DPM Score Regressor for Detecting Occluded Humans from Depth Images

Tsuyoshi Usami*, Hiroshi Fukui†, Yuji Yamauchi‡, Takayoshi Yamashita§ and Hironobu Fujiyoshi¶
*Email: usami915@vision.cs.chubu.ac.jp
†Email: fhiro@vision.cs.chubu.ac.jp
‡Email: yuu@vision.cs.chubu.ac.jp
§Email: yamashita@cs.chubu.ac.jp
¶Email: hf@cs.chubu.ac.jp
Chubu University, Kasugai-shi, Aichi, Japan

***Abstract*—A part-based object detection method called deformable part models (DPMs) is known as a robust method for detecting objects that have detection method for posture variation. In the detection stage, the DPMs assume that all parts are visible. If some parts of people are partially occluded by an object such as a table or wall, detecting them becomes difficult. This paper proposes a robust method for detecting occluded humans to regress scores with a reduced influence of occlusion. We apply 3D raster scanning to depth images for finding occluded regions. We compute occlusion rates in each part of humans as the occlusion rate from regions. We regress from the DPM detection scores, DPM root scores, DPM part scores, and occlusion rates as explanatory variables to enable detecting the scores of humans with no occlusion. Using these detection scores makes detecting humans easier. Experimental results show that the precision of the proposed method was improved by almost 20% compared with that of conventional DPMs.**

I. Introduction

Dalal *et al.* have proposed a method combining histograms of oriented gradients (HOG) features and a support vector machine (SVM)[1]. This method has been applied to object detection in other categories as well as to human detection, and is widely used. HOG features are local ones that focus on the luminance gradient. There is a characteristic that absorbs fluctuations in lighting and local position. Therefore, various method of HOG-based object detection have been proposed to achive high accuracy[2][3][4]. Among them, Felzenszwalb *et al.* proposed deformable part models (DPMs)[5]. DPMs are a part-based approach. This method captures appearance features that include not only a person's whole body but also such parts as the hands and feet. DPMs are highly accurate for human detection because they obtain robustness against posture variation by learning the positional relationship between the respective parts. Learning the models of DPMs assumes that all of the parts of all can be observed. However, a problem occurs when parts of the human body are occluded, making detection by DPMs difficult.

Methods of adjusting the identification according to the occlusion region by determining the occlusion area have been proposed[4][6][7]. Wang *et al.* proposed a method to switch the part detector that is applied according to the occlusion region by clustering the regions[4]. Enzweiler *et al.* proposed a method of assigning weights using the occlusion rates for the part-based classifier by detecting the occlusion region from the distance information and movement information[6]. Ikemura *et al.* proposed a method to assign weights against the weak classifiers of Real AdaBoost using the occlusion rates obtained from extractions of the occlusion area[7]. However, these methods that assign weights for the scores of the discriminator may not be detected correctly due to decreases in the detection score that occur when the occlusion rate is high.

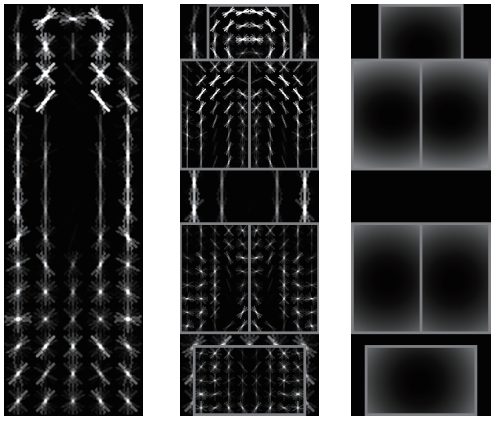

Fig. 1. Visualization of the DPMsl

In this paper, we propose a score calculation method to reduce the effect of occlusion using support vector regression (SVR)[8] with occlusion rates and DPM scores as explanatory variables. In this method, we get the occlusion rates and DPM scores from depth images. Thus, even if occlusion occurs in many areas of a person, an output close to the original score can still be obtained.

II. Deformable Part Models [5]

DPMs are an object detection method that is part-based and that corresponds to posture variation. In this chapter, we describe the DPM discriminant function and the problem of occlusion.

A. The discriminant function in DPMs

As shown in Figure 1, the DPM approach is composed of a root filter, part filter, and spatial model. As shown in Figure 1(a), the root filter portrays the appearance features of the human body. As shown in Figure 1(b), the part filter captures the appearance features that enable discrimination of humans, such as the head and legs. As shown in Figure 1(c), the spatial model represents the positional relationship between the parts. The deformation cost provided by the spatial model increases if parts move from the reference positions.

Detection scores of DPMs with n parts are obtained from the identification function shown in (1) by using these three kinds of models.

$$score(p_0) = F_0' \cdot \phi(H, p_0) + \sum_{i=1}^{n} \max_{x_i, y_i} (F_i' \cdot \phi(H, p_i) - d_i \cdot \quad \phi_d(dx_i, dy_i)) + b \qquad (1)$$

The first term is the score of the root filter, the second term is the score of the part filter, and the third term is bias. F_i' $(i = 0, ..., n)$ is the weight vector of each filter, and $\phi(H, p_i)$ $(i = 0, ..., n)$ is the feature vector of the detection window corresponding to each filter. Inner product $F_i' \cdot \phi(H, p_i)$ $(i = 0, ..., n)$ of these two vectors is the score in each filter. i is the root filter when $i = 0$ and the part filter when $i > 0$. (dx_i, dy_i) is the amount of movement from the reference position of each part, as shown in (2).

$$(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i) \qquad (2)$$

In addition, (dx_i, dy_i) is a quadratic function representing the movement direction and amount of the part movement, as shown in (3).

$$\phi_d(dx_i, dy_i) = (dx_i, dy_i, dx_i^2, dy_i^2) \qquad (3)$$

$\phi_d(dx_i, dy_i)$ is used in the calculation of the deformation cost$d_i \cdot \phi_d(dx_i, dy_i)$.

B. Problem due to occlusion

DPMs learn a model on the assumption that all the parts can be observed. For that reason, if the occlusion occurs in part of the human body, the detection becomes difficult. Therefore, we investigated the effect of occlusion in DPMs. Figure 2(a) shows a position of the root filter and part filter detected by DPMs. When performing detection against Figure 2(a), the detection score of the DPMs is 1.519. Then, the detection score of the DPMs when artificial occlusion occurs as shown in Figure 2(b), is -0.361. Table I shows each part scores, including these detection scores. As shown in Table I, the score of the root and parts 4-6 where occlusion has occurred greatly decreases, and it can be seen that the detection score is low. Thus, the detection score of DPMs significantly decreases when occlusion occurs in the part area.

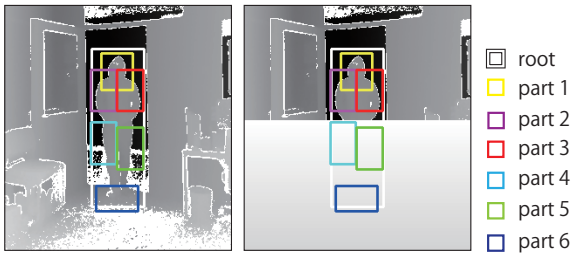

Fig. 2. Generated of the pseudo occlusion

III. PROPOSED METHOD

Herein, we describe the proposed method for score calculation using regression with DPM scores and occlusion rates. Figure 3 shows the flow of the method. It learns the DPM approach, and weights and biases of SVR. In the discrimination, we get explanatory variables from the input depth image and obtain scores that decrease the effect of occlusion using SVR.

A. Learning of DPMs and SVR

In the learning of the proposed method, we obtain the model of DPMs and weights and biases of SVR.

1) Learning of DPMs: Because of the simultaneous need to learn the weight vector of each filter F_i' $(i = 0, ..., n)$, a four-dimensional vector that defines the deformation cost d_i $(i = 1, ..., n)$ and bias b, DPMs learn by using a latent support vector machine (LSVM). The LSVM obtains β that minimizes the objective function $L_{D(Z)}(\beta)$. (4) defines objective function $L_{D(Z)}(\beta)$.

$$L_{D(Z)}(\beta) = \frac{1}{2} \parallel \beta \parallel^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i)) \qquad (4)$$

In (4), the first term is optimized, and the second term is the loss function. As shown in (5), β is a set of parameters obtained by learning.

$$\beta \quad = \quad (F_0', ..., F_n', d_1, ..., d_n, b) \qquad (5)$$

$D(Z)$ is a set of learning samples (x_i, y_i).

2) Learning of SVR: In the method, the response variables are detection scores with no occurrence of occlusion, and the explanatory variables are each of the scores and the occlusion rates. Therefore, the learning of SVR needs two scores: one where occlusion occurred and one where it did not. Thus, as shown in Figure 4, we generate a learning sample by applying an artificial occlusion. The explanatory variables are composed of the detection scores, root scores, part scores, and occlusion

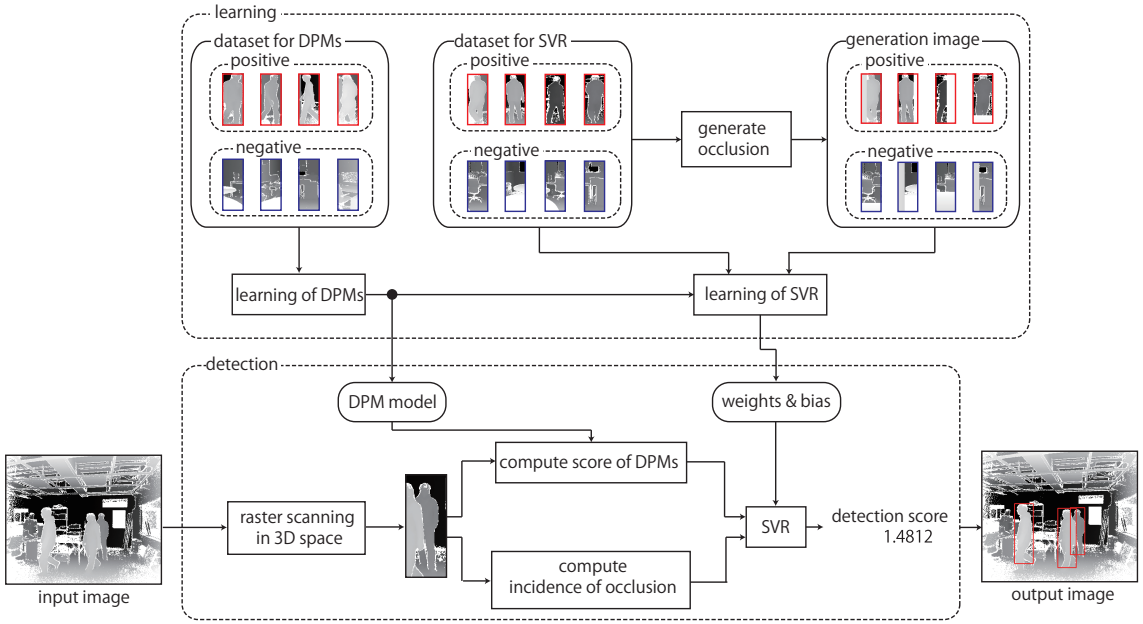

Fig. 3. Flow of the proposed method

TABLE I
CHANGES IN THE SCORE OF DPMS DUE TO OCCURRENCE OF OCCLUSION

	detection score	root	part1 (head)	part2 (right shoulder)	part3 (left shoulder)	part4 (right foot)	part5 (left foot)	part6 (bottom foot)
no occlusion	1.519	1.535	0.984	0.494	0.343	0.265	0.313	0.340
occlusion	-0.361	0.651	0.984	0.494	0.343	-0.018	-0.025	-0.057

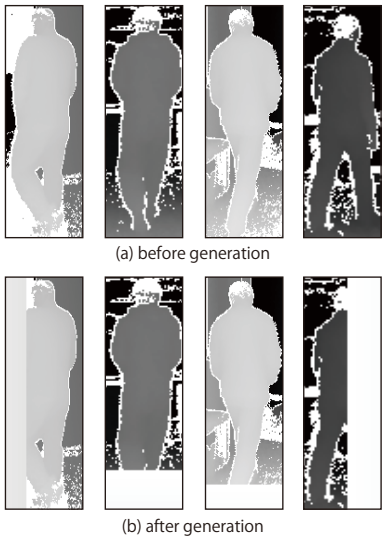

Fig. 4. Images where occlusion occured

rates. (6) shows the explanatory variables in the case of DPMs with 6 parts.

$$\begin{aligned}\mathbf{x} = (&score(p_0), F_0' \cdot \phi(H, p_0),\\ &F_1' \cdot \phi(H, p_1) - d_1 \cdot \phi_d(dx_1, \quad dy_1), ...,\\ &F_6' \cdot \phi(H, p_6) - d_6 \cdot \phi_d(dx_6, \quad dy_6),\\ &O(p_1) * (F_1' \cdot \phi(H, p_1) - d_1 \quad \cdot \phi_d(dx_1, dy_1)), ...,\\ &O(p_6) * (F_6' \cdot \phi(H, p_6) - d_6 \quad \cdot \phi_d(dx_6, dy_6)))\end{aligned} \tag{6}$$

p_i is a variable that contains the coordinates (x_i, y_i) and size $(width_i, height_i)$ of the upper left corner of the parts filter. We obtain the occlusion rate $O(p_i)$ of i parts using the (7).

$$O(p_i) = \frac{\sum_{k=y_i}^{y_i+height_i} \sum_{l=x_i}^{x_i+width_i} \alpha(k,l)}{width_i * height_i} \tag{7}$$

$\alpha(k,l)$ is a function that represents the presence or absence of occlusion occurring on the coordinate (k,l). The output of $\alpha(k,l)$ is 1 when occlusion occurs and 0 when it does not.

The learning of the SVR is to determine the weights and biases for the objective function so that it is minimized. (8) shows the expression for minimizing the objective function [9].

$$\arg\min_{\mathbf{w},b} C \sum_{i=1}^{N} E(t_i - f(\mathbf{x}_i)) + \frac{1}{2}|\mathbf{w}|^2 \tag{8}$$

$E(t_i - f(\mathbf{x}_i))$ is an error function, as shown in (9).

$$E(\alpha) = \begin{cases} 0 & (\alpha \leq th) \\ \alpha - th & (\alpha > th) \end{cases} \tag{9}$$

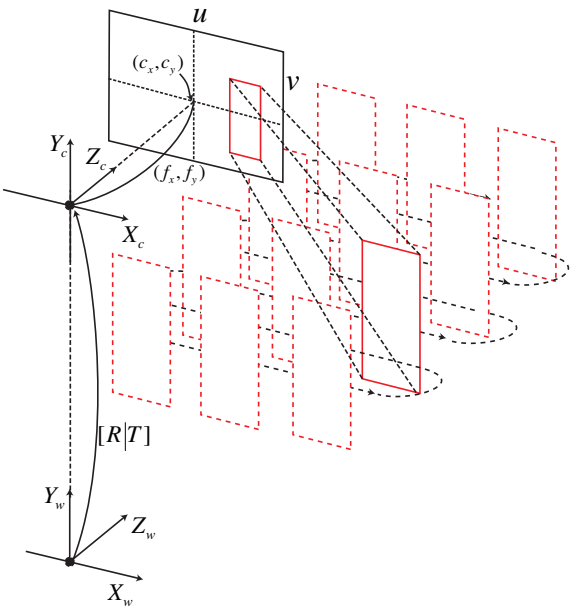

Fig. 5. 3D raster scanning

th is a threshold value of the allowable error.

B. 3D raster scanning[7]

In this study, we ran 3D raster scanning because of the use of depth images. As shown in Figure 5, detection windows were placed in 3D space and scanned along the floor ($Y_w = 0$). 3D raster scanning performs efficient detection and does not work against the agonistic position. Herewith, improvement in the detection accuracy can be expected.

In 3D raster scanning, the detection window placed on a three-dimensional space needs to be projected onto an image. (10) shows a conversion equation from world coordinate (X_w, Y_w, Z_w) to local coordinate (u, v).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{A}[\mathbf{R}|\mathbf{T}] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{10}$$

Extrinsic parameter $[\mathbf{R}|\mathbf{T}]$ is composed of rotation matrix $\mathbf{R}$ and translation matrix $\mathbf{T}$. In this study, the position and height of the camera are fixed. The position of the camera is fixed so that the world coordinate and camera coordinate are parallel. Therefore, the rotation matrix is a unit vector. In addition, for the origin of the world coordinate and the floor surface of the camera position, the translation matrix is $[0, -\text{height of camera}, 0]^T$. Thus, the extrinsic parameter of the camera in this study is (11).

$$[\mathbf{R}|\mathbf{T}] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1.4 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{11}$$

Intrinsic parameter $\mathbf{A}$ is composed of the focal length represented in pixels (f_x, f_y) and the center coordinate of the camera (c_x, c_y), as shown in (12).

$$\mathbf{A} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

(13) shows substituting the (11) and (12) with (10).

$$\begin{cases} u = \dfrac{X_w}{Z_w} f_x + c_x \\ v = \dfrac{Y_w - 1.4}{Z_w} f_y + c_y \end{cases} \tag{13}$$

We conduct perspective projection conversions from the detection window set at any of the world coordinates to image coordinates by (13).

C. The judgment of occlusion and calculation of the occlusion rate

If occlusion occurs in the detection object, obstacles exist in front of the detection target. Therefore, if the value of the target pixel is less than the distance of the detection window, we can determine obstacles exist. The expression of performing occlusion judgment of target pixel (k, l) is the following.

$$\alpha(k,l) = \begin{cases} 1 & (Z_w - z(k,l)) > t \\ 0 & (Z_w - z(k,l)) \leq t \end{cases} \tag{14}$$

$z(k, l)$ is the distance value of coordinate (k, l), and Z_w is a distance value from the camera to the detection window. t is a threshold, and it is 30 cm in this study. The occlusion rate is obtained using the same expression as when learning.

D. Regression by SVR at the time of detection

Explanatory variables to be input into the SVR are 14-dimensional parameters, as shown in (6). The occlusion rate is obtained from each part detection window in the 3D raster scanning. The score reduced effect of occlusion is determined by the identification function $f(\mathbf{x})$ of (15).

$$f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w} + b \tag{15}$$

$\mathbf{w}$ is the weight vector of the SVR, and b is the bias term. From the aforementioned, we obtain the score reduced effect of occlusion using regression.

IV. EVALUATION EXPERIMENT

We conducted evaluation experiments to compare DPMs and proposed method. We evaluated the effectiveness of the proposed method in two experiments using image generated pseudo occlusion and actual images.

A. Evaluation of robustness to occlusion

We evaluated the performance against occlusion of the proposed method and the conventional one.

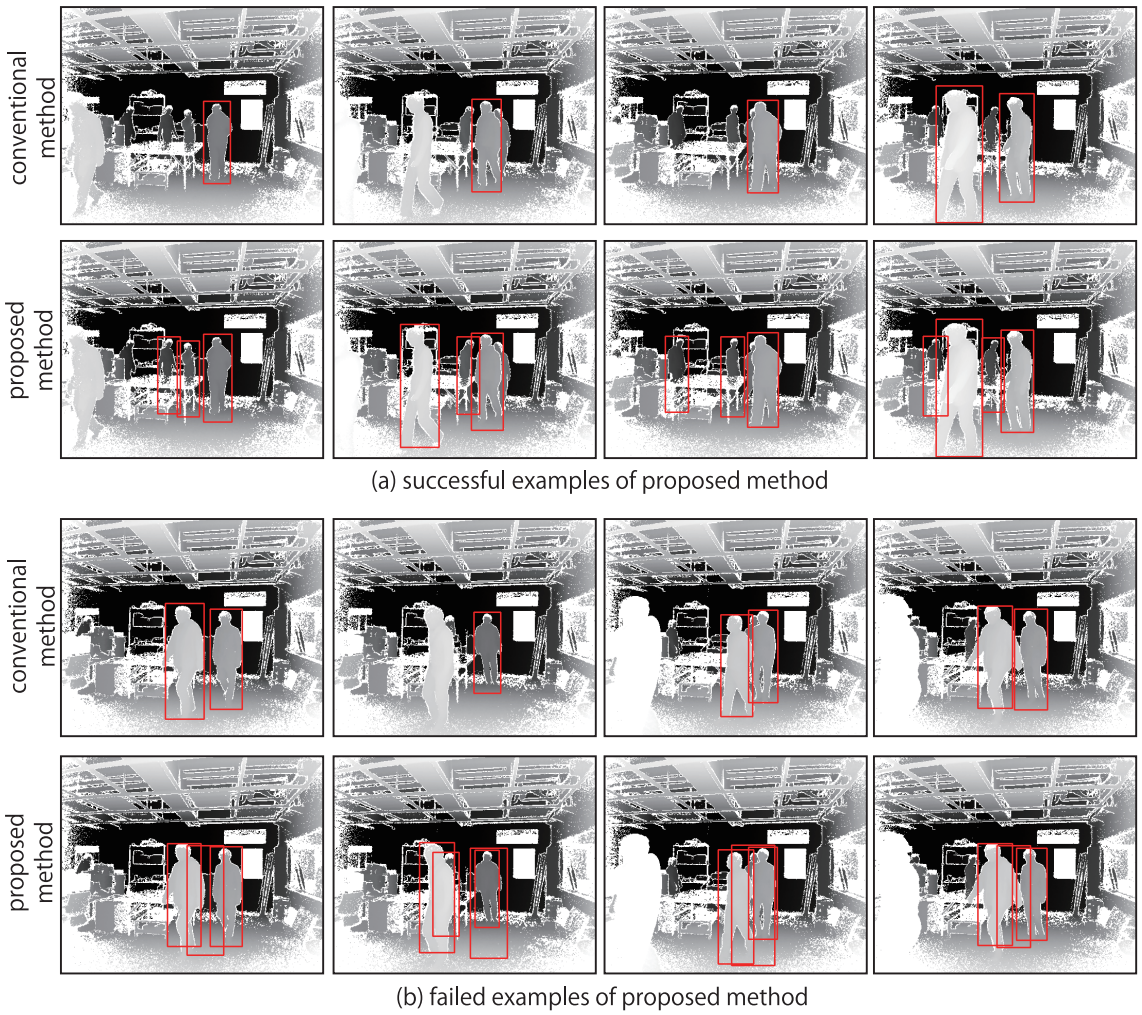

Fig. 7. Detection example of conventional method and proposed method

1) Overview of the experiment: In this experiment, we evaluated the detection rates when the occlusion rates changed. The threshold used for detection was a value obtained by learning the DPMs. Image-generated pseudo occlusion and depth images taken with Kinect V2 were used in the experiment. We used 905 positive samples and 1008 negative samples for DPM learning. Positive samples for DPM learning were depth images obtained by cutting the human area where occlusion did not occur. We used 1209 positive samples that generated pseudo occlusion and 1300 negative samples for SVR learning. We used the generation pattern and frequency of occlusion from [10] to generate of occlusion. We used 800 positive samples with generated pseudo occlusion for the evaluation.

2) Results of the experiment: Figure 7 shows the comparison results of the detection rate for each occlusion rate. The two methods do not have a big difference in the detection rate until 10% occlusion. The proposed method obtained a 10% higher detection rate than the conventional method at more than 20% occlusion.

B. Evaluation of detection performance due to actual images

We evaluated in the actual images where occlusion occurred.

1) Overview of the experiment: In this experiment, we compared the detection rates and false detection rates using an evaluation dataset containing images of people where occlusion occurred. The threshold used for detection was a value obtained by learning the DPMs.

The DPMs and SVR for learning used a dataset of Section IV-A1. We used 649 depth images obtained by KinectV2 in the evaluation dataset. There was a 1088 human area in this evaluation dataset.

2) Results of the experiment: We evaluated the real images where occlusion occurred, as shown in Figure II. The false detection rate of the proposed method was a little higher than that of conventional method, but the detection rate of the proposed method was almost 20% higher than that of conventional method.

Figure 6 shows examples of the conventional method's and the proposed method's detection.

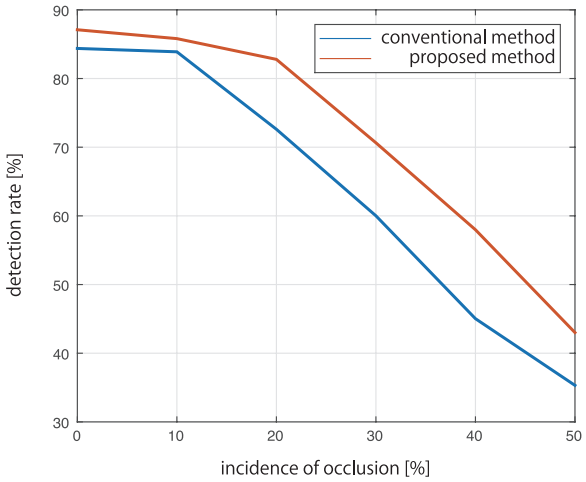

Fig. 6. Comparison of detection rate with increasing occlusion area

TABLE II
PERFORMANCE COMPARISON WITH THE OCCLUSION AREA

	detection rate[%]	false detection rate[%]
conventional method	46.67	0.31
proposed method	65.62	4.47

As shown in Figure 6(a), the proposed method can detect a person that cannot be detected by the conventional one. However, as shown in Figure 6(b), false detection that does not occur in the conventional method occurs in the proposed method. The scores of the target that are false positive are higher than the threshold, but they are lower than the scores of the target that are true positive. Therefore, we can decrease the occurrence false detections by changing the threshold.

V. CONCLUSION

This paper proposed a score calculation method that reduces the effects of occlusion using regression. The method can detect occluded humans using not only DPM scores but also occlusion rates in explanatory variables. In future research, we plan to enable achieving higher accuracy of detection performance by reviewing the learning samples and parameters.

REFERENCES

[1] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, CVPR, vol. 1, pp. 886
UTF2013893, 2005.
[2] A. Bosch, A. Zisserman and X. Munoz, “Representing shape with a spatial pyramid kernel”, ICIVR, 2007.
[3] P. Ott and M. Everingham, “Implicit color segmentation features for pedestrian and object detection”, ICCV, 2007.
[4] X. Wang, H. X. Tomy and Y. Shuicheng .“An HOG-LBP human detecor with partial occlusion handling”, ICCV, pp. 32
UTF201339, 2009.
[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, “Object detection with discriminatively trained part based models”, PAMI, vol. 32, no. 9, pp. 1627
UTF20131645, 2010.
[6] M. Enzweiler, A. Eigenstetter, B. Schiele and D. M. Gavrila, “Multi-cue pedestrian classification with partial occlusion handling”, CVPR, pp. 990
UTF2013997, 2010.
[7] S. Ikemura and H. Fujiyoshi, “Real-time human detection using relational depth similarity features”, ACCV, pp. 25
UTF201338, 2011.
[8] B. Debasish, P. Srimanta and P. D. Candra, “Support vector regression”, NIP, vol. 11, no. 10, pp. 201
UTF2013224, 2007
[9] C. Chih-Chung and L. Chih-Jen, “A library for support vector machines”, ACM TIST, vol. 2, no. 2, pp. 27:1–27:27, 2011.
[10] P. Dollar, C. Wojek, B. Schiele and P. Perona, “Pedestrian detection: an evaluation of the state of the art.” PAMI, vol. 34, no. 4, pp. 743
UTF2013761, 2012.