# Fast Human Detection Combining Range Image Segmentation and Local Feature Based Detection

Toru Ubukata*, Masatoshi Shibata*, Kenji Terabayashi†, Alessandro Moro‡,
Takehiro Kawashita*, Gakuto Masuyama§ and Kazunori Umeda§
*School of Science and Engineering, Chuo University, Japan
{shibata,kawashita}@sensor.mech.chuo-u.ac.jp
†Faculty of Technology, Shizuoka University, Japan, tera@eng.shizuoka.ac.jp
‡Ritecs, Japan, alessandromoro.italy@ritecs.co.jp
§Faculty of Science and Engineering, Chuo University, Japan
{masuyama,umeda}@mech.chuo-u.ac.jp

***Abstract*—This paper proposes a human detection method that combines range image segmentation and human detection based on image local features. The method uses a stereo vision system called Subtraction Stereo, which extracts a range image of foreground regions. An extracted range image is segmented for each object by Mean Shift Clustering. Human detection based on local features is applied to each segment of foreground regions to detect humans. In this process, regions to scan a detection window for extracting local features are restricted. In addition, the size of the detection window is obtained using the distance information of a range image and camera parameters. Therefore, processing time and false detection can be reduced. Joint HOG features are used as the image local features. When applying the Joint HOG based human detection, occlusion of multiple humans is considered in construction of a classifier and in integration of detection windows, which improves the detection performance for the occluded humans. The proposed method is evaluated by experiments comparing with the method using Joint HOG features only. 11fps fast human detection is achieved.**

## I. Introduction

Human detection from images is a key technology for surveillance, marketing, etc., and many studies have been done on this topic. The typical approach is to extract local features and apply a statistical learning method such as Boosting [1] or Support Vector Machine (SVM) [2] for the features [3] [4]. One of the most popular local features is Histogram of Oriented Gradients (HOG) proposed by Dalal and Triggs [5]. The HOG feature is robust for illumination condition because it is based on edge information. Other examples of the local feature are Haar-like features [6], EOH features [7], Edgelet features [8]. Wang et al. [9] combine HOG features and Local Binary Pattern (LBP) [10] to detect humans. Linear SVM is used for learning, and good performance is achieved for the INRIA dataset [11], though real-time processing time is not realized.

In the methods using local features, a detection window is scanned on the whole image to extract local features. The human size in an image is unknown in general and thus multiple scans with different window size are required, and consequently, long processing time is required to extract local features in many times. Wu et al. [12] use range information obtained by a stereo camera and restrict the scan region in face detection. Ikemura et al. [13] proposed a method to effectively scan a detection window using the range information obtained by a Time-of-flight (TOF) camera. We take a strategy similar to [12] and [13].

In this paper, we propose a fast and accurate human detection method using a stereo camera. We combine range image segmentation and human detection based on local features. To obtain a range image, we use Subtraction Stereo [14]. The method obtains a range image of only foreground regions. We segment each foreground region using Mean Shift Clustering [15] by improving the method proposed in [16]. Then we apply human detection based on local features to each segment of foreground regions to detect humans. In this process, regions to scan a detection window for extracting local features are restricted, and the range to change window size can be restricted. Therefore, processing time and false detection can be reduced. We adopt Joint HOG features [17] as the local features. When applying the Joint HOG based human detection, we consider occlusion to improve the detection of an occluded human region.

This paper is organized as follows. In section 2, we explain foreground detection using Subtraction Stereo. In section 3, we show a method to segment foreground range image regions. Then we propose a human detection method by combining range image segmentation and human detection based on local features in section 4. In section 5, we show experimental results in indoor and outdoor scenes to evaluate the effectiveness of the proposed method. We conclude this paper in section 6.

## II. Foreground detection

### A. Subtraction Stereo

We use Subtraction Stereo [14] to detect foreground regions in an image. Fig.1 shows the flow of the Subtraction Stereo. It is based on background subtraction and stereo matching. Foreground regions are extracted with a background subtraction method first, and then stereo matching is applied to the extracted foreground regions. Therefore, the processing regions for stereo matching are restricted to the foreground, and we can obtain the range image of foreground regions with less correspondence error and less processing time.

The background subtraction method needs to deal with illumination change and change of background itself. We

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.734

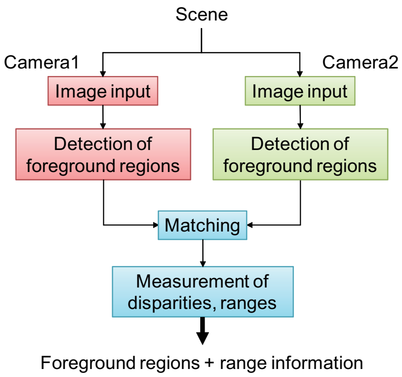
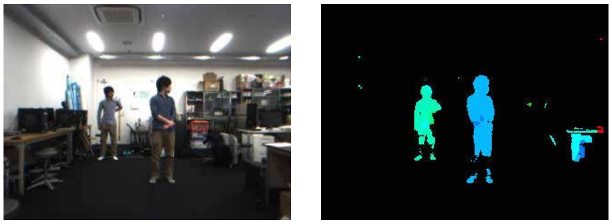

Scene
Camera1
Camera2
Image input
Image input
Detection of foreground regions
Detection of foreground regions
Matching
Measurement of disparities, ranges
Foreground regions + range information

Fig. 1. Flow of Subtraction Stereo [14]



(a) Input image

(b) Disparity image obtained by Subtraction Stereo

Fig. 2. Foreground detection by Subtraction Stereo

implement a non-parametric background update that deals with the histogram of intensities at every pixel. A background image is automatically updated periodically.

An example of the output image by Subtraction Stereo is shown in Fig.2.

### B. *Shadow Detection*

Foreground regions detected by Subtraction Stereo contain shadow regions, which should be discriminated. We detect shadow regions using the method in [18]. The method defines the evaluation function to detect shadow regions based on the changes of intensity and hue in the neighborhood.

Fig.3 shows an example of detected foreground regions using the above foreground detection procedures. Blue regions indicate the detected foreground regions and green regions are detected as shadow regions.

## III. Range Image Segmentation of Foreground Regions

Humans often overlap each other in an image. We project pixels of foreground regions on a ground plane and apply segmentation to the projected foreground regions. To do this, we adopt Mean Shift Clustering.

### A. *Projection of Range Image on a Plane*

Foreground regions obtained by Subtraction Stereo can be separated to objects by a standard labelling technique. However, each labelled object may contain multiple humans because of overlap in an image and should be divided to each human's region. Let us define each object region as $F_i (i = 1, \cdots, n)$, where $n$ is the number of objects. Each pixel of $F_i$ has three-dimensional (3D) information that is measured by a stereo camera. We project each pixel's 3D coordinate on the ground plane. Then we divide the plane by cells with a certain size (5cm×5cm in the following experiments) and make a two-dimensional (2D) histogram of the number of pixels.

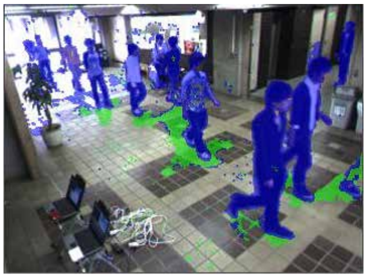


Fig. 3. Detected foreground regions with shadow detection

### B. *Segmentation by Mean Shift Clustering*

Humans are supposed to exist at the peaks of the histogram. We apply Mean Shift Clustering [15] to estimate the number and positions of the peaks automatically. Mean Shift Clustering is a nonparametric technique that does not require prior knowledge of the number of clusters and does not constrain the shape of the clusters [19], [20].

Given the position vector of a cell $c$ in the projection plane as $\mathbf{P_c}$, the mean shift vector $\mathbf{m}(\mathbf{v})$ at the centroid $\mathbf{v}$ is represented as

$$\mathbf{m}(\mathbf{v}) = \frac{\sum_c \mathbf{P_c}\, H_c\, g\Big( \|\frac{\mathbf{v} - \mathbf{P_c}}{\sigma}\|^2 \Big)}{\sum_c H_c\, g\Big( \|\frac{\mathbf{v} - \mathbf{P_c}}{\sigma}\|^2 \Big)} - \mathbf{v} \qquad (1)$$

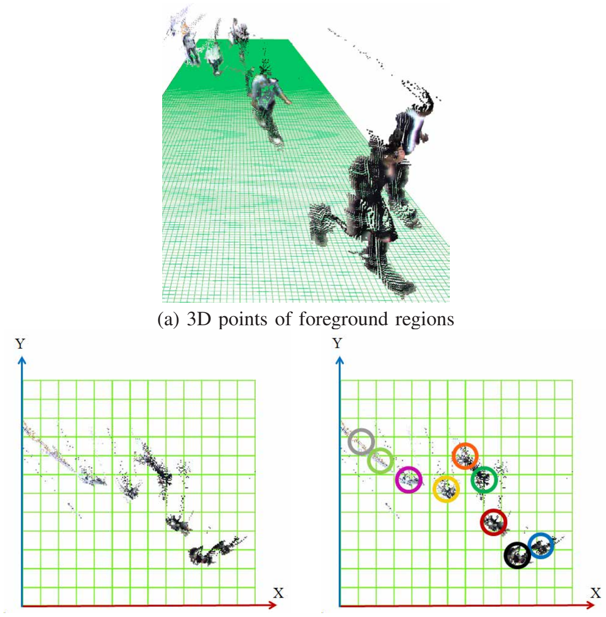
where $H_c$ is the frequency of the histogram at cell $c$, $g$ is the Gaussian kernel, and $\sigma$ is the standard deviation of $g$.

A range image is clustered by the following steps using the mean shift vector $\mathbf{m}(\mathbf{v})$.

1) estimate the initial positions and number of kernels according to the size of projection
2) move each kernel by iterative calculation of (1) and estimate the position of peaks of histograms
3) integrate neighbor kernels and make cells within a certain distance from the centroid a same cluster

Foreground regions are segmented by back-projecting the projected points in the cells to the range image for each cluster. We define each region segmented from a foreground region $F_i$ as $SF_{i,j} (j = 1, \cdots, m_i)$, where $m_i$ represents the number of segments of $F_i$.

Fig.4 illustrates the procedure. Fig.4(a) shows 3D points of foreground regions in Fig.3. Fig.4(b) shows the projection of the 3D points on the 2D plane. Fig.4(c) illustrates the



(a) 3D points of foreground regions

(b) Projection on 2D plane

(c) Mean Shift Clustering

Fig. 4. Range image segmentation

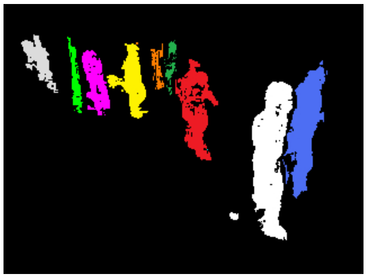


Fig. 5. Segmented regions by range image segmentation

segmentation in 2D by Mean Shift Clustering. The circles represent Gaussian kernels. With the back-projection of 2D points, segmented regions are obtained as shown in Fig.5

## IV. Human Detection

We construct a human detection system based on Joint HOG features. Fast human detection is achieved by combining the range image segmentation results. Occlusion is detected from the segmentation results and human detection at the occluded regions is improved.

### *A. Human Detection Based on Joint HOG Features*

We adopt an object detection method proposed by Mitsui and Fujiyoshi [17]. In this method, joint features combined from multiple HOGs, which are referred to as Joint HOG, are used with two-stage boosting. HOG [5] is one of the most popular and useful low-level features for object detection. HOG features are obtained by calculating gradients. An image region in a detection window is divided into cells and gradient orientation is calculated at each cell. HOG feature consists of the histogram of the gradient orientations.

Joint HOG considers co-occurrence [21] of HOG features at two different cells and has better detection performance especially for symmetric or continuous shapes [17]. Features to represent co-occurrence are calculated for every combination of cells. Effective combinations of cells are selected by the first Real AdaBoost [22] and Joint HOG features are extracted. Then the second Real AdaBoost is applied to select Joint HOG features that are effective for classification, and a strong classifier $H(\mathbf{X})$ is constructed as follows.

$$H(\mathbf{X}) = \sum_{t=1}^{T} h_t(\mathbf{X}) \tag{2}$$

where $\mathbf{X}$ is the selected Joint HOG feature, $T$ is the number of second stage training, and $h_t(\mathbf{X})$ is the weak classifier that is obtained at the first stage training.

### *B. Scan and Integration of Detection Window*

In our method, scan of a detection window is applied to segmented regions $SF_{i,j}$. In this process, we decrease the number of scans by dynamically adjusting the window size using the distance information of each region.

The window size is estimated using inverse proportional relation between the size of human in a 2D image and the distance from the camera to the human. Height $R_h$ and width $R_w$ of a window are estimated using the para-perspective projection model as follows.

$$R_h = \frac{k_h}{W_Y(i,j)}(\cos\theta - y\sin\theta) \tag{3}$$

$$R_w = \frac{k_w}{C_Z(i,j)} \tag{4}$$

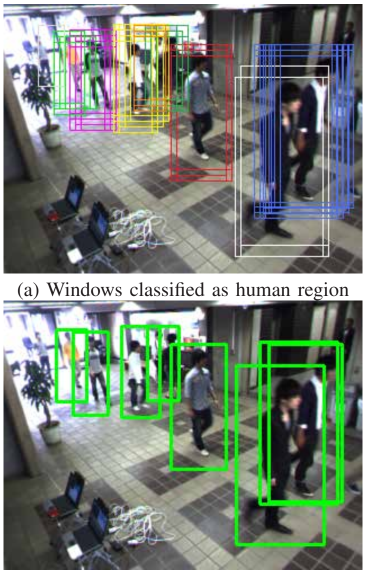
where $W_Y(i,j)$ is the distance from the camera and the centroid of the region $SF_{i,j}$ in the world coordinate system, $C_Z(i,j)$ is the distance from the camera and the centroid of the region $SF_{i,j}$ in the camera coordinate system, $\theta$ is the camera's tilt angle, and $y$ is the normalized vertical image coordinate. Coefficients $k_h$ and $k_w$ can be obtained experimentally.

Fig.6(a) is obtained by applying the proposed scan method to Fig.5. The scan method was applied to each region $SF_{i,j}$ independently. Each window means to be classified as human region. It is shown that error detection from the background, which is inevitable in a standard human detection method, is restrained.

Generally, multiple windows are detected for each human as shown in Fig.6(a). Therefore, we integrate neighboring detection windows. When regions of multiple humans have overlap, windows for the humans tend to be integrated because they are closely located. To avoid the over-integration, we perform the integration for each segmented region $SF_{i,j}$, not for the whole image. That is, only the windows with the same color are integrated in Fig.6(a). Fig.6(b) is the integrated result, which shows that the integration process avoids over-integration and works well.

### *C. Consideration of Occlusion*

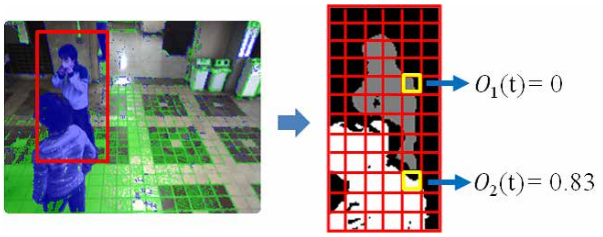
When multiple humans overlap in an image, the back human is occluded. We detect the occlusion as follows. When



(a) Windows classified as human region

(b) Integration of windows

Fig. 6. Detection window



(a) Detection window

(b) Occlusion ratio

Fig. 7. Occlusion detection

multiple regions $SF_{i,j}$ are included in a detection window, we compare the distances to each region $W_Y(i,j)$ and detect occlusion [13]. In Fig.7, the (white) region that is in front of the scan target region (gray) is regarded as occluding region.

Joint HOG features obtain the output of a weak classifier $h_t(\mathbf{X})$ from the features of the combined two cells. We obtain the occlusion ratio (the ratio of white region in Fig.7(b)) for each cell. We modify the weak classifier by using the obtained occlusion ratio of each cell and define the final classifier $H'(\mathbf{X})$ as follows.

$$H'(\mathbf{X}) = \sum_{t=1}^{T} \{h_t(\mathbf{X}) \cdot (1 - O_1(t)) \cdot (1 - O_2(t))\} \quad (5)$$

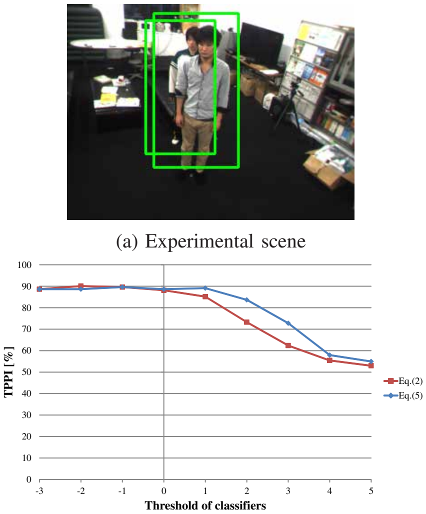
where $O_1(t)$, $O_2(t)$ represent the occlusion ratio of two cells used in $h_t(\mathbf{X})$. As occlusion ratio becomes larger, the output of corresponding weak classifier becomes smaller and so the output of the classifier at the occluded region can be restrained.

By thresholding the output of the final classifier $H'(\mathbf{X})$, a target region is classified to a human or not.

(a) Positive

(b) Negative

Fig. 8. Some samples used for training



(a) Experimental scene

(b) Comparison of classifiers

Fig. 9. Evaluation of classifiers for occlusion

## V. Experiments

### A. Experimental Conditions

We used NICTA Pedestrian Dataset [23] for training of the classifier. The numbers of positive and negative samples were 7,892 and 30,000 respectively. Fig.8 shows the examples of samples. We trained 10 and 300 times for the first and second stages respectively to construct a classifier. We used a Point Grey Research Bumblebee2 stereo camera with the image size of $320 \times 240$. A PC with an Intel Core2 Duo CPU (3.06 GHz) was used.

TPrate (True Positive rate) and FPrate (False Positive rate) in the experimental results represent the correctly and wrongly detected rates respectively.

### B. Evaluation of Classifier Considering Occlusion

We compared the classifiers (2) and (5) to evaluate the consideration of occlusion. We tested them for 400 scenes with occlusion as shown in Fig.9(a). Fig.9(b) shows the classification results with different thresholds for the classifiers. Table I shows the classification results for the minimum threshold without false detection. It can be said from Fig.9(b) and Table I that performance of detection is increased by the consideration of occlusion.

### C. Evaluation of the Proposed Method

To evaluate the proposed method, we compared it with the method using Joint HOG features only [17], in which a

TABLE I. COMPARISON WITH AND WITHOUT CONSIDERATION OF OCCLUSION

| Classifier | TPrate [%] | FPrate [%] |
|---|---|---|
| H'(x) (Eq.(5)) | 89.1 | 0.0 |
| H(x) (Eq.(2)) | 71.3 | 0.0 |

TABLE II. EVALUATION RESULT IN AN INDOOR SCENE

| Method | TPrate [%] | FPrate [%] |
|---|---|---|
| Proposed | 83.1 | 3.1 |
| Reference | 63.5 | 65.9 |

TABLE III. EVALUATION RESULT IN AN OUTDOOR SCENE

| Method | TPrate [%] | FPrate [%] |
|---|---|---|
| Proposed | 68.8 | 1.1 |
| Reference | 48.8 | 68.5 |

detection window is scanned on the whole image. We mention the method as "reference method." We used scenes that are different from the training data set as the test data. We made experiments in an indoor scene with complicated background and an outdoor scene with large illumination variation. The number of frames were 2,000 in each scene. The threshold was set empirically for a different indoor scene with simple background and the value was used for every scene.

Examples of human detection and evaluation results are given in Fig.10 and Table II for the indoor scene and Fig.11 and Table III for the outdoor scene. In each figure, (a) shows the results by the proposed method and (b) by the reference method.

It can be said from Tables II, III that the detection rate of the reference method is not high for scenes with complicated backgrounds. The reason is that when human-like shape such as the cross at the center of Fig.10(b) appears in the background, the region is detected falsely. On the contrary, the proposed method succeeds to decrease the false detection by restricting the regions to detect humans. In addition, humans who are not detected by the reference method as seen in Figs.10(b), and 11(b) are detected by the proposed method as seen in Figs.10(a), and 11(a). This is because occlusion is considered as described in IV-C.

In the outdoor scene, the proposed method could detect humans under difficult illumination or background condition such as in the shadow as shown in Fig.11(a). However, the proposed method sometimes failed to detect humans when they are too close to the camera and the detection window could not cover the whole body.

### *D. Processing time*

Table IV shows the processing time of each process. It is shown that the processing speed of the proposed method is much faster than the reference method because of the restriction of the regions to detect humans. The proposed method works at about 11fps and thus online human detection is possible.

TABLE IV. PROCESSING TIME

| Process | Proposed method [ms] | Reference method [ms] |
|---|---|---|
| Capture | 16.3 | 16.3 |
| Background subtraction | 0.4 | - |
| Stereo matching | 18.8 | - |
| Shadow detection | 6.7 | - |
| Segmentation | 11.5 | - |
| Joint HOG | 30.4 | 502.2 |
| Others | 4.0 | - |
| Total | 88.1 | 518.5 |

## VI. CONCLUSION

In this paper, we have proposed a human detection method that combines range image segmentation and an object detection method based on local features. First, foreground regions are extracted using Subtraction Stereo, and then each region is segmented using Mean Shift Clustering. A classifier using Joint HOG features [17] is applied to each segment and humans are detected. Human detection with high detection accuracy and fast processing speed is achieved by combining the two methods and restricting the regions scanned by human detection windows. The size of the detection window is obtained using the distance information of a range image and camera parameters, which also contributes to reduce the processing time. Furthermore, occlusion is considered in constructing a classifier and in integration of detection windows, which improves the detection performance for the occluded humans. Experiments showed that the proposed method has better accuracy and speed than the method using Joint HOG features only. 11fps fast human detection is achieved.

In the future, we intend to introduce detection of human parts and combine it with whole body detection to construct a classifier that is robust to the change of human appearance.

## ACKNOWLEDGMENT

This study was partly supported by CREST, JST.

## REFERENCES

[1] Y. Feund, M. Jones, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Computational Learning Theory, Eurocolt, pp.11-20, 1995.

[2] B. E. Boser, I. M.Guyon, V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," Proc. of 5th Annual Workshop on Computational Learning Theory , pp.144-152, 1992.

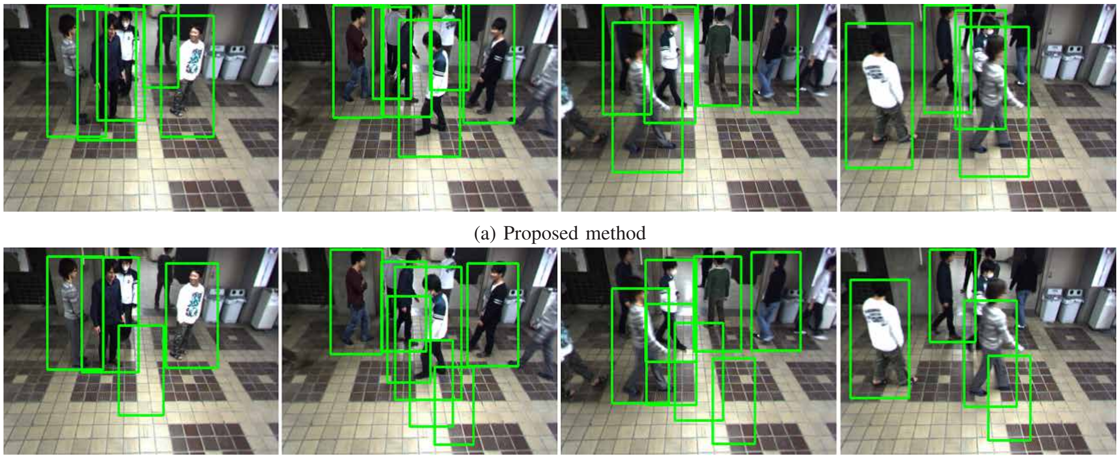
[3] C. Papageorgiou, et al., "A Trainable System for Object Detection," Int. J. of Computer Vision, vol.38, no.1, pp.15-33, 2000.

[4] P. Viola, et al., "Detecting Pedestrians Using Patterns of Motion and Appearance," Proc. of IEEE Int. Conf. on Computer Vision, pp.734-741, 2003.

[5] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CA, USA, pp.886-893, 2005.

[6] P. Viola, J. Jones, "Rapid object detection using a boosted cascade of simple features," Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp.511-518, 2001.

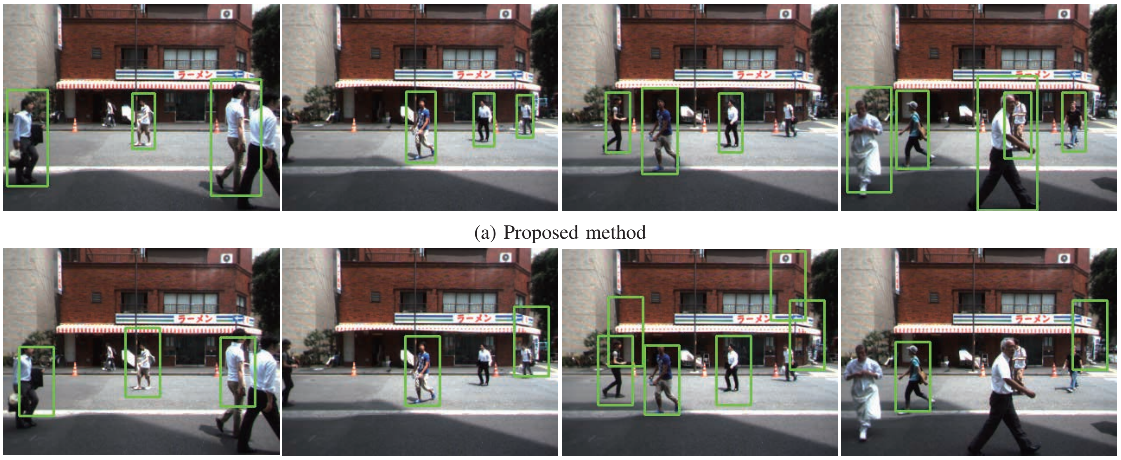
[7] K. Levi, Y. Weiss, "Learning object detection from a small number of example: The importance of good feature," Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Vol.2, pp.53-60, 2004.



(a) Proposed method

(b) Reference method

Fig. 10. Examples of detection in an indoor scene with complicated background



(a) Proposed method

(b) Reference method

Fig. 11. Examples of detection in an outdoor scene with large illumination variation

[8] B. Wu, R. Nevatia, "Detection of multiple, partially occluded human in a single image by bayesian combination of edgelet part detectors," Proc. of IEEE Int. Conf. on Computer Vision, vol.1, pp.90-97, 2005.

[9] X. Wang, T. X. Han, S. Yan, "An HOG-LBP Human Detector with Partial Occlusion Handling," Proc. of IEEE Int. Conf. on Computer Vision, pp. 32-39, 2009.

[10] T. Ahonen, A. Hadid, and M. Pietikinen, "Face description with local binary patterns: Application to face recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.28, no.12, pp.2037-2041, 2006.

[11] INRIA Person Dataset, http://pascal.inrialpes.fr/data/human/

[12] H. Wu, K. Suzuki, T. Wada, and Q. Chen, "Accelerating Face Detection by Using Depth Information," Proc. of the 3rd Pacific Rim Symp. on Advances in Image and Video Technology, pp.657–667, 2009.

[13] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features," Proc. of the 10th Asian Conf. on Computer Vision, Vol.IV, pp.25-38, 2010.

[14] K. Umeda, et al., "Subtraction Stereo -A Stereo Camera System That Focuses On Moving Regions-," Proc. of SPIE-IS & T Electronic Imaging, Vol.7239 Three-Dimensional Imaging Metrology, 723908, 2009.

[15] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.17, no.8, pp.790-799, 1995.

[16] T. Ubukata, K. Terabayashi, A. Moro, K. Umeda, "Multi-Object Segmentation in a Projection Plane Using Subtraction Stereo," Proc. of 20th Int. Conf. on Pattern Recognition (ICPR2010), pp.3296-3299, 2010.

[17] T. Mitsui and H. Fujiyoshi, "Object detection by joint features based on two-stage boosting," Proc. of IEEE 12th Int. Conf. on Computer Vision Workshops (ICCV Workshops), pp.1169 - 1176, 2009.

[18] A. Moro, et al., "Auto-adaptive threshold and shadow detection approaches for pedestrians detection," Proc. of AWSVCI, pp.9-12, 2009.

[19] D. Comaniciu, V. Ramesh, P. Meer, "Real-time tracking of non-rigid objects using mean shift," Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp.142-149, 2000.

[20] A. Yilmaz, "Object Tracking by Asymmetric Kernel Mean Shift with Automatic Scale and Orientation Selection," Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp.1-6, 2007.

[21] T. Mita, T. Kaneko, B. Stenger, O. Hori, "Discriminative Feature Co-occurrence Selection for Object Detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.30, no.7, pp.1257-1269, 2008.

[22] R. E. Schapire and Y. Singer, "Improved Boosting Algorithm Using Confidence-rated Predictions," Machine Learning, No.37, pp.297-336, 1999.

[23] G. Overett, L. Petersson, N. Brewer, L. Andersson, N. Pettersson, "A new pedestrian dataset for supervised learning," Proc. of IEEE Intelligent Vehicle Symposium, pp.373-378, 2008.