

Gesture Recognition using a Range Camera

M.B. Holte and T.B. Moeslund

Laboratory of Computer Vision and Media Technology
Aalborg University, Denmark
Email: tbm@cvmt.dk

Abstract. This paper introduces use of range information acquired by a CSEM SwissRanger SR-2 camera for one and two arms gesture recognition. The range data enables motion detection and 3D representation of gestures. Motion is detected by double difference range images and filtered by a hysteresis bandpass filter. Gestures are represented by shape contexts in the form of spherical histograms. This representation allows a simple matching by binary comparison of the histogram bins. Although this approach is still on a early state, results indicate successful recognition of a set of commonly used gestures.

1 Introduction

Most gesture recognition approaches use intensity images to extract necessary motion information. However, recently 3D motion information has been introduced to overcome the inherent problem of recognizing 3D gestures in 2D images, e.g., [4] obtain 3D data to produce motion history volumes by use of multiple calibrated, and background-subtracted cameras. Our approach aims at recognition of one and two arm gestures by use of a single range camera (CSEM SwissRanger SR-2). This type of camera has prior been used for obstacle detection [1] and environment reconstruction [2].

1.1 System overview

An overview of the range based gesture recognition system can be seen in figure 1 and is described in the following sections.

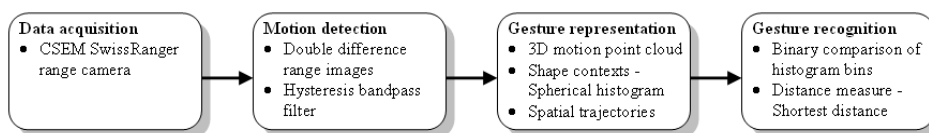


Fig. 1. An overview of the range based gesture recognition system.

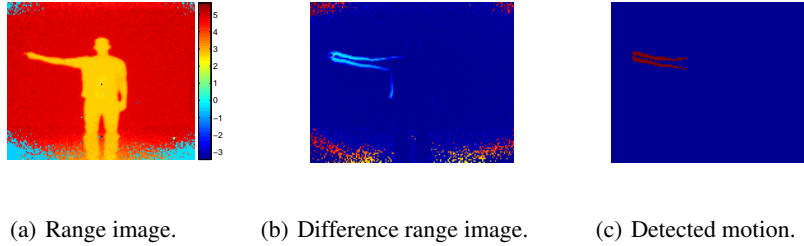


Fig. 2. (a) shows a range image, where the pixel values correspond to a distance. (b) shows the difference range image used for motion detection. (c) is the resulting motion detected in 2D after hysteresis bandpass filtering and creation of a double difference image.

2 CSEM SwissRanger SR-2 Range Camera

The CSEM SwissRanger SR-2 range camera [3] is based on the Time-Of-Flight (TOF) principle. The camera emits radio-frequency modulated light in the near-infrared spectrum, which is backscattered by the scene and detected by a CMOS CCD. The camera can deliver range and intensity images of 160×124 pixels with an active range of 7.5 m. We have achieved a frame rate of around 13 frames per second. The depth accuracy is typically in the order of a few centimeters, depending of the distance range and illumination. Figure 2(a) shows a range image of a “point right” action.

3 Motion Detection

Motion is detected by use of image subtraction. Specifically, double difference images are used. These images are created in 2D image coordinates with the range/depth represented as pixel values. However, to extract the relevant motion data from the arm movements, the data needs to be filtered properly. The range data captured by the SwissRanger camera includes a high degree of noise, resulting in jitter and errors causing pixels with extremely high range values. To handle the noise effects, each of the two difference images (figure 2(b)) are filtered with a hysteresis bandpass filter before they are ANDed together to create a double difference image.

3.1 Hysteresis Bandpass Filter

The hysteresis bandpass filter operates in 2D and use four threshold values T_1, T_2, T_3 and T_4 . The values that falls within the motion range $[T_2, T_3]$ are most suited for arm movements. By introducing a hysteresis which adds pixels in the range $[T_1, T_4]$ if they are connected with pixels from $[T_2, T_3]$, the extracted motion becomes less fragmented and includes more pixels of the full motion, while excluding noisy image regions. To avoid smaller motion regions caused by noise or unwanted motion along the body, the regions that have been classified within the range $[T_2, T_3]$ are filtered by size and connected component analysis.

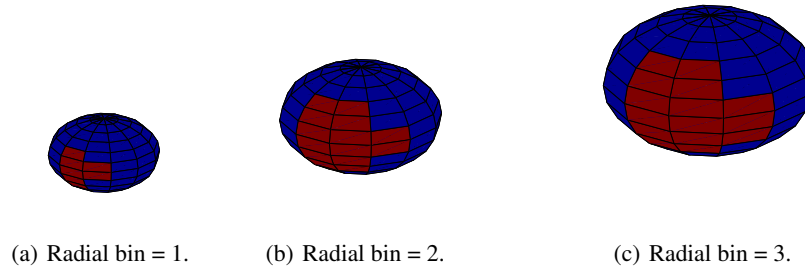


Fig. 3. Shape context for a “point right” gesture represented by the spherical histogram.

When the relevant motion has been extracted, the data is mapped to a 3D world coordinate system, resulting in a 3D motion point cloud representing the arm movements.

4 Gesture Representation

A 3D motion point cloud is represented efficiently by use of shape contexts. A shape context is based on a spherical histogram. This histogram is centered in a reference point, currently selected manually as the center of the upper body. The histogram is divided linearly into 12 azimuthal (east-west) bins and 12 colatitudinal (north-south) bins, while the radial direction is divided into 3 bins. A bin is either on (“1”) or off (“0”) depending on the number of 3D points in a particular bin. This results in an n ($= 12 \times 12 \times 3$) dimensional binary feature vector for each frame, see figure 3. A gesture is now represented by ANDing the shape contexts calculated for the temporal duration of the gesture.

It should be noted that this kind of representation requires known start and end points of a particular gesture.

5 Recognition of Gestures

A gesture is recognized by matching the current shape contexts histogram with a known set of histograms for each possible gesture. The matching is performed by a binary comparison of the histogram bins, resulting in a distance for each gesture. The distance measure is computed using XOR.

The gesture with the shortest distance (nearest neighbor) is selected as the best match.

6 Results

The purpose of this test is to give an indication of the system’s performance and not to produce final results. Six gestures have been used to make a small scale test of the current system: Point right, Move left, Move closer, Raise arm, Wave, and Clap.

We have recorded data of these six gestures performed by two different persons and matched each of the gestures captured of one of the persons with the six gestures from the other. All the matches are correct and have the distances shown in figure 4.

	1	2	3	4	5	6
1. Point right	11	39	41	40	25	88
2. Move left	29	13	19	46	31	62
3. Move closer	37	37	35	48	39	82
4. Raise arm	28	32	40	9	14	73
5. Wave	20	26	32	23	14	73
6. Clap	77	55	65	88	79	12

Fig. 4. Matching results for the six actions.

7 Conclusion

In this paper we have presented a gesture recognition approach operating on range data acquired with a CSEM SwissRanger SR-2 camera. This allows 3D representation of gestures based on shape contexts and a simple binary matching. By introducing a hysteresis bandpass filter, it is possible to extract useful motion information by double difference range images, although the noisy nature of the range data. A small scale test indicates promising results.

7.1 Future Work

The range based gesture recognition system is still in a preliminary phase and some further work are required. Some basic parts of the approach need to be investigated further. First of all an automatic method to estimate the reference point for histogram centering is required. Furthermore, an intelligent pre-selection of the range data of interest would be an advantage.

The current hysteresis bandpass filter operates in 2D difference images. It would be interesting to extend the filter to operate directly in the 3D range data if possible.

A more sophisticated matching strategy might improve the performance, e.g., Edit distance. Finally, we need to conduct a large scale test of the system.

References

1. R. Bostelman, T. Hong, R. Madhavan, and B. Weiss. 3d range imaging for urban search and rescue robotics research. In *Safety, Security and Rescue Robotics, Workshop, 2005 IEEE International*, pages 164–169, june 2005.

2. P. Michel, J. Chestnutt, S. Kagami, K. Nishiwaki, J. Kuffner, and T. Kanade. Online environment reconstruction for biped navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'06)*, pages 3089–3094, Orlando, Florida, May 2006.
3. T. Oggier, M. Stamm, M. Schweizer, and J. Pedersen. User manual swissranger 2 rev. b. Version 1.02, March 2005.
4. D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3), November/December 2006.