Occlusion Handling Human Detection with Refocused Images

Hirokatsu Kataoka, Shuhei Ohki, Kenji Iwata and Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan.
Email: hirokatsu.kataoka@aist.go.jp

***Abstract*—The paper presents *a novel robust human detection method based on camera array system* to broaden the application range for human detection. Currently, even by using a deep neural network (DNN), it is difficult to detect a hardly occluded human. In the camera array system, we consider how to distinctly show a human occluded by an environmental condition. The generated *refocused images* by the camera array system allow us to remove the effect of the noises. Although refocused images have not been utilized in conventional human detection, we believe that the refocused images are beneficial for improving the detection performance, especially in severe conditions. To execute the experiments, we have collected Refocused Human DataBase (RHDB) with the camera array system. By using HOG+SVM with a monocular camera (at an almost random rate of 54.8%), the refocused images made the +10.1% improvement (64.9%) by noticeably showing a human. The combined representation of refocused images and AlexNet achieved 94.6% on the RHDB. Moreover, our final model recorded 98.0% with an attention-layer and fine-tuned parameters. The result is much higher than the traditional human detection with HOG+SVM and a monocular camera (+43.2%).**

I. Introduction

In robotics research, active control for collision avoidance between humans and robots has been studied for a long time. Controlling technologies are rapidly developing due to the recent growth of autonomous driving. However, an advanced occlusion handling method, which finds humans in fields of agricultural crops, is required in addition to typical fields such as traffic roads and indoor scenes. To avoid collisions, we need to develop human detection for more complicated scenes.

In the paper, we tackle a very challenging issue, which is detection of hardly occluded humans in fields containing crops or heavy brush. We focus on the crops in agricultural fields. Figure 1 shows a hard positive and extremely hard positive agricultural scene in comparison with an easy positive scene. Especially in hardly occluded situations, conventional approaches such as the histogram of oriented gradients (HOG) [1] and its improved models [2], [3] might not capture an effective feature in an agricultural field because those methods rely on contours and edge shapes to detect an object.

The convolutional neural network (CNN) proposed by LeCun [4] and its improved models [5], [6] perform automatic training for an appropriate representation depending on the characteristics of the image samples in a database. The model construction by a CNN has achieved outstanding performances in image recognition tasks [7]. In a pre-experiment, we tried to execute a human detection task, but a simple CNN-based approach failed to train an effective feature in hardly occluded agricultural scenes. We believe that another system or device must be considered in order to capture improved images, because a simple CNN setting is limited under extremely hard positive conditions.

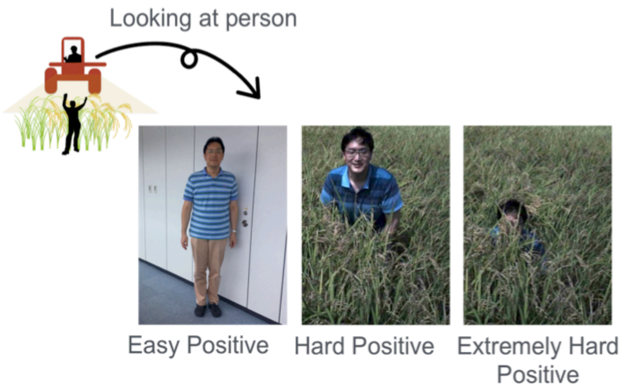


Fig. 1. Comparison of conventional and agricultural scenes. The left figure shows a conventional detection scene. In the two images on the right, the safety combine harvester system shows detection of a human in an agricultural scene. We have to detect a human at any time in not only the conventional easy positive scene but also in the hard and extremely hard positive scenes.

Our solution is to apply a *camera array system* that arranges multiple cameras in parallel (see Figure 2). The camera array system can exclude the effects of various edges that are obstructed. We use the multi-camera geometry to refocus on a specific human and then synthesize the refocused images. The operation allows us to greatly improve the image quality (namely signal-to-noise ratio) as the result of occlusion removal.

We employ the refocused images from a camera array system as training and inference images. In our experiment, we improved the results from 54.8% (monocular camera + HOG + support vector machine (SVM)) to 94.6% (camera array system + CNN). For the final model with an attention-layer and fine-tuned parameters, we achieved 98.0% on a self-collected database. Our rate of occluded human detection almost surpassed the human-level rate.

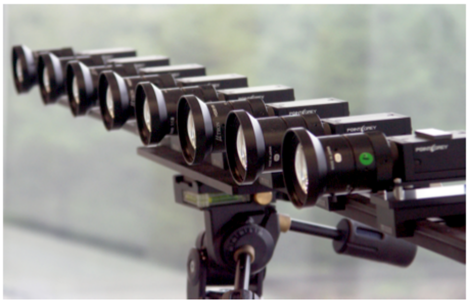


Fig. 2. Camera array system

II. Related work

A. Human detection

Some representative databases for human detection include the INRIA person [1] and Caltech pedestrian [8], [9] datasets. Dalal *et al.* [1] released the INRIA person dataset. Moreover, they simultaneously proposed the epoch-making algorithm, HOG, as a gradient descriptor. Undoubtedly, their work including the descriptor and database has accelerated the human detection field. Dollar *et al.* pursued the problem of human detection by using the Caltech pedestrian dataset [8], [9]. Their detailed analysis was beneficial for improving the descriptors, classifier, and model, and for removing several difficulties regarding analysis of the benchmark.

Benenson *et al.* [10] considered over 40 approaches and compared the results. Recent research has revealed that a detector with CNN [11] is better than conventional approaches such as random forests [10] and the deformable part model (DPM) [12]. The performance of automatic feature learning outperformed hand-tuning algorithms for a large number of data. Zhang *et al.* [13] improved the annotation of the Caltech pedestrian dataset, but they claimed that a more sophisticated annotation was required to improve detection. According to the results in deep learning, we must consider the dataset quality in addition to the detection model.

This paper examines human detection in an agricultural field. Unlike human detection in a traffic road, we must consider the hard occlusion problem. Existing approaches in human detection partially solve the hard occlusion problem with data-driven parameter optimization. We directly give a solution for hard occlusion by applying a photography-based camera array system. At the same time, we propose a human database that is captured by the camera array system.

B. Camera array

Computational photography, which is used for high-performance image processing, contains refocusing technique for changing the focal length by postprocessing photos [14]. Computational photography makes it possible to record the light distribution in the real world, whereas a conventional camera cannot record much light information. Two types of camera systems can record a light field. One is a camera with arranged lens elements, and the other is a multi-camera array. In multi-camera arrays, commercially available cameras can be arranged so that fields of view overlap each other to some degree. Such a feature is considered to be advantageous for detecting humans in an occluded environment. By calculations based on the image data taken by multi-camera arrays from a plurality of viewpoints, a light field can be obtained and the focus can be changed after taking enough photos to produce an accurate 3D model of a subject.

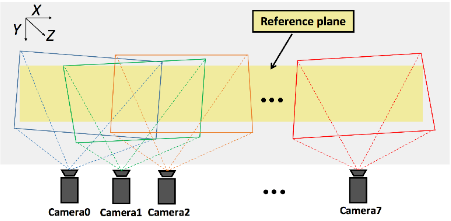


Fig. 3. Image rectification of camera array system

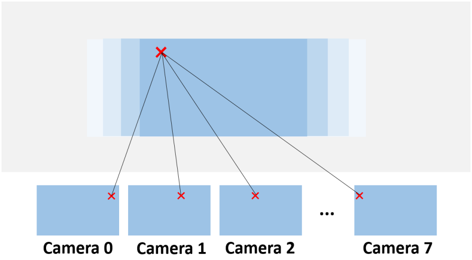


Fig. 4. Synthesis of refocused images

In this study, a multi-camera array consisting of eight conventional cameras, as shown in Figure 2, was used to improve the image system for input into the CNN.

III. Camera-array system

Eight cameras of the same specification are arranged parallel to each other on a rail, as shown in Figure 2, but these cameras have variations in their physical arrangement. To synthesize the refocused images from the camera array system, first, internal parameters and external parameters of each camera are obtained by calibration. The internal parameters, such as the focal length, are fixed parameters for each camera, and the external parameters are parameters dependent on the mounting position and the angle of the cameras. From these parameters, a transformation matrix of the rectified image is obtained. Next, as shown in Figure 3, the rectified images of each camera are mapped to the reference plane by using the transformation matrix. Finally, it is possible to easily synthesize refocused images by adjusting the amount of shift in the x direction according to the virtual focal length, as shown in Figure 4.

IV. Human detection

Our human detection method is mainly composed of a camera array system and a feature descriptor. The improved camera sensor shown in Section III allows us to effectively

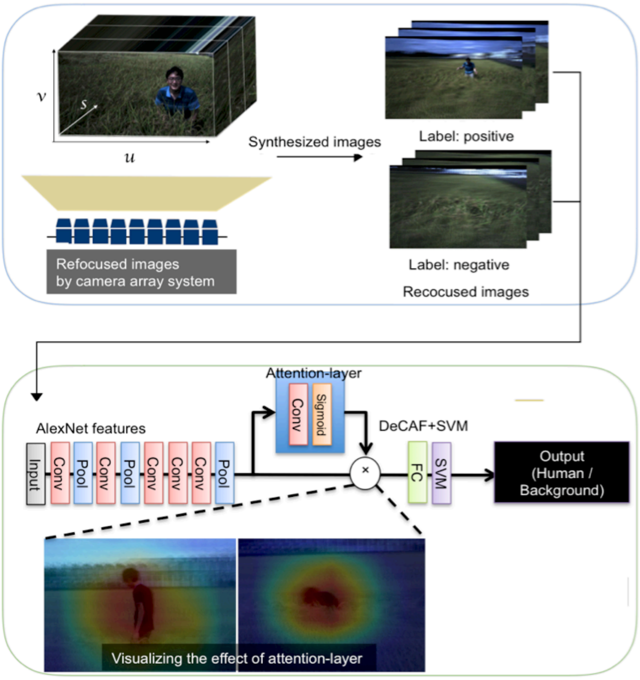


Fig. 5. Flowchart of human detection

make a good image to see a human. To detect a human in the image, we capture several features, such as deeply learned features.

Figure 5 shows the flowchart of our human detection. The two-step human detection method, namely (i) camera array system and (ii) feature description, is proposed to detect a hardly occluded human. Intuitively, by combining the camera array system and the deep features, we can produce a good image to see a human in the first step. Then, we can train a well-organized feature from an area of a convmap's attention in the second step. The process flow is different from conventional human detection in terms of refocused images; therefore, we show the method in detail.

Input image with camera array system. The camera array system enables us to directly input a 3D light field. The input has enough flexibility to grab a feature regardless of the distance between a human and an object, which is a combine harvester in our example. The detailed camera array system is shown in Section III.

Model architecture for feature description. We try to employ several deep features. We mainly evaluate DeCAF [**?**] and end-to-end CNN models [5][1]. In DeCAF, we extract features from the intermediate layer on a CNN. We use AlexNet [5], which consists of 8 layers, and extract the features from the first fully connected (fc) layer. Using these features, we classify the images by a SVM. The AlexNet model is pre-trained by the ImageNet dataset [7].

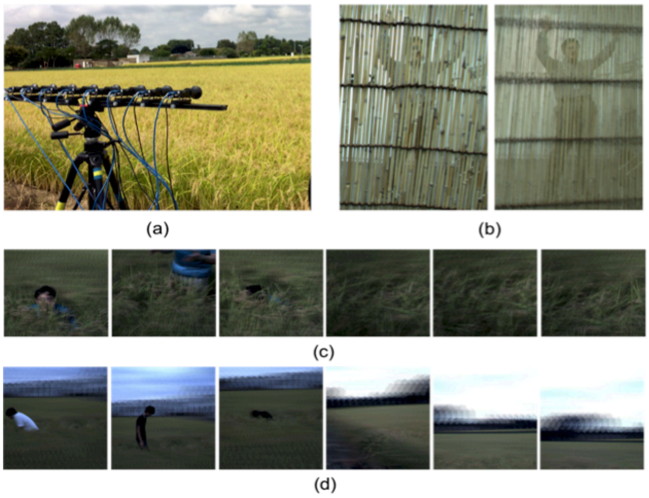


Fig. 6. Dataset creation: (a) Example position of camera array system (b) PRESET: experimental setup to confirm effectiveness of refocus (c) SET1: captured agricultural field 1 (d) SET2: captured agricultural field 2.

Attention-layer for convolutional networks. To construct a well-organized model for setting human detection, we focus on the feature extraction around the area of the human(s). Because we set binarized detection, namely, a human is in the image or not, we assign an attention-layer to effectively extract a human's feature. As shown in Figure 5 and 8, the attention-layer is constructed of the additional convolutional and sigmoid layers after the last convolutional layer in AlexNet. The output of the attention-layer is defined as $A_c = \Sigma_{x,y} a_c(x,y)$, which is multiplied at each channel output as follows:

$$M_c^{'}(x,y) \quad = \quad \Sigma_c a_c(x,y) M_c(x,y) \qquad (1)$$

where M and $M^{'}$ are the convolutional maps and attentioned convolutional maps with c channels.

V. Refocused Human DataBase (RHDB) with Camera Array System

In this section, we describe our dataset creation and training procedure.

Dataset creation. To conduct an experiment of human detection, we set our camera array system on the side of an objective field, as shown in Figure 6(a). We took images in a laboratory setup: PRESET (Figure 6(b)) and 2 different agricultural fields corresponding to SET1 (Figure 6(c)) and SET2 (Figure 6(d)). The movement of the camera was allowed to move only in the vertical and horizontal directions, not the back and forth directions, and the distance between the human and the camera was varied at approximately 2.0–10.0 meters.

In this way, we created the RHDB with the camera array system. The numbers of positive and negative images are shown in Table I.

At the beginning, to verify the effect of the camera array system, we created a PRESET data taken by the camera array and a single camera. The PRESET contains a human captured

[1] Figure 5 shows DeCAF, which classifies a human or a background by a support vector machine (SVM).

TABLE I

CONFIGURATION OF RHDB: NUMBER OF IMAGES, INPUT IMAGE SIZE AND CHANNEL USAGE IN EACH SET

SET	#Positive without/with data aug.	#Negative without/with data aug.	Image size (pixel)	Channel
PRESET	1,000/10,000	1,000/10,000	450x350	RGB
SET1	4,769/47,690	4,769/47,690	700x700	RGB
SET2	9,705/97,050	4,977/49,770	700x700	RGB

in a laboratory environment which has a shelter between the human and camera(s). We set the cameras stably in the laboratory environment.

Then we captured objects such as an autonomous combine harvester in an agricultural field of rice plants as a practical application. Both SET1 and SET2 contain hard occluded situations. To assume the movement of the combine harvester, we varied the viewpoints and movements of the camera array. Multiple humans took varying postures through the datasets.

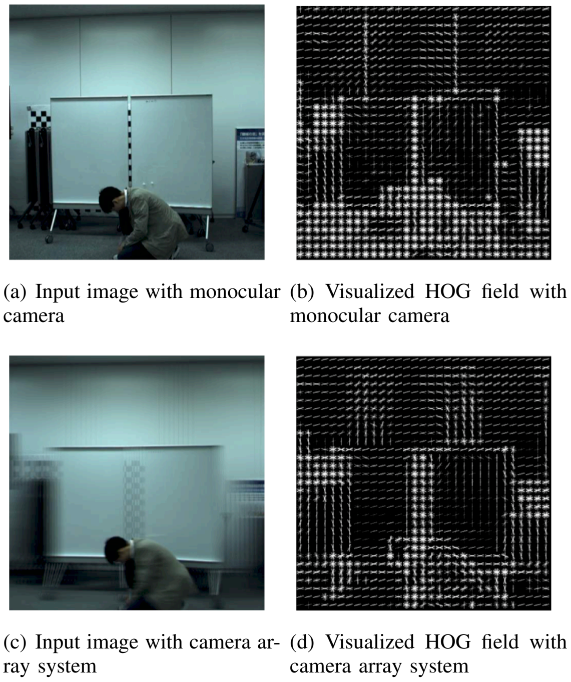
Training setup. We split training and testing samples by considering different viewpoints and time zones. This is why we avoid detection of any dataset biases (e.g., training only background without understanding pedestrian's feature). The amount of images is 3/4 for the training set and 1/4 for the testing set in all datasets. Spatial jittering and flipping are applied for data augmentation. The number of images is increased x10 with the data augmentation. In AlexNet, the input was 224 pixels × 224 pixels × 3 channels. The initial learning rate was set to 0.001, and updating was set to a factor of 0.1 per 10 epochs; thus, learning was completed after 30 epochs. We assigned a high dropout ratio in each of the fully connected (fc) layers. We set both the first and second fc layers to 0.8.

VI. EXPERIMENT

We carried out several experiments on the Refocused Human DataBase (RHDB) by dividing PRESET (laboratory setup), SET1 and SET2 (two different agricultural fields). We describe the detailed experiments step by step:

A. *Evaluating the effect of refocus on RHDB*

Starting from the HOG descriptor, we evaluate with and without the camera array system to create refocused images on RHDB-PRESET, RHDB-SET1 and RHDB-SET2 (see Table II). Generally, we can confirm the improvements in human detection by using the refocused images. We set two types of block sizes–32x32 [pixel] and 64x64 [pixel] for the HOG. The two settings have 29,241 and 5,184 dimensions, respectively. As shown in Table II, the HOG result is improved +3.6% (HOG (32)) and +6.7% (HOG (64)) with the refocused images on RHDB-PRESET. At the same time, the result of the refocused images increased +10.1% (HOG (32)) and +10.1% (HOG (64)) on RHDB-SET1. In the experiment, we showed the impact of the refocused images with the camera array system. Especially on RHDB-SET1, the refocused images highly improved the accuracies from both settings of the HOG descriptor with an almost chance rate (= 50%).



(a) Input image with monocular camera (b) Visualized HOG field with monocular camera

(c) Input image with camera array system (d) Visualized HOG field with camera array system

Fig. 7. Visualizations of monocular camera and camera array system in visualized edge spaces.

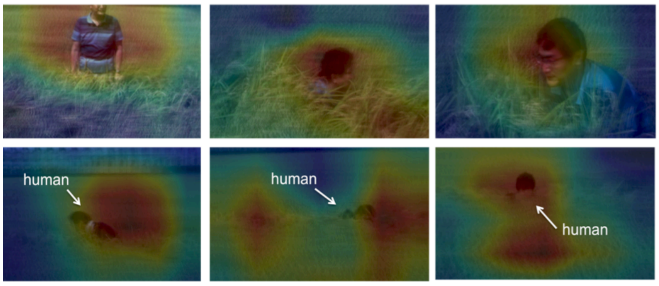


Fig. 8. Visualizing the effect of attention layer

Figure 7 shows the visualized edge spaces with and without refocus by the camera array system. Although we cannot confirm the effect of the camera array in the input images (Figure 7(a) and (c)), improvement of the edge spaces is clearly seen. In Figure 7(d), the human is (relatively) easy to distinguish from the background. On the contrary, HOG with a monocular camera (see Figure 7(b)) is confusing and depends on the floor's texture.

TABLE II
EFFECTS OF REFOCUSED IMAGES: A COMPARISON BETWEEN MONOCULAR CAMERA AND CAMERA ARRAY SYSTEM FOR INPUT IMAGES

	Refocus / Monocular (%)	Refocus / Monocular (%)	Refocus / Monocular (%)	#dim
	@PRESET	@SET1	@SET2	
HOG (64)	**86.9** / 80.2 (+6.7)	**71.34** / 57.33 (+14.01)	**62.4** / 52.3 (+10.1)	5,184
HOG (32)	**94.5** / 90.9 (+3.6)	**72.00** / 61.28 (+10.72)	**64.9** / 54.8 (+10.1)	29,241
AlexNet	**100** / **100** (+0.0)	**90.80** / 87.58 (+3.22)	**94.6** / 91.2 (+3.4)	4,096

B. Evaluating the effect of DNN

We apply DNN representation to improve the accuracies on the refocused database. In the experiment we employ AlexNet and activation features of the 6-th fully connected layer as DeCAF. DeCAF is combined with the SVM to classify a human (positive) or a background (negative). Table II shows the accuracies of DeCAF with AlexNet. We also confirm the improvement with the refocused images and the increase in comparison with the HOG descriptor. Because of the simple background and easy positive in RHDB-PRESET, AlexNet-DeCAF recorded 100% on both settings, i.e., with and without refocused images. Although we cannot check the effect of refocused images with AlexNet-DeCAF on RHDB-PRESET, it is sufficient to see the performance of the DNN. We can see the improvement of refocused images on RHDB-SET1, which increased +3.4% from the previous highly accurate rate.

For practical use, we explore the well-organized tuning for DNN with refocused images in the following subsections.

C. Exploration study for DNN + refocused images

Table III shows the comparisons of DeCAF+SVM versus the end-to-end model, and the number of fully connected units. We use RHDB-SET1 and RHDB-SET2 to confirm a more practical experimental setting. As shown above, the activated feature is referred by DeCAF [**?**].

In the number of activations in fully connected (fc) layers, we varied the connected units as {16, 32, 64, 128, 256, 512, 1024, 2048, 4096}. From the experimental results, around 64 units seem to be a better way to understand a human in refocused images. Experimentally, the number is fit to the problem, which is 2-category classification between a human and a background.

By comparing the end-to-end model and DeCAF, DeCAF is better in most of the cases. The DeCAF improved +5.65% (64-dim: 91.86% and default 4,096-dim: 86.21%) on RHDB-SET1 and +5.80% (64-dim: 91.25% and default 4,096-dim: 97.05%).

D. Final results with attention layer in AlexNet

Table IV shows our final model with the attention layer in AlexNet. We evaluated the scores of precision, recall and F-score in addition to accuracy in the experiment. Although AlexNet without the attention layer (DeCAF) performed with the best accuracy (%) on RHDB-SET1, AlexNet with the attention layer improved in other evaluations. Figure 8 visualizes the effect of the attention layer with class-activated mapping (CAM) [16]. The human class is visualized in both figures. We can confirm the human-specified activation to be not only hard positive (Figure 8 left) but also extremely hard positive (Figure 8 right). The improvement with the attention layer is shown, even though the baseline with AlexNet (DeCAF) is high enough.

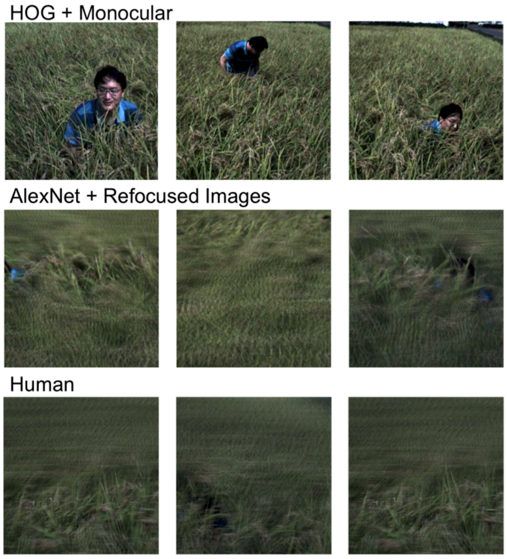


Fig. 9. Examples of false negative frames for (top) monocular camera & HOG, (middle) refocused images & AlexNet, and (bottom) human-level performance.

Moreover, we evaluated human-level performance, as shown in Table IV. We did the same training and testing as the computer did on RHDB-SET1 and RHDB-SET2. According to the bottom of Figure 9, a human tester only fooled the extremely hard positives which are doubted images whether an annotator can put a positive label. Surprisingly, in terms of performance coefficients, our proposed method (best rate) surpassed the average of human testers in precision, recall and f-score, and partially outperformed the human average in accuracy.

E. Qualitative assessment

In this subsection we show human detection examples. Figure 9 (top) shows false negative samples by using HOG. Figure 9 (middle) shows false negative frames with AlexNet. In these frames, we hardly find a human. These frames are the so-called hard positive samples; however, the annotation is defined as “human is in the image or not”. Then we must put a positive label in the examples because the images include

TABLE III
EXPLORATION STUDY FOR ACTIVATION UNITS IN ALEXNET

Activations in fc layers	AlexNet (End-to-End) @SET1	AlexNet (End-to-End) @SET2	AlexNet (DeCAF) @SET1	AlexNet (DeCAF) @SET2
32	81.43	93.65	89.06	95.61
64	88.26	**96.73**	**91.86**	**97.05**
128	**89.43**	96.51	87.77	87.00
256	86.33	84.91	86.71	90.33
512	87.30	80.96	87.34	82.78
1024	87.13	87.93	86.16	83.93
2048	86.58	87.52	85.28	84.69
4096	86.67	91.50	86.21	91.25

TABLE IV
COMPARISON OF ATTENTION-BASED APPROACHES AND HUMAN-LEVEL PERFORMANCES

	Accuracy (%) @SET1 / @SET2	Precision @SET1 / @SET2	Recall @SET1 / @SET2	F-score @SET1 / @SET2
AlexNet (DeCAF; ours)	**91.86** / 97.05	93.01 / 97.09	91.87 / 97.06	91.82 / 97.07
AlexNet w/ attention layer (End-to-End; ours)	91.69 / 97.46	**96.35 / 98.76**	**95.73 / 98.73**	**95.83 / 98.74**
AlexNet w/ attention layer (DeCAF; ours)	91.69 / **98.03**	91.34 / 98.12	89.52 / 98.04	89.41 / 98.05
Human Average	**94.93 / 97.86**	**95.08 / 97.90**	**94.93 / 97.87**	**94.93 / 97.87**

a part of a human's body. The annotation problem should be discussed as well as the DNN architecture. Conversely, the detection performance of AlexNet is very high. The model architecture only makes a mistake if an input image is (extremely) hard positive. On the contrary, Figure 9 shows false negative frames by HOG. The simple hand-crafted method cannot detect a human from such apparent positive samples under the hard occluded condition.

VII. CONCLUSION

To cope with the problem of detecting a hardly occluded human, we propose a method of using refocused images taken by a camera array system. We synthesized images obtained by the camera array system by refocusing the human under an occluded environment in order to reduce the occlusion. We also proposed an attention layer, which specifies a human's area to obtain better representation for human detection. From the HOG + monocular camera (at an almost random accuracy rate of 54.8%), the refocused images made a +10.1% improvement (64.9%). The combined representation with the refocused images and AlexNet achieves 94.6% on the self-collected Refocused Human DataBase (RHDB). Finally, our proposed model achieved 98.0% with the attention layer and fine-tuned parameters. This is much higher than de-facto-standard human detection with HOG+SVM and a monocular camera (+43.2%).

In the future, the feature representation must be improved. Although the current model inputs refocused images, we verified the joint optimization of refocused and feature parameters from multi-view images. As described above, by making both the camera array system and the learning characteristics to be more appropriate, we can enhance the accuracy of human detection.

REFERENCES

[1] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection" in CVPR, 2005.

[2] T. Watanabe, S. Ito, and K. Yokoi,"Co-occurrence histograms of oriented gradients for pedestrian detection," Information Processing Society of Japan (IPSJ) Transactions on Computer Vision and Applica-tions, Vol.2, pp.39-47 (2010).

[3] H. Kataoka, K. Tamura, K. Iwata, Y. Satoh, Y. Matsui, Y. Aoki, "Extended Feature Descriptor and Vehicle Motion Model with Tracking-by-detection for Pedestrian Active Safety," IEICE Transactions on Information and Systems, Vol.E97-D, No.2, 2014.

[4] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, 1998.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," NIPS, 2012.

[6] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recoginition," in ICLR, 2015.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," in IJCV, 2015.

[8] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian Detection: A Benchmark", in CVPR, 2009.

[9] P. Dollar, C. Wojek, B. Schiele, P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," in TPAMI, vol.34, no.4, pp.743-761, 2012.

[10] R. Benenson, M. Omran, J. Hosang, B. Shiele, "Ten years of pedestrian detection, what have we learned?", in ECCVW, 2014.

[11] P. Sermanet, K. Kavukcuoglu, S. Chintala, Y. LeCun, "Pedestrian Detection with Unsupervised Multi-Stage Feature Learning," in CVPR, 2013.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," in TPAMI, 2010.

[13] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, "How Far are We from Solving Pedestrian Detection?", in CVPR, 2016.

[14] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High Performance Imaging Using Large Camera Arrays," in SIGGRAPH 2005.

[15] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, "Learning Deep Features for Discriminative Localization," in CVPR, 2016.