Nuremberg, Germany
Prof. Dr.-Ing., Joachim Hornegger, Institute of Pattern Recognition, University of Erlangen-Nuremberg, Germany
Prof. Dr. med., Johannes Kornhuber, University Hospital Erlangen, Germany

Abstract

This paper presents a new approach for gesture recognition using a 3-D camera. The approach is based on projections onto the image axes (x-/y-axis) and additional depth features which are calculated within the region of interest (ROI). The 3-D camera provides a gray scale image and distance values to each object of the observed scene. These distance values allow for fairly simple background elimination. A nearest neighbour classifier is used to categorize the acquired image data in real-time. The method was tested with a ‘Leave-One-Out’ evaluation including 12 gesture images of 34 persons. The achieved recognition rate is 94.61%.

1 Introduction

As a statistical expansion, there will be approximately 9.1 million people in the year 2050 in Germany who will be 80 years old or elder [1]. This will be almost three times as high as it is today.
For the purpose of enabling the seniors to live autonomously for a longer period of time, convenient human machine interfaces (HMI) are necessary. E.g., a senior can control a device by a simple input like speech or gestures, instead of getting up and moving to the device. Similarly, these inputs are easier than pressing a little button on a remote-control for the specific device.
Several divisions on HMIs exist: tactile, touch, gesture and speech interfaces. Each of them has certain advantages as well as disadvantages. In this paper, we focus on a gesture interface. The advantages of our gesture controlled interface are: it works contactless, it is user-independent and the natural handling if the gestures are chosen well for the respective function.
The utilization of a 3-D camera bares simplifications, e.g., the recorded distance data can be used to easily extract the hand and arm from the background. Additional, a gray scale image is recorded by the camera.

This paper describes a 3-D camera based real-time gesture recognition system. The article is organized as follows. Section 2 lists state-of-the-art work for gesture recognition in general and for TOF technology regarding gesture recognition or position tracking. The TOF principle is also explained in section 2. Then, the algorithm is illustrated in section 3 and evaluated in section 4. The last section addresses the conclusion and outlook.

2 State-of-the-art

2.1 Gesture Recognition

Triesch and von der Malsburg [8] utilize an elastic graph matching technique to classify hand gestures. Six hand gestures are used to control the robot. Nefian et al. [9] introduce an approach for recognizing dynamic and static gestures with a 3-D statistical model. Four static gestures are presented in their paper, but evaluated together with the dynamic ones. Zernike moments and pseudo-Zernike moments for static gesture recognition are used by Chang et al. [10].
Walchshäusl [11] gives a good introduction of possible gesture recognition approaches, e.g., template matching with sum of absolute gray scale differences or with gray scale correlation. He also mentions several types of moments as features: area, center of gravity, central moments and hu-moment invariants. Other possible features for classification could be circumference, aspect ratio or Fourier descriptors.
Wysoski et al. [12] introduce feature vector model based on boundary histograms.
The partitioning between hand and arm also plays an important role: calculating the crossing between the size of the arm until the bigger palm appears or using ellipse- or tangent-fitting are common approaches to accomplish this task.

2.2 Time-of-Flight (TOF) Technology

Comparable to the laser radar technique, the depth information of a pixel is measured by the phase shift of transmitted and reflected infra-red light. Arrays of light emitting diodes (LEDs), mounted at the camera, send out modulated infra-red light. This light is reflected by objects of the scene and detected by the sensor. The phase shift $\Delta\mathbf{t}$ is calculated between the emitted signal at time $\mathbf{t_s}$ and the received signal at time $\mathbf{t_r}$ and finally the measured distance is computed by:

$$d = \frac{c\Delta t}{2},$$

where **c** is the speed of light [4].

Liu and Fujimura [5] describe an approach for recognizing dynamic gestures. Their classification is based on hand detection, shape and trajectory analysis. They only utilize the depth information of the camera for hand and head detection. Breuer et al. [7] also use a time-of-flight camera to recognize hand movements. They transform the measured data into 3-D point clouds and use principle component analysis (PCA) as a first estimate. Unfortunately, their approach delivers just two till three frames per second.
Du et al. [6] use a time-of-flight camera for tracking the finger position: the depth values are used for the detected central columns.

3 Classification of Gestures

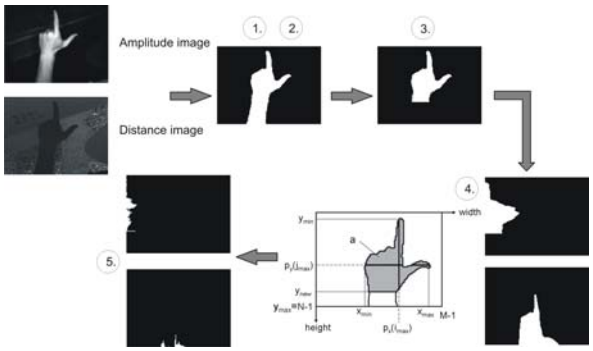
In this section, we describe our approach for recognizing different gestures with a TOF camera (**Fig.** 1). The core of the method is to use the projections of the hand onto the image axes (x-/y-axis). These features are simple to compute and allow real-time classification of the gestures. The region of the arm is discarded since this area does not contain any useful information concerning the gestures. To distinguish between gestures with same projections

but different alignments, e.g., gesture G and L (**Fig.** 4), the provided depth information of the camera is included in the classification process.
The detection of features in 2-D is computationally cheaper than using the reconstructed data in 3-D.

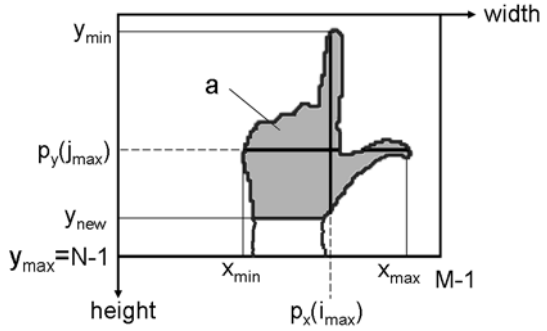
Figure 1 Overview of the presented algorithm; the camera provides an amplitude and a distance image; first, extraction of the hand and the arm (1.), computation of the bounding box (2.), removal of the arm (3.), projection onto the image axes (4.), calculation of additional depth features (4.) and finally, classification (5.).

The algorithm can be divided into five steps:

1. **Segmentation** of the hand and arm:



The hand and the arm are segmented by a simple iterative seed fill algorithm via depth values. The depth range has to be adjusted to the situation/scene: background can be easily excluded. The hand should be extracted because the arm region will negatively influence the depth values and will have no effect on the gestures.



2. **Computation** of the bounding box:
The segmented region is projected onto the image axes (x-/y-axis) to determine the bounding box of the object [3]. Projections onto the x-axis will be referred to as $\mathbf{p_x}$, onto the y-axis as $\mathbf{p_y}$ (**Fig.** 2). The resolution is denoted by M x N. The four borders of the bounding box ($\mathbf{x_{min}}$, $\mathbf{x_{max}}$, $\mathbf{y_{min}}$, $\mathbf{y_{max}}$) are determined by a threshold $\boldsymbol{\theta}$. Using the indices, we can determine at which side the arm enters the field of view (FOV). For the chosen scenario, the arm enters the FOV from the bottom side.

Figure 2 Labeling of the used variables. Dyed area a: segmented hand without the boundary pixels (contour); $\mathbf{p_x}$: projection onto the x-axis (width); $\mathbf{p_y}$: onto the y-axis (height); $\mathbf{i_{max}}$: index of the maximal projection value of $\mathbf{p_x}$; $\mathbf{j_{max}}$: of $\mathbf{p_y}$; $\mathbf{x_{min}}$ and $\mathbf{x_{max}}$: minimal and maximal border of $\mathbf{p_x}$; $\mathbf{y_{min}}$ and $\mathbf{y_{max}}$ respectively.

3. **Extraction** of the hand:
The "average" length of a human hand is denominated by **l** [mm]. The index $\mathbf{y_{ini}}$ specifies the initial cut-off position for the chosen scenario. It is computed by projecting an "average" hand of length **l** [mm] with distance **d** [mm] from the 3-D camera into the image plane. With the focal length **f** [mm] and pixel dimension **s** [mm], we get:

$$y_{ini} = y_{\min} + \frac{f \cdot l}{d \cdot s}.$$

For the experiments, we made two assumptions:

- To cut off the hand at the right position, the arm is not covered at the wrist by a pullover or a watch.
- The person can wear a ring which should not affect the classification.

The first point is important for the shape of the gesture. If the gesture is a stretched hand, $\mathbf{y_{new}}$ is nearly $\mathbf{y_{ini}}$. Otherwise $\mathbf{y_{ini}}$ has to be adjusted, e.g., if the hand forms a fist: if the height projections $\mathbf{p_y}$ fulfill the criterion $p_y(i) - p_y(i+\alpha) < \beta$ (i ranges from $\mathbf{y_{ini}}$ to $\mathbf{y_{min}}$), $\mathbf{y_{ini}}$ is decreased by one. The resulting $\mathbf{y_{ini}}$ is $\mathbf{y_{new}}$, and $\mathbf{y_{max}}$ respectively. The criterion should express the small bottleneck at the wrist: $\boldsymbol{\alpha}$ adjusts the interval of the two projection bins which are compared and to determine the difference between the arm and the hand by projections (scanlines), the parameter $\boldsymbol{\beta}$ expresses the minimal barrier.

4. **Projection** of the hand onto the image axes:
The truncated hand is projected onto the image axes, subsequently the projections are enhanced with a median filter to eliminate outliers. The pre-processed projections are still denoted by $\mathbf{p_x}$ and $\mathbf{p_y}$. At that time the additional depth information is integrated into the classification process. For the truncated hand, three components are calculated: $\mathbf{a_{min}}$, $\mathbf{a_{avg}}$ and $\mathbf{a_{max}}$. $\mathbf{a_{min}}$ describes the minimum of the inner reliable

pixels of the segmented hand region a (respectively the average $\mathbf{a_{avg}}$ and the maximum $\mathbf{a_{max}}$).

5. **Classification**:
The classification is realized with majority decision over **m** frames and **k** nearest neighbours over all training gesture samples. For each gesture in the training set, the smoothed projections $\mathbf{p_x}$ and $\mathbf{p_y}$ are saved as well as the additional depth information $\mathbf{a_{min}}$, $\mathbf{a_{avg}}$ and $\mathbf{a_{max}}$. The current gesture is normalized to the reference gestures' domain (**R**). The distance $\mathbf{d_{class}}$ for the current gesture to one of the reference gestures is computed by:

$$d_{class} = \left| \frac{a_{avg} - a_{\min}}{a_{\max}} - \frac{a_{R,avg} - a_{R,\min}}{a_{R,\max}} \right| + \frac{x_{diff}}{x_{\max} - x_{\min}} + \frac{y_{diff}}{y_{\max} - y_{\min}},$$

where $\mathbf{x_{diff}}$ is the difference for the width over all M bins, $\mathbf{y_{diff}}$ respectively. The majority decision is first done for **k** nearest neighbours for each frame. If a ties occurs, the minimal distance $\mathbf{d_{class}}$ decides. Subsequently, the decision is carried out over **m** frames in the same manner.

Figure 3 Setup of the camera.

4 Setup and Evaluation

In the experimental setup, a PMD[vision] 19k camera of PMDTechnologies GmbH was used [2]. The resolution of this camera is 160 x 120 pixels. The pixel dimension **s** is 40 micrometers in each dimension and the focal length **f** is 12 mm. The PMD[vision] 19k camera provides up to 15 frames per second and has viewing angle of 40 degrees. For the experiments the depth range for the seed fill algorithm was between 70 and 110 cm. As "average" hand length **l**=210 mm was used. The parameter **α** was set to three, **β** to two. The threshold **θ** was set to 4 pixels, assuming that projection values smaller than the threshold are unreliable border pixels, e.g., at the tip of the thumb. The implemented median filter has kernel size three. **Fig.** 3 shows the setup of the experimental environment.

Figure 4 The 12 static gestures which have to be recognized by the system (Labeling: row-wise gesture A till gesture K).

The algorithm was tested with 12 different static gestures (**Fig.** 4). The data set contains 12 gestures of 34 persons, mixed male and female. Eight persons wear a ring during the acquisition. As evaluation criterion the 'Leave-One-Out' method was applied: one person is excluded from the data set and tested against all other reference gestures. The overall classification rate is computed by:

$$\frac{\sum_{g=1}^{\#gestures} a_{gg}}{\#persons \cdot \#gestures},$$

here $\mathbf{a_{gg}}$ is the number of correctly classified gestures in each class. The results of the classification rate can be seen in **Table** 1: the classification rate is 94.61%.

5 Conclusion and Outlook

This article introduces a new approach for static gesture recognition based on a PMD[vision] 19k camera.
The gesture recognition system helps elderly people to control easily everyday life devices. Here, 12 different gestures were selected which can be connected, e.g., to turn on/off the television or illumination. The gestures can be performed suitable. The features for the classification are based on projections and additional depth information. 34 persons executed the 12 different gestures and the 'Leave-One-Out' method for evaluation was chosen. 94.61% is the recognition rate for this gesture set. The

big advantages of the proposed method are: the real-time applicability and the user-independence.
An improvement for the proposed approach could be, e.g., a rotation-invariant extension. Thereby, the recognition rate would increase. Skewed performed gestures would be aligned and normalized.

	A	B	C	D	E	F	G	H	I	J	K	L
A	31	0	0	0	0	0	0	0	0	3	0	0
B	0	32	0	0	0	0	0	0	0	0	2	0
C	0	0	33	0	0	0	1	0	0	0	0	0
D	0	0	0	34	0	0	0	0	0	0	0	0
E	0	0	0	0	33	0	1	0	0	0	0	0
F	0	0	0	0	0	34	0	0	0	0	0	0
G	1	0	0	0	0	0	32	0	0	0	0	1
H	0	0	0	0	0	0	0	33	0	0	1	0
I	0	0	0	0	0	0	0	0	33	0	1	0
J	4	0	0	0	0	0	0	0	0	30	0	0
K	0	0	0	0	0	0	0	4	0	0	30	0
L	0	0	0	0	0	1	1	0	0	1	0	31

Table 1 The top row denotes the recognized gesture class; first column denotes the actual performed gesture; the main diagonal shows the correct detected gestures.

6 Literature

[1] Germany's Population by 2050: Results of the 11th coordinated population projection, Federal Statistical Office, Wiesbaden, Nov. 2006

[2] PMDTechnologies GmbH, Nov. 2007, http://www.pmdtec.com/

[3] Hornegger, J., Niemann, H.: Probabilistic Modeling and Recognition of 3-D Objects, International Journal of Computer Vision, Vol. 39, No. 3, pp. 229-251, Sept. 2002

[4] Sünkel, T.: Erkennung isolierter komplexer Handgesten in 2 1/2 D Videosequenzen mit Hidden Markov Modellen, Master thesis, Institute of Pattern Recognition, University of Erlangen-Nuremberg, Germany, June 2006

[5] Liu, X., Fujimura, K.: Hand Gesture Recognition using Depth Data, IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, pp. 529-534, May 2004

[6] Du, H., Oggier, T., Lustenberger, F., Charbon, E.: A Virtual Keyboard Based on True-3D Optical Ranging, Proceedings of the British Machine Vision Conference, Oxford, U.K., Vol. 1, pp. 220-229, 2005

[7] Breuer, P., Eckes, C., Müller, S.: Hand Gesture Recognition with a novel IR Time-of-Flight Range Camera – A pilot study, Lecture Notes in Computer Science, 3rd international Conference MIRAGE 2007, Rocquencourt, France, pp. 247-260, March 2007

[8] Triesch, J., von der Malsburg, C.: Robotic Gesture Recognition, Lecture Notes in Computer Science, Gesture Workshop, Bielefeld, Germany, pp. 233-244, Sept. 1997

[9] Nefian, A.V., Grzeszczuk, R., Eruhimov, V.: A statistical upper body model for 3D static and dynamic gesture recognition from stereo sequences, International Conference on Image Processing, Vol. 3, pp. 286-289, 2001

[10] Chang, C.-C., Chen J.-J., Tai W.-K., Han C.-C.: New approach for static gesture recognition, Journal of Information Science and Engineering, Vol. 22, No. 5, pp. 1047-1057, Sept. 2006

[11] Walchshäusl, L.: Klassifizierung dynamischer Gesten im Kontext multimodaler Infotainmentsysteme, Master thesis, Chair for Computer Aided Medical Procedures and Augmented Reality, Technical University Munich, Germany, Dec. 2004

[12] Wysoski, S.G., Lamar, M.V., Kuroyanagi, S., Iwata, A.: A rotation invariant approach on static-gesture recognition using boundary histograms and neural networks, International Conference on Neural Information Processing, Vancouver, Canada, Vol. 4, pp. 2137-2141, Nov. 2002