Real-Time Human Detection using Relational Depth Similarity Features

Sho Ikemura, Hironobu Fujiyoshi

Dept. of Computer Science, Chubu University.
Matsumoto 1200, Kasugai, Aichi, 487-8501 Japan.
si@vision.cs.chubu.ac.jp, hf@cs.chubu.ac.jp
http://www.vision.cs.chubu.ac.jp

Abstract. Many conventional human detection methods use features based on gradients, such as histograms of oriented gradients (HOG), but human occlusions and complex backgrounds make accurate human detection difficult. Furthermore, real-time processing also presents problems because the use of raster scanning while varying the window scale comes at a high computational cost. To overcome these problems, we propose a method for detecting humans by Relational Depth Similarity Features(RDSF) based on depth information obtained from a TOF camera. Our method calculates the features derived from a similarity of depth histograms that represent the relationship between two local regions. During the process of detection, by using raster scanning in a 3D space, a considerable increase in speed is achieved. In addition, we perform highly accurate classification by considering of occlusion regions. Our method achieved a detection rate of 95.3% with a false positive rate of 1.0%. It also had a 11.5% higher performance than the conventional method, and our detection system can run in real-time (10 fps).

1 Introduction

There has recently been interest into the implementation of techniques that will assist in comprehending the intentions of people within spaces such as offices, homes, and public facilities. In order to implement techniques of monitoring people in this manner, it is necessary to know where people are within such a space, and it has become a challenge to implement human detection that is highly accurate and also fast. There has been much research in the past into human detection, and various different methods have been proposed[1][2][3][4][5]. Among human detection methods that use conventional visible-light cameras, there are methods that involve statistical training with local features and boosting. Gradient-based features such as HOG[1], EOH[2], and edgelets[5] are often used as local features, and there have been reports that these enable the implementation of highly accurate human detection. However, gradient-based features obtained from visible-light camera images encounter difficulties in perceiving the shapes of human beings when there are complex backgrounds and when people overlap each other, and the detection accuracy can drop as a result. To counter

this problem, Ess et al. have proposed a highly-accurate human detection method for confusing scenes, using depth information obtained by stereo cameras[6]. However, the acquisition of depth information by stereo cameras necessitates correspondence calculations between images by camera calibration and stereo matching, so the processing costs are high and real-time detection is difficult. In addition, since the sizes of the humans within the image is unknown, conventional human detection methods also have problems in that repeated raster scans while varying the scale of the detection window increases the computational cost and makes real-time processing difficult.

This paper proposes a real-time human detection method that uses depth information obtained from a time-of-flight (TOF) camera and can cope with overlapping people and complex scenes. The proposed method uses depth information obtained from the TOF camera to calculate relational depth similarity features (RDSFs) that determine depth information for local regions, and constructs a final classifier by Real AdaBoost. It uses the thus-constructed classifiers to detect humans, and implements faster raster scanning of detection windows in a 3D space and also improves the detection accuracy by considering occlusion regions.

2 Local features based on depth information

A TOF camera is a camera that measures the distance to an object by measuring the time taken for infrared light that is emitted from LEDs located around the camera to be reflected by the object being detected and observed by the camera. In this study, we use a MESA SR-3100 as the TOF camera. The SR-3100 can acquire depth information in real-time from 0.3 m to 7.5 m (with a resolution of 22 mm at 3 m), but it cannot photograph outdoors so it is limited to use indoors. When detecting humans, it is considered effective to use depth information obtained by a TOF camera to perceive the depthwise relationships of human bodies and the background. Thus this method proposes the use of a relational depth similarity feature obtained from depth distributions of two local regions.

2.1 Relational depth similarity features

A relational depth similarity feature is used to denote the degree of similarity of depth histograms obtained from two local regions. As shown in Fig. 1, we divide each depth image into local regions that are cells of 8 x 8 pixels, and select two cells. We compute depth histograms from the depth information of each of the two cells selected in this way, then normalize them so that the total value of each depth histogram is 1. If each bin of the two normalized depth histograms p and q created from the thus computed m bins is p_n and q_n, we compute the degree of similarity S between them from the Bhattacharyya distance[7] and use that

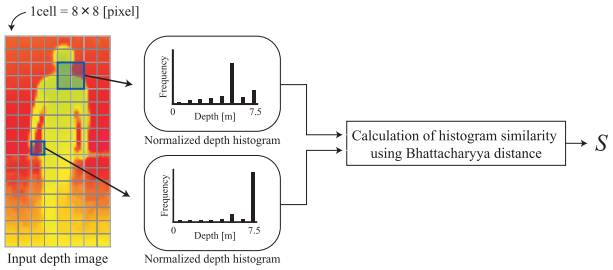

Fig. 1. Calculation of RDSF.

as an RDSF.

$$S = \sum_{n=1}^{m} \sqrt{p_n q_n} \tag{1}$$

Since the RDSF is a feature obtained from the degree of similarity of depth information for two regions, it becomes a feature that expresses the relative depth relationship between the two regions.

2.2 Varied rectangular region sizes

Based on the processing of Section 2.1, we calculate an feature vector of RDSF by calculating the degrees of similarity for all combinations of rectangular regions, as shown in Fig. 2. During this process, we use Equation (2) for normalization. In this case, if p_n is the nth bin of the depth histogram, the nth bin p'_n of the normalized depth histogram can be obtained from the following equation:

$$p'_n = \frac{p_n}{\sum_{i=1}^{m} p_i} \tag{2}$$

With this method, the detection window size is set to 64 x 128 pixels so it can be divided into 8 x 16 cells. There are 492 rectangular regions obtained by varying the cell units of the rectangular region from 1 x 1 to 8 x 8 to compute depth histogram. To calculate the RDSF from combinations of the 492 rectangular regions, ${}_{492}\mathrm{C}_2 = 120,786$ feature candidates are extracted from within one detection window.

2.3 Faster depth histogram calculations by integral histograms

To reduce computational costs during the feature calculations, this method uses integral histograms[8] to compute the depth histograms rapidly. We first quantize the depth of each pixel to a 0.3-m spacing. Since this study divides the distances 0

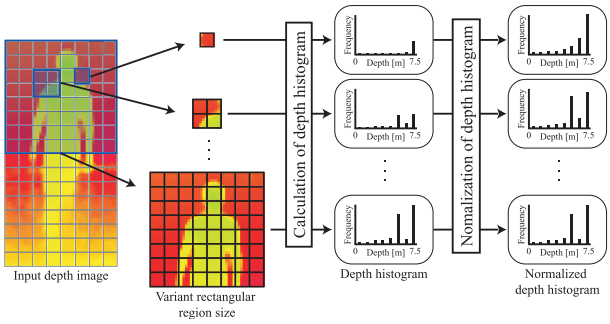

Fig. 2. Normalization of depth histograms for various rectangular region sizes.

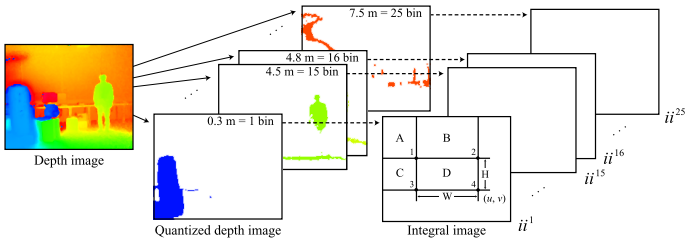

Fig. 3. Calculation of integral histogram.

m to 7.5 m by a 0.3-m spacing, that means we compute depth histograms formed of 25 bins. We then create 25 quantized images in corresponding to bin n, as shown in Fig. 3, and compute an integrated image $ii^n(u, v)$ from the quantized images $i^n(u, v)$, using Equations (3) and (4):

$$s^n(u, v) = s^n(u, v-1) + i^n(u, v) \tag{3}$$

$$ii^n(u, v) = ii^n(u-1, v) + s^n(u, v) \tag{4}$$

In this case, $s^n(u, v)$ denotes the sum of pixels in the rows of bin n and $ii^n(u, v)$ denotes the sum of s^n of the columns. Note that we assume that $s^n(u, -1) = 0$ and $ii^n(-1, v) = 0$. In the calculation of the nth bin D^n of the depth histogram from the region D in Fig. 3, it would be sufficient to obtain the sum from four points of the nth integrated image ii^n, from the following equation:

$$\begin{aligned} D^n = (ii^n(u, v) + ii^n(u-W, v-H)) - \\ (ii^n(u-W, v) + ii^n(u, v-H)) \end{aligned} \tag{5}$$

This makes it possible to rapidly obtain the value of the nth bin of the depth histogram, irrespective of the size of the region.

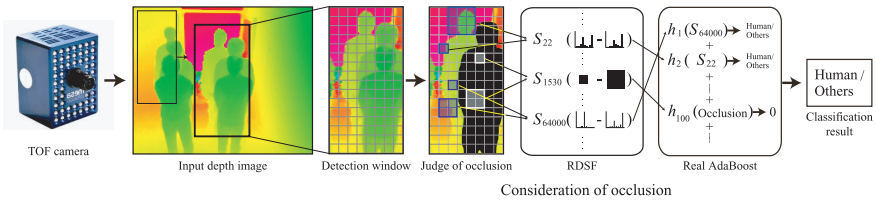

Fig. 4. Flow of human detection using depth information.

3 Human detection using depth information

The flow of human detection in accordance with the proposed method is shown in Fig. 4. The proposed method first performs a raster scan of the detection windows in a 3D space. It then computes the RDSFs from the detection windows. It judges whether there are occlusions in the calculated features, and uses Real AdaBoost to classify whether each detection window is of a human or a non-human object. Finally, it integrates the detection windows that have been classified as human by mean-shift clustering in the 3D space, to determine the location of human.

3.1 Construction of classifiers by Real Adaboost

The proposed method uses Real AdaBoost[9] in the human classification. Real AdaBoost obtains degrees of separation from the probability density functions for each dimension of features in positive classes and negative classes, and selects the features that enable the greatest separation between positive and negative classes as weak classifiers. Since the degrees of separation are handled as evaluated values during this process, the output of the classification results can be done as real numbers. If a weak classifier selected by the training is $h_t(x)$, the final classifier $H(x)$ that is constructed is given by the following equation:

$$H(x) = \mathrm{sign}(\sum_{t=1}^{T} h_t(x)) \tag{6}$$

3.2 Raster scanning in 3D space

Conventional human detection methods involve repeated raster scans while the scale of the detection window is varied, so there are many detection windows that do not match the dimensions of humans. With the proposed method, we determine the detection windows to correspond to the sizes of humans, by using depth information to perform raster scans in a 3D space, which speeds up the processing. With $y_w = 0$, 60 x 180 [cm] detection windows in the $x_w - z_w$ plane are subjected to raster scanning, as shown in Fig. 5. The 3D coordinates of each detection window obtained by the raster scan in the 3D space are projected onto image coordinates $[u, v]^{\mathrm{T}}$, using Equation (7), and a feature is calculated

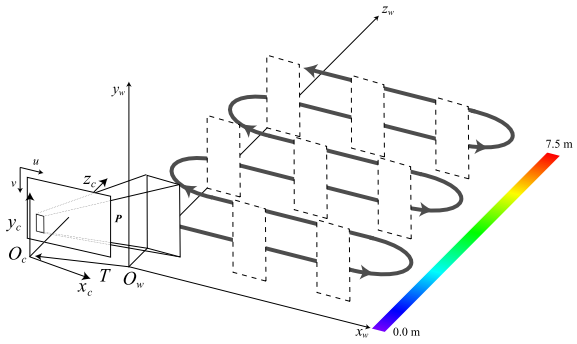

Fig. 5. Raster scanning in 3D space.

from the depth information within the region corresponding to the projected coordinate position.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \boldsymbol{P} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{7}$$

$$\boldsymbol{P} = \boldsymbol{A}\,[\boldsymbol{R}|\boldsymbol{T}] \tag{8}$$

The matrix $\boldsymbol{P}$ is a perspective projection matrix which is computed from an intrinsic parameter obtained by camera calibration, a rotation matrix $\boldsymbol{R}$ that is an extrinsic parameter, and a translation vector $\boldsymbol{T}$. Simple camera calibration can be done to enable the TOF camera to acquire the global coordinates (x_w, y_w, z_w) within a 5 x 4 x 7.5 [m] space corresponding to the image coordinates (u, v). In this case, the TOF camera uses a CCD sensor to observe infrared light, so the intrinsic parameter $\boldsymbol{A}$ is similar to that of an ordinary camera. In addition, Mure-Dubois, et al. have compared published intrinsic parameters and the results of actual calibrations and confirmed that the intrinsic parameters are close[10].

3.3 Classification adjusted for occlusions

In a confusing scene in which a number of people are overlapping, occlusions can occur in the human regions that are being observed. Depth information extracted from an occlusion region is the cause of the output of erroneous responds for weak classifiers. Thus we ensure that any output of weak classifiers that perceive such occlusion regions is not integrated into the final classifier without modification. Since the proposed method performs a raster scan of a detection window in a 3D space, the position of the window in global coordinates is known. In this case, we determine that any object region that is closer to the camera than the detection

Fig. 6. Examples of occluded regions.

window is an occlusion, by comparing depth information acquired from the TOF camera, and use that in the classification. In this study, we assume that when there is natural overlapping between people, the depth from one person who is in the front and another person who is in the rear will be at least 0.3 m, and that any object region that is at least 0.3 m closer than the detection window for a person being detected is an occlusion.

Extraction of occlusion regions We use the depth z_w of the detection window during the raster scanning of the 3D space to determine the threshold for occlusion judgement. If we assume that each pixel in the detection window is (u, v) and the depth map thereof is $d(u, v)$, the occlusion label $O(u, v)$ at each set of coordinates is given by the following equation:

$$O(u,v) = \begin{cases} 1 & \text{if } d(u,v) < z_w - 0.3 \text{ m} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The extracted occlusion regions are shown in Fig. 6 as black areas.

Classification from consideration of occlusion regions If we assume that the proportion of occlusion regions existing within a rectangular region B_t, which is the target of the tth weak classifier $h_t(x)$, is an occlusion rate OR_t, it can be obtained from the following equation:

$$OR_t = \frac{1}{B_t} \sum_{(u,v) \in B_t} O(u,v) \tag{10}$$

Using the thus-computed occlusion rate OR_t, the final classifier $H'(x)$ from consideration of occlusion regions is expressed by Equation (11). During this time, the output of weak classifiers that have been computed from regions in which occlusions occur can be restrained by applying the proportion of occlusions as weighting to the weak classifiers.

$$H'(x) = \text{sign}(\sum_{t=1}^{T} h_t(x) \cdot (1 - OR_t)) \tag{11}$$

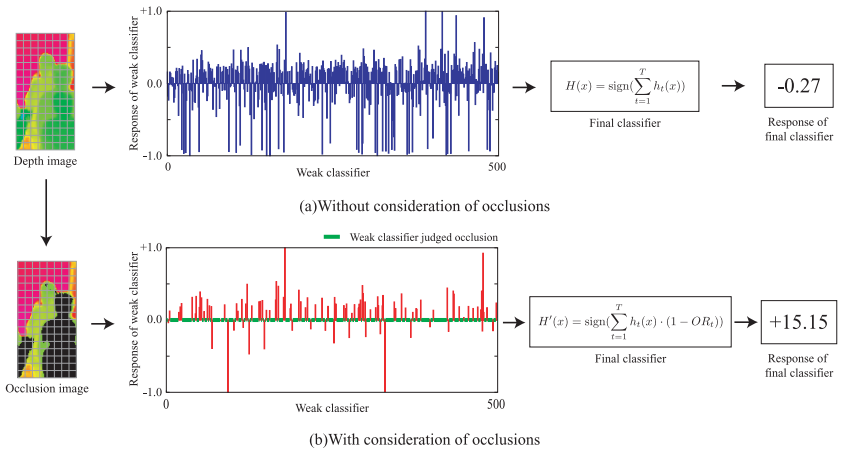

Fig. 7. Examples of classifications with and without consideration of occlusions.

If the classification by final classifiers is done without considering occlusion regions, as shown in Fig. 7(a), the output of a large number of weak classifiers is a disadvantage and as a result, non-human objects are mistakenly classified. On the other hand, Fig. 7(b) shows an example in which the output of final classifiers is obtained from consideration of occlusion rates, in which humans are classified correctly.

3.4 Mean-shift clustering in 3D space

In conventional human detection with a visible-light camera[3], the detection windows that have been classified as representing humans are integrated by mean-shift clustering[11] and taken as detection results. However, with mean-shift clustering alone in an image space, detection windows could be erroneously integrated if humans overlap in them, as shown in (b) and (d) of Fig. 8. With the proposed method, mean-shift clustering is done within a 3D space, as shown in (c) and (e) of Fig. 8. Since this enables the separation of clusters by depth information, even when humans are overlapping, the erroneous integration of detection windows can be suppressed.

3D mean-shift clustering calculates the mean-shift vector $m(\mathbf{x})$ from Equation (12). In this case, $\mathbf{x}$ denotes the center coordinate of the detection window and $\mathbf{x}_i$ denotes the 3D coordinate of each data item. k is a kernel function and h is the bandwidth, which in this study is taken to be $h = 0.3$ m.

$$m(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i k\left(\left|\left|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right|\right|^2\right)}{\sum_{i=1}^{n} k\left(\left|\left|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right|\right|^2\right)} - \mathbf{x} \tag{12}$$

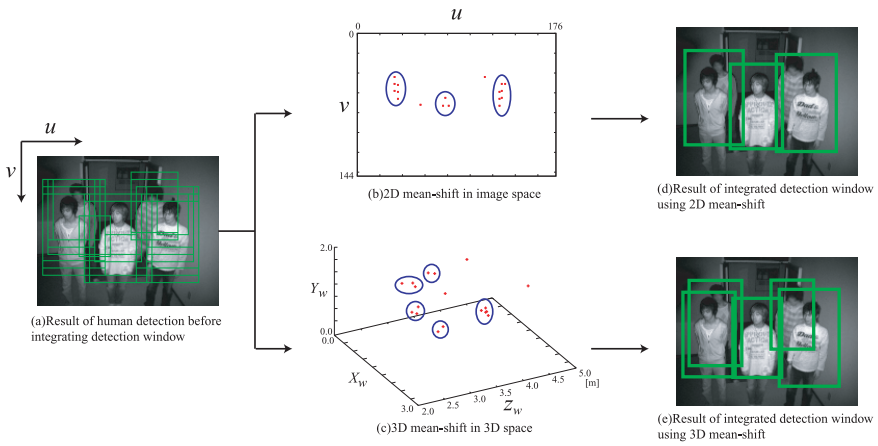

Fig. 8. Integration of detection windows by using mean-shift clustering.

4 Evaluation of the proposed method by classification experiments

We performed evaluation experiments to confirm the validity of the proposed method.

4.1 Database

For the database, we used sequences shot by a TOF camera. We installed the TOF camera at a height of approximately 2.5 m indoors, and targeted scenes of people walking and scenes in which a number of people overlap. We used 1346 positive examples for training and 10,000 negative examples for training, taken from sequences that had been shot indoors. In the evaluation, we used 2206 positive samples for evaluation and 8100 negative samples for evaluation. Since the TOF camera was set up to take pictures up to a maximum distance of 7.5 m indoors, it was difficult to use it to photograph the whole bodies of a number of humans. For that reason, the top 60% of the whole bodies of the humans were the targets for these experiments. Part of the database that was used for evaluation is shown in Fig. 9.

4.2 Feature evaluation experiments

Using the database for evaluation, we performed human classification experiments and compared them by feature classification accuracy. In the experiments, we compared features by using HOG features extracted from depth images, RDSFs, and both HOG features and RDSFs. Note that since these experiments were intended to evaluate features, there was no classifications adjusted

for occlusions. We use receiver operating characteristic (ROC) curves in the comparison of the experiment results. A ROC curve plots false positive rate along the horizontal axis and detection rate along the vertical axis. It is possible to compare detection rate with respect to false positive rate by varying the classifier thresholds. The detection performance is better towards the top left of the graph.

The results of feature evaluation experiments are shown in Fig. 10. RDSFs gave a detection rate of 95.3% with a false positive rate of 1.0%, which is an improvement of 11.5% over the classification rate of HOG features of depth images. A comparison of RDSFs alone and features obtained by both HOG features and RDSFs showed that the detection accuracy was the same. This shows that RDSFs are mainly (98%) selected during the weak classifier training and HOG features do not contribute to the classification. Examples of missed classifications are shown in Fig. 11. It is clear that the samples that tended to be erroneously classified involved large variations in shape or occlusions.

4.3 Evaluation experiments adjusted for occlusions

To demonstrate the validity of occlusion adjustment in the proposed method, we evaluated it by human classification experiments.

The results of evaluation experiments with and without occlusion adjustment are shown in Fig. 12. RDSFs with occlusion adjustment gave a detection rate of 97.8% with a false positive rate of 1.0%, which makes it clear that this method enables highly accurate classification even when occlusions occur. In addition, the detection rate improved even with HOG features with occlusion adjustment. This shows that it is possible to suppress the effects of occlusion regions by using occlusion rates to weight weak classifiers that are valid for the classification, to obtain output of the final classifiers.

Fig. 9. Examples of test data.

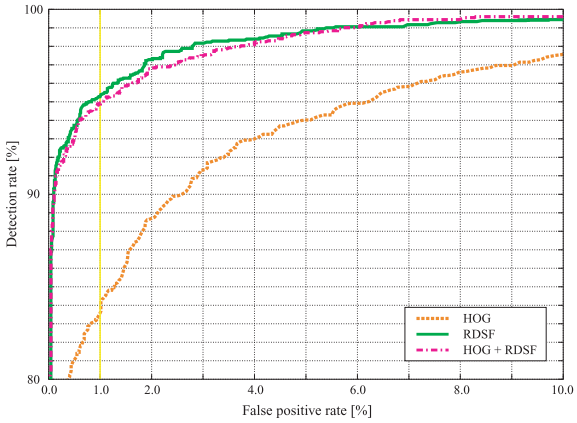

Fig. 10. Results of using features.

4.4 Features selected by training

Features that weak classifiers have selected in the Real AdaBoost training are shown in Fig. 13. With the HOG features of (a), features are selected in such a manner that the boundary lines between humans and the background such as the edges of the head and shoulders are perceived. It is clear that features of the upper half of bodies with few shape variations are often selected, as the tendency of training of up to 50 rounds. Next, with the RDSFs of (b), the selection is such that combinations of adjoining human regions and background regions are perceived in the first and third training rounds. Differing from the perception of boundaries at lines, such as with HOG features, boundaries are perceived by regions with RDSFs. This is considered to make the method robust in positioning humans. Boundaries were also perceived in the ninth and eleventh training rounds, but separated rectangular regions were selected, which differs from the first and third rounds. This is considered to make it possible to absorb variations in boundary position, since there are large variations in the lower halves of human bodies. In each of the second, fifth, and seventh training rounds, regions that tend to determine vertical or lateral symmetrical depth relationships of the human shape were selected. In each of the thirty-fourth and fortieth training rounds, two regions in the interior of the human were selected. When there are rectangular regions of the same depth, those two rectangular regions can be taken to represent a surface of human. The tendency with up to 50 rounds of training makes it clear that large rectangular regions were selected during the initial training rounds to determine the depth relationships of large-scale human regions. As the training proceeded, the selection was such that local depth relationships were perceived by selecting smaller rectangular regions.

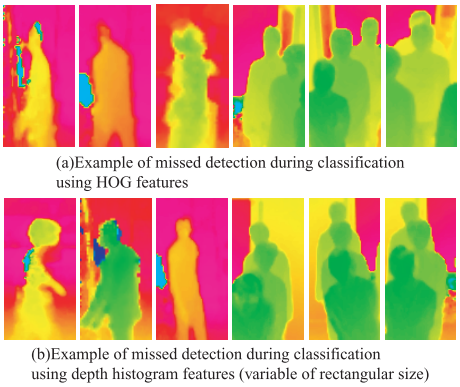

Fig. 11. Examples of missed detection during classification.

Table 1. Computational costs per frame [ms].

	Proccesing cost of 1 detection window	Total (361 windows)
Feature calculation	0.067	24.31
Classification	0.125	45.34
Integration of windows	–	31.97
Total	–	101.62

4.5 Real-time human detection

The examples of human detection using raster scanning of detection windows in a 3D space are shown in Fig. 14. From (a), we see that an object of a similar height to people is not detected erroneously and only the people are detected. From (b) and (c), we see that the 3D position of each person can be detected accurately, even when there is overlapping of people who are facing in different directions. The processing times for one frame (361 detection windows) when an Intel Xeon 3-GHz CPU was used are shown in Table 1. Since the proposed method performs the processing in approximately 100 ms, it enables real-time detection at approximately 10 fps.

5 Conclusions

In this paper, we proposed a real-time human detection method that uses depth information obtained from a TOF camera and can cope with overlapping people and complex scenes. The results of evaluation experiments show that this method enables a 11.5% improvement in detection rate over the use of HOG features, which is a conventional method. In addition, we have confirmed that

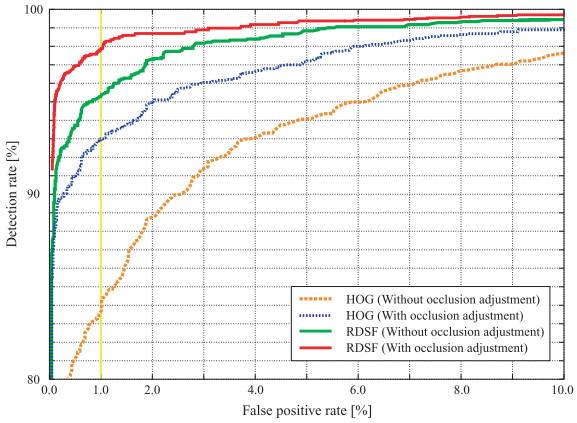

Fig. 12. Results of occlusion consideration.

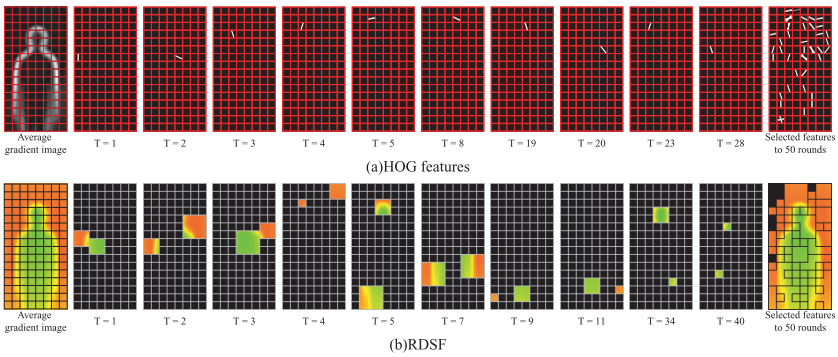

Fig. 13. Features selected by learning.

the proposed method enables highly-accurate classifications even when occlusions occur, by calculating the occlusion rate within the detection window, and performing classifications from consideration of occlusion regions. Since the proposed method requires a total processing time of only approximately 100 ms for the computation and classification of features and the integration of detection windows, it enables real-time detection at approximately 10 fps. In the future, we plan to perform motion analysis using depth information and its variation with time.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR **1** (2005) 886–893

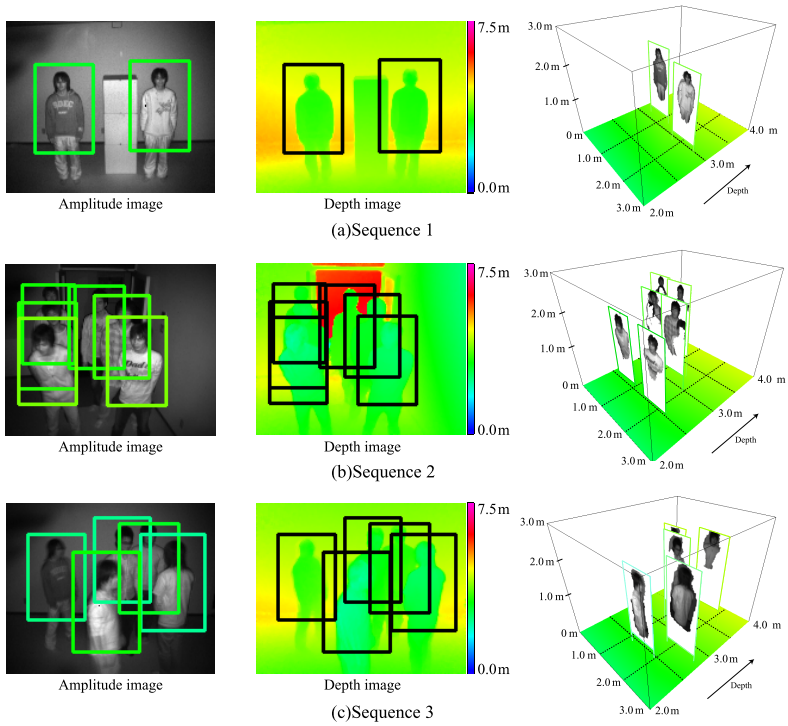

Fig. 14. Example of human detection.

2. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. CVPR **2** (2004) 53–60
3. Mitsui, T., Fujiyoshi, H.: Object detection by joint features based on two-stage boosting. International Workshop on Visual Surveillance (2009) 1169–1176
4. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet feature. CVPR (2007) 511–518
5. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. ICCV **1** (2005) 90–97
6. A. Ess, B.L., Gool, L.V.: Depth and appearance for mobile scene analysis. ICCV (2007)
7. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by probability distributions. Bull. Calcutta Math. Soc. **35** (1943) 99–109
8. Porikli, F.: Integral histogram: A fast way to extract histograms in cartesian spaces. CVPR (2005) 829–836
9. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning **37** (1999) 297–336
10. Mure-Dubois, J., Hugli, H.: Fusion of time of flight camera point clouds. ECCV workshop on M2SFA2 (2008)
11. Comaniciu, D., Meer, P.: Mean shift analysis and applications. ICCV (1999) 1197–1203