INTEGRATION OF GENERATIVE LEARNING AND MULTIPLE POSE CLASSIFIERS FOR PEDESTRIAN DETECTION

Hidefumi Yoshida[1], Daisuke Deguchi[1], Ichiro Ide[1], Hiroshi Murase[1],
Kunihiro Goto[2], Yoshikatsu Kimura[2] and Takashi Naito[2]
[1]*Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601 Japan*
[2]*Toyota Central Research & Development Laboratories, Inc., Nagakute, Aichi, 480-1192, Japan*

Keywords: Pedestrian Detection, Generative Learning, HOG, SVM.

Abstract: Recently, pedestrian detection from in-vehicle camera images is becoming an important technology in ITS (Intelligent Transportation System). However, it is difficult to detect pedestrians stably due to the variety of their poses and their backgrounds. To tackle this problem, we propose a method to detect various pedestrians from in-vehicle camera images by using multiple classifiers corresponding to various pedestrian pose classes. Since pedestrians' pose varies widely, it is difficult to construct a single classifier that can detect pedestrians with various poses stably. Therefore, this paper constructs multiple classifiers optimized for variously posed pedestrians by classifying pedestrian images into multiple pose classes. Also, to reduce the bias and the cost for preparing numerous pedestrian images for each pose class for learning, the proposed method employs a generative learning method. Finally, the proposed method constructs multiple classifiers by using the synthesized pedestrian images. Experimental results showed that the detection accuracy of the proposed method outperformed comparative methods, and we confirmed that the proposed method could detect variously posed pedestrians stably.

1 INTRODUCTION

Recently, many research groups have proposed methods to detect pedestrians from an in-vehicle camera image for driving assistance. The most successful methods to detect pedestrians are methods that employ Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) (Dalal and Triggs, 2005; Enzweiler et al., 2009). Since the HOG is robust against lighting condition changes and local geometric changes, and the SVM classifier has a high generalization ability, this combination is now widely used for detecting objects from images for various applications. However, this method requires numerous pedestrian images for training the classifier. Then, gathering various samples comprehensively is not feasible and its cost is quite expensive. In addition, since pedestrians' pose varies widely, it is difficult to detect various pedestrians by using a single classifier.

To overcome these problems, this paper proposes a method to detect variously posed pedestrians by using multiple classifiers optimized for each pedestrians' pose. Although each classifier needs to be trained by numerous pedestrian images corresponding to each pose, it is very difficult to gather various appearances and also time-consuming to prepare these images. Therefore, the proposed method reduces the bias and the cost for preparing these images by introducing a "generative learning" method. Here, generative learning is a method to train a classifier by synthesizing various training samples. This method was successfully applied in several applications, such as generic objects detection (Murase, 1996), traffic sign detection (Doman et al., 2009), pavement marker detection (Noda et al., 2009), and pedestrian detection (Enzweiler and Gavrila, 2008). The generative learning method synthesizes various images by modeling appearances of target objects in actual conditions. Thus, we can control the appearances of them. Although this method enables us to synthesize various images without manual intervention, the quality of the synthesized images is highly dependent on the generation model. Therefore, as used in (Enzweiler and Gavrila, 2008), this paper employs Statistical Shape Models (SSM, (Cootes et al., 1995)) to synthesize variously posed pedestrian images. The main contributions of this paper are:

Yoshida H., Deguchi D., Ide I., Murase H., Goto K., Kimura Y. and Naito T..
INTEGRATION OF GENERATIVE LEARNING AND MULTIPLE POSE CLASSIFIERS FOR PEDESTRIAN DETECTION.
DOI: 10.5220/0003817305670572
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 567-572
ISBN: 978-989-8565-03-7

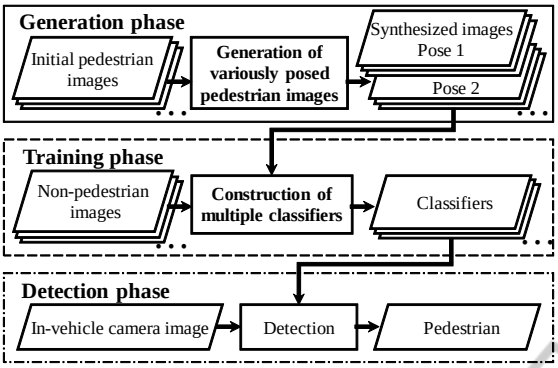


Figure 1: Process flow of the proposed method.

1. Generation of numerous pedestrian images labeled with their poses from only a small number of them for training.
2. Construction of multiple classifiers optimized for each pedestrians' pose.

This paper is organized as follows. In section 2, we describe the process flow of the proposed method and explain the procedures for the synthesis of pedestrian images and the construction of multiple classifiers. Section 3 describes an experiment using in-vehicle camera images. Finally, we conclude this paper in section 4.

2 METHOD

Figure 1 shows the process flow of the proposed method. As seen in Fig. 1, the proposed method consists of three phases; (1) the generation phase, (2) the training phase, and (3) the detection phase.

In the generation phase, inputs are only a small number of pedestrian images, but numerous pedestrian images are synthesized from them. Here, the proposed method employs SSM as a generation model for obtaining variously posed pedestrian images with various textures. This phase is divided into the shape generation, the texture generation, and the background synthesis steps.

Next, the proposed method constructs multiple classifiers in the training phase. Multiple classifiers consist of a classifier optimized for each pedestrians' pose which is trained by using pedestrian images synthesized in the previous phase.

The last is the detection phase that detects pedestrians from in-vehicle camera images by using the trained multiple classifiers. In this phase, outputs of multiple classifiers are combined and used for the final judgment of the pedestrian detection.

The following sections explain details of each phase.

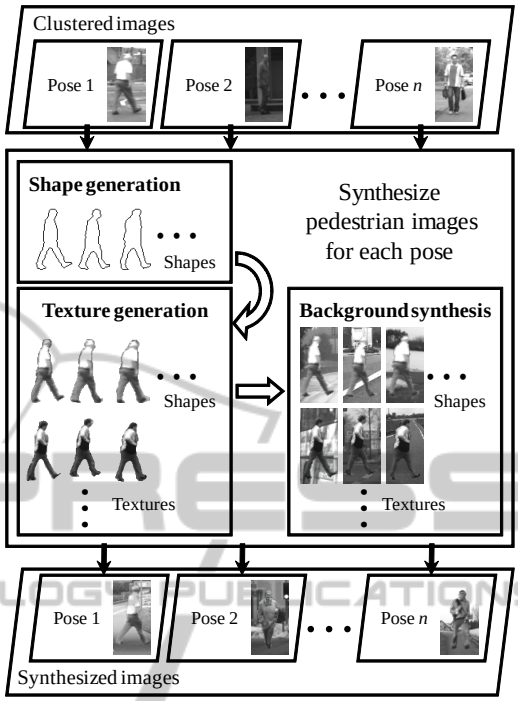
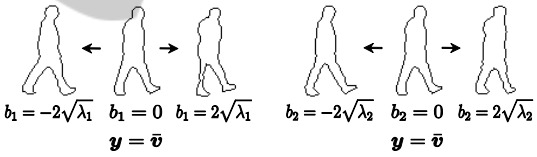


Figure 2: Overview of the generation phase.



(a) $s = 1$ (b) $s = 2$

Figure 3: Examples of the synthesized pedestrian shapes by using SSM. These shapes are synthesized by changing the weight b_s. The images (a) and (b) represent the synthesized shapes by using a different principal component s. They satisfy the condition $b_i = 0 \ \ (i \neq s)$. Images placed at the center of each figure correspond to the mean shape $\bar{\mathbf{v}}$. The left and the right images in each figure correspond to the synthesized shape $\mathbf{y}$ using Eq.(1).

2.1 Generation Phase

As seen in Fig. 2, the proposed method synthesizes variously posed pedestrian images with various textures from the initial pedestrian images classified into a pose class. To synthesize various pedestrian images corresponding to each pose class, the proposed method employs the framework proposed in (Enzweiler and Gavrila, 2008). Inputs of this phase are a small number of pedestrian images classified into each pose class. This is done by extracting the contours of pedestrians from the input images, and then

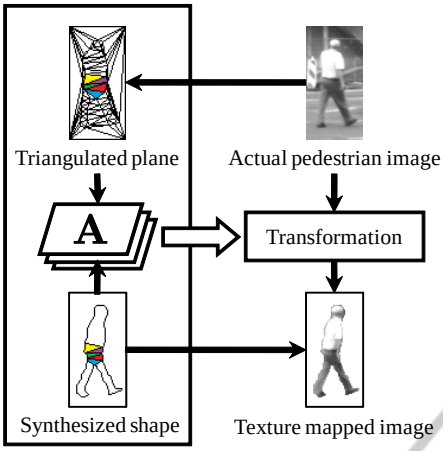


Figure 4: Texture mapping from an actual pedestrian image to a synthesized shape. Here, **A** is an affine transformation matrix.

by clustering them according to the distance between the extracted contours (Gavrila and Giebel, 2001). Finally, the proposed method considers each cluster as a pedestrian "pose", and uses them in the following process.

2.1.1 Shape and Texture Generation

The proposed method synthesizes various pedestrian shapes by using a Statistical Shape Model (SSM) (Cootes et al., 1995), as shown in Fig. 3. This generation process is applied to each pose class.

In the SSM, the synthesized shape **y** can be represented as

$$\mathbf{y} = \bar{\mathbf{v}} + \mathbf{Pb}, \quad (1)$$

where $\bar{\mathbf{v}}$ is the mean vector corresponding to the shape of each posed pedestrian, and **Pb** represents the shape perturbation. Matrix **P** consists of eigenvectors obtained by applying PCA to pedestrian shapes in each pose class, and these eigenvectors are selected by evaluating eigenvalues so that the cumulative contribution ratio of eigenvalues exceeds 99%.

Textures of pedestrians are synthesized by applying a procedure similar to that in the shape generation step. In this step, luminance values within a pedestrian region are represented by **v**. First, the proposed method applies the Delaunay triangulation algorithm to the control points placed at the contour of a pedestrian, and then obtains a set of triangles as shown in the upper left image in Fig. 4. Then, the proposed method computes an affine transformation matrix **A** for each triangle by referring to the result of the shape generation step. This transformation transforms vertices of each triangle from an input pedestrian image to the synthesized shape. Then, the texture inside each

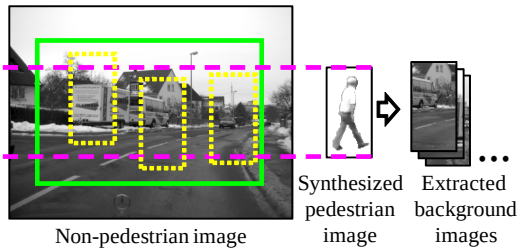


Figure 5: Examples of extracted background images.

triangle is mapped onto the synthesized shape by using this transformation matrix **A**. Finally, the proposed method applies this texture mapping process for all triangles obtained by the Delaunay triangulation algorithm.

After applying the above process, variously textured pedestrian images for the same pose can be obtained. By using these images, the proposed method synthesizes various textures for each pose. First, the proposed method represents intensities of each image as an intensity vector. Then, by applying the SSM algorithm to the intensity vectors, a new pedestrian texture is obtained.

2.1.2 Background Synthesis

As the last step, the proposed method combines the synthesized pedestrian image with various background images. In this step, the proposed method extracts background images from in-vehicle camera images containing no pedestrian by changing the parameters such as the clipping position and the size of the clipping rectangle. Since we can assume that a pedestrian does not float in the sky nor lie on the road, the proposed method sets the parameters for background extraction so that an image is not composed of only the sky or a road surface. Figure 5 shows examples of the extracted background images. Finally, the proposed method uses alpha blending for synthesizing a pedestrian image super-imposed on a background image.

2.2 Training Phase

In this phase, the proposed method constructs multiple classifiers optimized for each pedestrians' pose by using the synthesized pedestrian images. Here, the multiple classifiers consist of simple two-class classifiers. The proposed method optimizes the performance of each classifier so that each classifier can detect each posed pedestrian.

First, the proposed method extracts HOG features from the synthesized pedestrian images and non-pedestrian images. Then, a linear SVM classifier is

constructed for each pedestrians' pose by using these features. Here, libSVM[1] is used for constructing the SVM classifiers.

2.3 Detection Phase

In the detection phase, pedestrians are detected from in-vehicle camera images by using the trained classifiers as seen in Fig. 6. In this phase, pedestrian detection is performed by sliding a detection window over the entire region of an image, and each detection window is evaluated by applying multiple classifiers. Here, the proposed method computes outputs of multiple classifiers for each detection window, and the maximum is used as a pedestrian likelihood $F(\mathbf{i})$. Thus,

$$F(\mathbf{i}) = \max\{f^1(\mathbf{i}), f^2(\mathbf{i}), ..., f^K(\mathbf{i})\}, \qquad (2)$$

where $f^k(\mathbf{i})$ is a two-class classifier corresponding to each posed pedestrian, and $\mathbf{i}$ represents an extracted HOG feature. Finally, if $F(\mathbf{i})$ is larger than a threshold ε, the proposed method outputs that the detection window contains a pedestrian.

3 EXPERIMENT

We evaluated the performance of the proposed method by using in-vehicle camera images. The following sections describe details of the dataset used in the experiment and the results of the proposed method.

3.1 Dataset

In this experiment, the proposed method was evaluated by using the "Daimler Pedestrian Detection Benchmark"[2] dataset which consists of 15,660 pedestrian images and 6,745 non-pedestrian images. We manually selected 200 pedestrian images in daylight conditions from this dataset as inputs of the generation phase. Also, we prepared 35,500 non-pedestrian images in various scales by gathering false positives from a weak detector constructed through a preliminary experiment on this dataset (Fig. 7). For validation, we prepared 1,016 in-vehicle camera images including 1,110 pedestrians. The resolution of the in-vehicle camera images was 640×480 pixels.

[1]LIBSVM A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2]http://www.gavrila.net/Research/Pedestrian_Detection/Daimler_Pedestrian_Benchmarks/Daimler_Pedestrian_Detection_B/daimler_pedestrian_detection_b.html

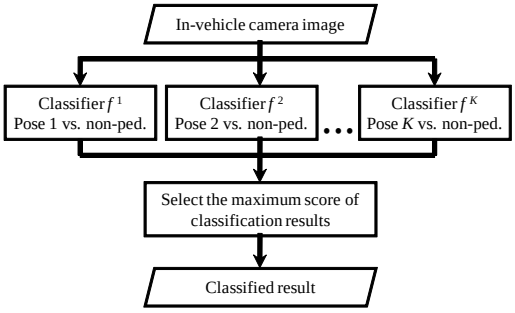


Figure 6: Overview of the multiple classifiers.

Figure 7: Examples of the pedestrian and non-pedestrian images used for training.

Figure 8: Examples of the synthesized pedestrian images.

3.2 Generation of Pedestrian Images

First of all, 200 pedestrian images were divided into eleven pose classes corresponding to each pedestrians' pose. In this step, the contours of all pedestrian images were extracted manually. Then, 15,660 pedestrian images were synthesized by using these images as seeds. We assumed a uniform distribution for the a parameter **b**. Figure 8 shows the examples of the synthesized pedestrian images.

3.3 Performance Evaluation

The performance was evaluated by ROC curves representing the relationship between the detection rate and the false positives per frame. The detection rate was measured by evaluating the overlap between the detection result and the ground-truth labeled manually. The ROC curves were drawn by changing the threshold ε introduced in section 2.3.

To confirm the performance of the proposed method, we compared the proposed method with

Table 1: Specifications of the proposed method and the comparative methods.

Methods	Generation of training images	Classifier	Initial inputs of pedestrian images	Num. of images used for construction of classifiers	
				Ped.	Non-Ped.
Proposed method	yes	multiple two-class	200	15,660	35,500
Comparative method 1	yes	simple two-class	200	15,660	35,500
Comparative method 2	no	simple two-class	15,660	15,660	35,500

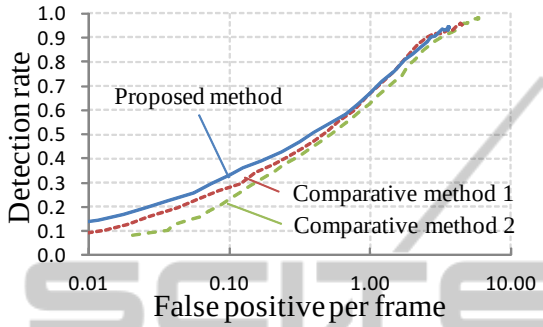


Figure 9: ROC curves of the proposed method and the comparative methods.

three comparative methods. Table 1 shows the specifications of the proposed method and the comparative methods. The proposed method used 15,660 synthesized pedestrian images (generated from only 200 pedestrian images) and 35,500 non-pedestrian images to construct the multiple classifiers. Although the comparative method 1 used the same images with the proposed method, it applied a simple two-class classifier. The comparative method 2 used 15,660 pedestrian images including the initial inputs of the proposed method obtained manually. These methods were simpler versions of a previous work (Enzweiler et al., 2009), where they did not employ bootstrapping iteration when gathering negative samples, compared to the original method. Since it is difficult to segment pedestrian regions manually from 15,660 pedestrian images due to its cost, we could not compare the performance with their multiple classifier versions using all pedestrian images.

3.4 Results and Discussions

Figure 9 shows the ROC curves of the three methods. The proposed method and the comparative method 1 outperformed the comparative method 2. Figure 10 shows examples of the detection results where the proposed method and the comparative method 1 could detect pedestrians correctly but the comparative method 2 could not. Here, each result is the result from a classifier giving the highest performance (F-measure) for each method.

As can be seen in the first and the second columns of Fig. 10, although the comparative method 2 could not detect pedestrians, the proposed method and the comparative method 1 could detect pedestrians correctly. In general, to detect pedestrians with complex backgrounds, the classifier should be trained by using various training samples including complex backgrounds. Since the comparative method 2 could not train various complex backgrounds, it could not detect such pedestrians. In contrast, since the proposed method and the comparative method 1 synthesized various pedestrian images with various backgrounds, these pedestrians could be detected correctly. Also, the proposed method and the comparative method 1 synthesized variously posed pedestrians for training. Therefore, these method could also detect pedestrians whose poses were not included in the initial pedestrian images. Thus, we can say that the generative learning method outperformed the simple gathering method for the controlled synthesis.

As can be seen in Fig. 9, the performance of the proposed method outperformed the comparative method 1. The detection results of these methods are shown in Fig. 10. From these results, we can say that the proposed method could detect variously posed pedestrians in comparison with the comparative method 1. Especially, it can be observed that the proposed method could detect not only walking pedestrians but also standing posed pedestrians. Since the proposed method constructed multiple classifiers optimized for various poses, the detection performance improved against variously posed pedestrians.

In the proposed method, outputs of the constructed multiple classifiers were simply combined by taking the maximum of the detection scores. However, the detection performance may be highly affected by an incorrect output of a classifier. Therefore, we will investigate other methods for combining the outputs from multiple classifiers.

4 CONCLUSIONS

This paper proposed a novel method for detecting variously posed pedestrians. The proposed method constructed multiple classifiers optimized for each pose

Figure 10: Comparison of the detection results; (a) Proposed method, (b) Comparative method 1, and (c) Comparative method 2.

of pedestrians. Also, the proposed method introduced a generative learning method to reduce the bias and the cost for preparing numerous pedestrian images for each pose.

Next, we evaluated the performance of the proposed method by applying it to in-vehicle camera images, where the proposed method outperformed the performance of the comparative methods. We also confirmed that the proposed method could detect pedestrians with various poses stably.

Future work includes the evaluation of the performance by changing the number of initial pedestrian images and the investigation of other methods for combining the outputs of multiple classifiers by considering the actual distribution of pedestrian poses.

ACKNOWLEDGEMENTS

We give a special thanks to the members of Murase laboratory at Nagoya University. Parts of this research were supported by JST CREST and MEXT Grant-in-Aid for Scientific Research. This work was developed based on the MIST library (http://mist.murase.m.is.nagoya-u.ac.jp/).

REFERENCES

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models. Their training and application. *Computer Vision and Image Understanding*, 61:38–59.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893.

Doman, K., Deguchi, D., Takahashi, T., Mekada, Y., Ide, I., and Murase, H. (2009). Construction of cascaded traffic sign detector using generative learning. In *Proceedings of 4th International Conference on Innovative Computing, Information and Control*, pages 889–892.

Enzweiler, M. and Gavrila, D. M. (2008). A mixed generative-discriminative framework for pedestrian classification. In *Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Enzweiler, M., Member, S., IEEE, and Gavrila, D. M. (2009). Monocular pedestrian detection: Survey and expreriments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195.

Gavrila, D. M. and Giebel, J. (2001). Virtual sample generation for template-based shape matching. In *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 676–681.

Murase, H. (1996). Learning by a generation approach to appearance-based object recognition. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 1, pages 24–29.

Noda, M., Takahashi, T., Deguchi, D., Ide, I., Murase, H., Kojima, Y., and Naito, T. (2009). Recognition of road markings from in-vehicle camera images by a generative learning method. In *Proceedings of the 11th IAPR Conference on Machine Vision Applications*, pages 514–517.