21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Human Detection by Haar-like Filtering using Depth Information

Sho Ikemura, Hironobu Fujiyoshi
Dept. of Computer Science, Chubu Univ. Aichi, 487-8501 Japan
si@vision.cs.chubu.ac.jp, hf@cs.chubu.ac.jp

Abstract

We propose a high-accuracy human detection method featuring a Haar-like filter expressing the human shape and using depth information obtained by capturing people from above with a time-of-flight (TOF) camera. This method extracts object regions by performing background subtraction against this depth information, and passes these extracted object regions through a Haar-like filter based on a human model expressing the convex shape of shoulder-head-shoulder. Human detection is achieved by integrating the results of this filtering by mean-shift clustering. The proposed method improves detection rate by 5.7% compared to a human-detection technique that simply applies mean-shift clustering to depth information obtained by background subtraction. We show that our method can detect humans in real time at a frame rate of about 19 fps.

1 Introduction

Depth-imaging sensors such as time-of-flight (TOF) cameras and the Kinect sensor device that can measure depth in real time have been attracting attention in recent years as candidates for application to human-sensing techniques. Conventional human detection methods combine gradient-based features such as histograms-of-oriented-gradients (HOG) features [?] with statistical learning methods such as AdaBoost and support vector machines (SVM) [?, ?, ?]. It has been reported that these methods are capable of high-accuracy human detection with high generalization since they construct human classifiers that detect localized human shapes. At the same time, complex backgrounds and human occlusion can make it difficult to discern object shapes in the case of gradient-based features extracted from images taken with visible-light cameras. A human detection method using depth information as a countermeasure to complex backgrounds and human occlusion has been proposed [?], but it is based on statistical learning and requires a large amount of training data to be prepared beforehand. Detection by that method also becomes difficult in environments that differ from the ones in which training data were prepared. These problems must be solved to make the method practical.

In response to the issues described above, human detection methods that capture humans from above without the use of statistical learning are being proposed [?, ?, ?, ?]. The advantage of capturing humans from above is that the effects of human occlusion are eliminated and change in background is minimal, which means that extraction of object regions can be easily performed by background-subtraction processing. The human detection system commercialized by Giken Trastem [?] achieves high-accuracy, real-time human detection by using a camera attached to a ceiling to capture people from above and by expressing human shapes by the vector focus point method. This method, however, requires a human model to be prepared beforehand, and as a result, detection accuracy drops in the case of people with features significantly different from the model. The Censys3D People Tracking System [?] from Point Grey Research, meanwhile, achieves human detection by capturing people from above with a stereo camera and using the depth information so obtained to detect the shape of the human head. However, as this method detects a person based on the height of a detected head, it is thought that the presence of an object of similar height may lead to erroneous detection.

In light of the above, we propose a human detection method without statistical learning that uses real-time depth images of pedestrians taken from above by a TOF camera and that applies Haar-like filtering based on a human model extracting the convex shape of shoulder-head-shoulder.

2 Proposed Method

The first step in the proposed method is to extract object regions by performing background subtraction against depth images. Then, to determine whether these extracted object regions correspond to people, the next step is to perform Haar-like filtering to extract the convex shape of shoulder-head-shoulder with respect to

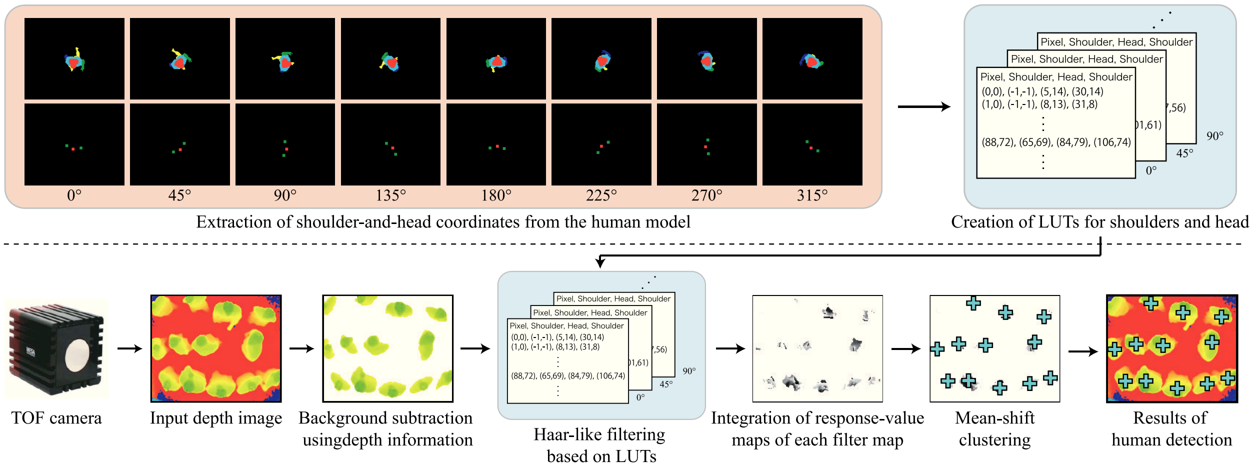

0°
45°
90°
135°
180°
225°
270°
315°
Extraction of shoulder-and-head coordinates from the human model
Pixel, Shoulder, Head, Shoulder
(0,0), (-1,-1), (5,14), (30,14)
(1,0), (-1,-1), (8,13), (31,8)
(88,72), (65,69), (84,79), (106,74)
0°
45°
90°
Creation of LUTs for shoulders and head
TOF camera
Input depth image
Background subtraction usingdepth information
Haar-like filtering based on LUTs
Integration of response-value maps of each filter map
Mean-shift clustering
Results of human detection

Figure 1. Flow of human detection by the proposed method.

each object region. Human detection can now be performed by using mean-shift clustering to integrate the points of the convex shape extracted by Haar-like filtering. The flow of human detection by Haar-like filtering based on a human model is shown in Figure 1.

2.1 Extraction of convex shape by Haar-like filtering

To determine whether an extracted object region is human or not, the proposed method performs Haar-like filtering that extracts the convex shape formed by the shoulder-head-shoulder combination in humans. A Haar-like filter [**?**] takes the difference in brightness between black and white regions to be its response value. The method described by Viola et al. achieves high-speed, high-accuracy face detection by using Haar-like filters to determine contrast in the human face [**?**].

When applying Haar-like filtering to depth images, the response value is taken to be the difference in the average depth between the black and white regions. Specifically, if the black region is higher than the white region, a positive response value is output, and conversely, if the white region is higher than the black region, a negative response value is output. To capture the difference in height between the human head and shoulders, the proposed method uses a Haar-like filter composed of white, black, and white regions (Figure 2). The response value $H(r1, r2)$ of this Haar-like filter is calculated by the following equation using average depth $S(r1)$ of black region $r1$ and average depth $S(r2)$ of white regions $r2$.

$$H(r1, r2) = S(r1) - S(r2)/2 \qquad (1)$$

Furthermore, to accommodate different human orientations, the proposed method uses multi-directional filtering in the four directions of 0, 45, 90, and 135 This

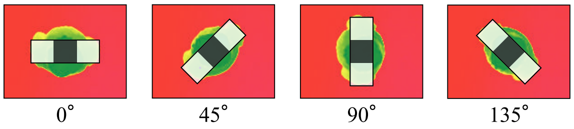


Figure 2. Haar-like filters used in the proposed method.

filtering process creates a convex filter map $F(u, v, d)$ by the threshold processing of Eq. (2) using response value $H(r1, r2)$ of the Haar-like filter calculated above. Here, (u, v) denotes image coordinates and d denotes filter direction.

$$F(u, v, d) = \begin{cases} 1 & \text{if } H(r1, r2, d) > \text{th} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

2.2 Haar-like-filter design based on a 3D human model

The fixed Haar-like filter of section 2.1 can be used to perform filtering against all image coordinates. There are times, however, when such fixed-filter filtering may encounter difficulties since the way in which the shoulders-and-head combination appears changes according to the position of the person (i.e. the place where the person is standing) in the image. Accordingly, to enable the proposed method to deal with variations in the position, orientation, and height of people, various arrangements of the black and white regions of the Haar-like filter are determined beforehand using a 3D human model. Specifically, for each pixel in the image as a center point, the orientation of the human model shown in Figure 3 was varied from 0to 360in 45increments for a total of eight orientations. The positions of the shoulders and head in this 3D human model were then extracted for each orientation at each pixel, and the resulting information was recorded in the form of look-up tables (LUTs). These LUTs contained the coordinates of the left-and-right shoulders and head of the human model for each center coordinate and orientation (angle) taken up by the human model. Additionally, to deal with variations in the height of people, the height of the human model was varied from 0.5 to 2.0 m in 0.1-m increments and LUTS were created in a similar manner.

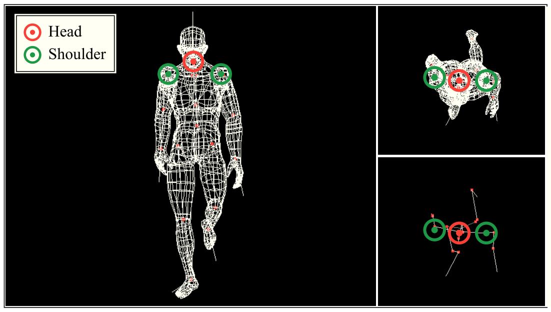


Figure 3. Extraction of shoulder and head positions from a 3D human model.

In this way, Haar-like filtering that takes into account the position, orientation, and height of people can be performed based on LUTs created beforehand. Thus, as shown in Figure 4, each coordinate in an extracted object region can be treated as a target coordinate, and potential shoulders-and-head positions can be determined by cross-referencing that target coordinate with the information recorded in the LUTs. In short, by adjusting the black and white regions of the Haar-like filter according to shoulders-and-head positions established beforehand in the above way, filtering that accommodates changes in human appearance can be performed.

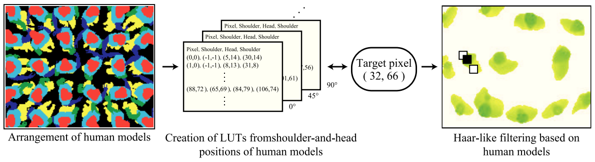

Pixel, Shoulder, Head, Shoulder
Pixel, Shoulder, Head, Shoulder
Pixel, Shoulder, Head, Shoulder
(0,0), (-1,-1), (5,14), (30,14)
(1,0), (-1,-1), (8,13), (31,8)
(88,72), (65,69), (84,79), (106,74)
,56)
1,61)
90°
45°
0°
Target pixel
(32, 66)
Arrangement of human models
Creation of LUTs fromshoulder-and-head positions of human models
Haar-like filtering based on human models

Figure 4. Haar-like filtering based on LUTs

2.3 Integration of multi-directional filter maps

In this step, the filter maps in four directions $F(u, v, d)$ obtained by Haar-like filtering are integrated into a single filter map. As shown by Eq. (3), if at least one of the convex-shape labels in the four-direction filter map $F(u, v, d)$ at coordinate (u, v) is 1, then filter map $F_{all}(u, v)$ after integration will likewise be equal to 1.

$$F_{all}(u, v) = \delta(\sum_{d} F(u, v, d) > 0) \tag{3}$$

2.4 3D mean-shift clustering of depth information

Human detection is now performed by clustering depth information based on the integrated filter map described above. The proposed method performs this clustering of 3D depth information by a mean-shift procedure [?] in which a mean-shift vector $m(\mathbf{x})$ is calculated by Eq. (4). Here, the 3D coordinate for which the label of the integrated filter map $F_{all}(u, v)$ is 1 is denoted as $\mathbf{x}_i$ and the 3D coordinate of the moving target point is denoted as $\mathbf{x}$. The symbol k denotes the kernel function and h denotes bandwidth. In this study, $h = 0.15$ m.

$$m(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i k\left(\left|\left|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right|\right|^2\right)}{\sum_{i=1}^{n} k\left(\left|\left|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right|\right|^2\right)} - \mathbf{x} \tag{4}$$

3 Evaluation of Proposed Method by Human-detection Experiment

We performed an experiment to assess the effectiveness of the proposed method as described below.

3.1 Database

We performed an experiment using a set of sequences capturing pedestrians from above using a TOF camera. Specifically, we used a database of three sequences taken in different environments as listed in Table 1. Sequence 1 targeted adults under crowded conditions. Sequence 2, meanwhile, targeted both adults and children against a background including a non-human object. Sequence 3, finally, targeted both adults and children walking up and down a staircase. Here, the natural variation in the height of the stairs caused human height to vary as well.

3.2 Comparison of detection accuracies

On performing this human-detection experiment, we compared detection accuracies of three detection methods- conventional method, proposed method 1, and proposed method 2- against this database of three evaluation sequences. The conventional method uses mean-shift clustering [?] to integrate the depth information of

Table 1. Database

	Height of TOF camera [m]	Body heights [cm]
Sequence 1	4.5	Adult : 165 ～ 185
Sequence 2	3.2	Adult : 165 ～ 175 Children : 100 ～ 120
Sequence 3 (staircase)	3.5	Adult : 165 ～ 175 Children : 100 ～ 120

Table 2. Detection accuracies.

	Sequence 1	Sequence 2	Sequence 3	Average
Conventional method	89.7	92.5	88.0	90.1
Proposed method 1	87.0	90.2	85.6	87.6
Proposed method 2	98.6	91.3	97.4	95.8

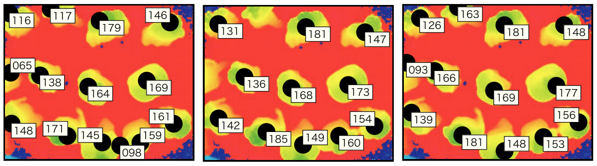
an object region extracted by background subtraction. Proposed method 1 performs human detection by filtering that arranges the black and white regions of the Haar-like filter in a linear manner, and proposed method 2 performs human detection by Haar-like filtering based on a human model. The accuracy of human detection is calculated by Eq. (5).

$$Detection\ rate\ [\%] = \frac{Detected\ [persons]}{Actual\ number\ [persons]} \quad (5)$$

Table 2 lists human-detection accuracies for the existing method, proposed method 1, and proposed method 2 for each of the three sequences described above. Examining these results, it can be seen that average detection accuracy for the three sequences by proposed method 1 is 2.5% lower than that of the existing method. The reason for this is that proposed method 1 uses a Haar-like filter with black and white regions aligned linearly, which cannot deal with changes in human appearance caused by a person standing at different positions. On the other hand, proposed method 2 improves average detection accuracy by 5.7% over that of the existing method. This is because proposed method 2 performs Haar-like filtering that uses a human model to take into account changes in human appearance at different positions.

3.3 Examples of human detection

Examples of human detection using the above three methods are shown in Figure 5. The black points in these examples represent human-detection points and the numerals indicate the heights of the detected people. For sequence 1, it can be seen that the existing method performs human detection with high accuracy but nevertheless fails in mean-shift clustering when two people are standing adjacent to each other resulting in an erroneous detection. Proposed method 1 has difficultly detecting people whose appearance has changed or people present at the edge of the image since it uses a linearly arranged Haar-like filter. Proposed method 2, however, performs high-accuracy human detection even for people whose appearance has changed or people present at the edge of the image since it performs Haar-like filtering based on a human model.



Conventional method Proposed method 1 Proposed method 2

Figure 5. Example of human detection.

4 Conclusion

We proposed a high-accuracy human detection method using depth information obtained by capturing people from above with a TOF camera and applying a Haar-like filter based on a 3D human model. It was shown that the proposed method improves detection rate by 5.7% compared to our existing method and that it can perform human detection in real time at a frame rate of about 19 fps.

References