15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Pedestrian Near-Miss Analysis on Vehicle-Mounted Driving Recorders

Teppei Suzuki
National Institute of Advance Industrial Science and Technology
Tsukuba City, Ibaraki, Japan
tsuzuki@aoki-medialab.org

Yoshimitsu Aoki
Keio University
Yokohama City, Kanagawa, Japan
aoki@elec.keio.ac.jp

Hirokatsu Kataoka
National Institute of Advance Industrial Science and Technology
Tsukuba City, Ibaraki, Japan
hirokatsu.kataoka@aist.go.jp

Abstract

Recently, a demand for video analysis on vehicle-mounted driving recorders has been increasing in vision-based safety systems, such as for autonomous vehicles. The technology must be positioned one of the most important task, however, the conventional traffic datasets (e.g. KITTI, Caltech Pedestrian) are not included any dangerous scenes (near-miss scenes), even though the objective of a safety system is to avoid danger. In this paper, (i) we create a pedestrian near-miss dataset on vehicle-mounted driving recorders and (ii) propose a method to jointly learns to predict pedestrian detection and its danger level {high, low, no-danger} with convolutional neural networks (CNN) based on the ResNets. According to the result, we demonstrate the effectiveness of our approach that achieved 68% accuracy of joint pedestrian detection and danger label prediction, and 58.6fps processing time on the self-collected pedestrian near-miss dataset.

1 Introduction

Since the histogram of oriented gradients (HOG) [1] was proposed by Dalal *et al.*, pedestrian detection has been noticed in computer vision community. The HOG can get an abstract object shape as a gradient histogram. To generate an object detector, the HOG is desirable to be combined with any classifiers such as SVM, AdaBoost and random forests. A pedestrian dataset must be also prepared for pedestrian detection, for example, the INRIA Person Dataset [1] and the Caltech Pedestrian Detection Benchmark [2], and pedestrian detection has been more accurate and faster in the decade.

On the other hand, since the AlexNet [3] achieved an outstanding performance in classification task on the Imagenet large-scale visual recognition challenge (ILSVRC) [4], the convolutional neural networks (CNNs) got frequently used various tasks that include classification, detection and semantic segmentation. The trend highly accelerates the vision-based system for autonomous vehicles. The KITTI benchmark [5] contains various tasks such as object detection, stereo matching and semantic segmentation toward a realization of autonomous vehicles. Thanks to the large-scale benchmark, a detection-based prevention and checking road surface are ready for a safety system. However, the conventional datasets do not contain any near-miss events, even though the top priority of safety system such as autonomous vehicles is to avoid a risk of car-to-pedestrian contact situation. To strengthen a current safety system for driving, a collection and analysis of near-miss scenes must be done since it is the most important task in the current traffic safety.

In this paper, we propose the pedestrian near-miss dataset on vehicle-mounted driving recorders and joint learning of pedestrian detection and danger level prediction with a CNN based on the ResNets.

Our contributions are as follows:

- We have collected the pedestrian near-miss dataset on a vehicle-mounted driving recorders, which contains a large number of near-miss scenes obtained by mounting driving recorders in over one hundred taxis.
- We propose a joint optimization of pedestrian detection and danger-label prediction in the framework of deep residual networks (ResNets). The proposed architecture outputs a detection rectangle and predicted danger-label with {high, low, no-danger}.
- We demonstrate the effectiveness of our approach that is an accurate prediction and real-time processing on the self-collected pedestrian near-miss dataset.

2 Related works

2.1 Pedestrian detection

There are many pedestrian detection approaches based on the HOG feature [1]. The HOG feature represents a feature vector with oriented gradients in an image, and various HOG-based works are existing. A combination of a HOG-based feature and any classifiers

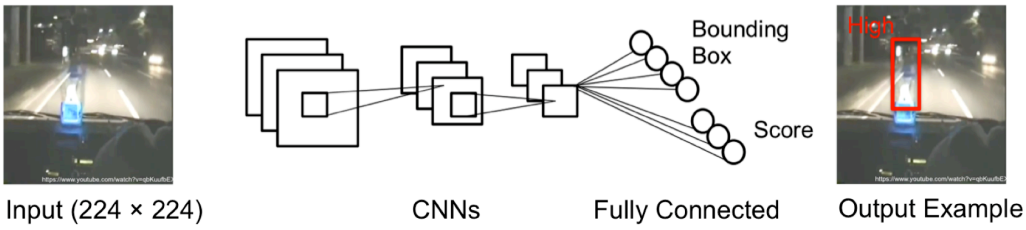

Figure 1. Proposal method.

such as SVM and random forests allow us to detect a pedestrian.

Here, the epoch-making CNN was occurred in 2012. Since the AlexNet [3] achieved overwhelming rate in classification task on the ILSVRC2012 [4], various approaches have been proposed in computer vision field. Researchers have proposed tunings for pedestrian detection. Above these lines, a prestigious pedestrian detector, scale-aware fast R-CNN (SAF-RCNN) [7] based on fast R-CNN [6] achieved an outstanding precision. Especially, the SAF-RCNN reached 9.68 % with miss rate on the Caltech pedestrian benchmark [2]. When humans do the same task, the miss rate is recorded as 5.60 %. In other words, the performance rate of computer is being approaching human's one. Moreover, Zhang et al.[9] proposed the improved model from the faster R-CNN [8] that achieved state-of-the-art on the representative datasets including the Caltech pedestrian.

However, these methods have some problems in terms of computational cost. Fast/faster R-CNN is more speeded-up approach than other CNN-based methods, however the speed-up tuning have saturated because the model difficulty. Although fast/faster R-CNN are designed for multi-class detection, our task is required a single-class detection and its danger-label. A simplified model should be proposed in order to be an efficient computation.

2.2 Temporal analysis

The dense trajectories (DT) [10] are used on a video analysis. To extract densely captured optical flows and combined feature vectors, the DT represents a sophisticated motion vector. Recently, the hand-crafted DT was replaced by two-stream ConvNets [11] which is a successful approach in action recognition. The model has two CNN models that one is trained on stacked flow images, another model is trained on RGB sequential images.

However, the both DT and two-stream ConvNets rely on temporal information especially with optical flow. In our experiments, the flow-based approach does not work well on the pedestrian near-miss dataset. We believe that a high-standard architecture from an RGB image is reasonable in the joint task.

3 Pedestrian NEAR-MISS dataset

In this section, we show the details of pedestrian near-miss dataset (PNM dataset) using this paper.

A near-miss scene is car-to-pedestrian contact situation in the paper. PNM dataset totally has 1,516

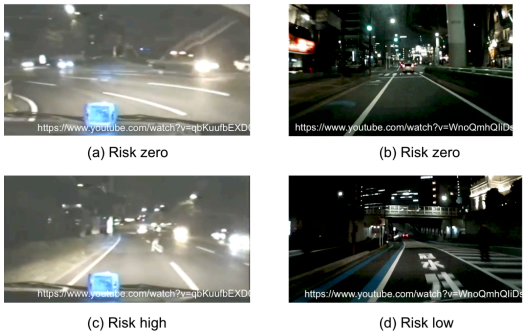

Figure 2. Example of NEAR-MISS situation

videos and 45,480 frames (each video contains 30 frames). We apply 1,316 videos for training sample, and 200 videos for test sample. The near-miss videos were obtaiend by vehicle-mounted driving recorders in over one hundred taxis. The video recording system was trigged if there was sudden braking, resulting in deceleration of more than 0.5 G; a 15-second image sequence was then recorded.

Each video is annotated risk rate {high, low} and each frame is annotated bounding box of pedestrian related to the near-miss situation. When a pedestrian is not in the frame, bounding box is not appeared. Our goal is to solve joint task of pedestrian detection and risk prediction, therefore when a pedestrian doesn't exist, the system outputs risk zero label at that frame. After all we predict 3 classes {high, low, zero}.

PNM dataset contains videos taken from various vehicles, places {intersection, city area, major road}, time {daytime, night} and weather {sunny, cloudy, rainy}.

We show the examples of near-miss scene in the Figure 2.

4 Joint pedestrian detection and risk prediction

By using a single model, we performs both pedestrian detection and its risk prediction. Our model is based on the deep residual networks (ResNet) [12, 13] from image input resized 224 [pixels] × 224 [pixels] x 3 [channels]. The ResNet has a bottleneck layer that is to perform identity mapping and this layer make a residual learning possible. As a result, a bottleneck layer enables to learn very deep convolutional architecture such that more than 100 layers. The reason why we use ResNet is to achieve sophisticated representation

for the both pedestrian detection and risk prediction in single CNN model. In Section 6, we show that ResNet has the best performance in some models.

A 2048-dimension vector is captured from ResNet. The vector is input to fully-connected layer for pedestrian detection and risk prediction. The output is consists of a 7-dimension vector which expresses bounding box {x, y, width, height} and risk scores {high, low, zero}. Our model architecture is shown in the Figure 1.

5 Implementation

We show the details of CNN training.

We use the 1,316 videos for training sample. Each video consists of 30 frames, namely training data totally contains 39,480 frames. Each frame is annotated danger label {high, low, no-danger} and bounding box {x, y, width, height}. However, when pedestrian doesn't exist in frame, bounding box is not annotated.

We define a loss function as follow:

$$L = \alpha l_{bbox}(b_g, b_p) + l_{class}(y, x) \tag{1}$$

where l_{bbox} is smooth L1 loss between bounding box of ground-truth b_g and bounding box of model output b_p and l_{class} is cross-entropy loss. y is the ground-truth label and x is prediction label. α is a trade-off term between mean squared error and cross-entropy loss. We set $\alpha = 0.01$. We update our model by back-propagating the gradient of the loss function. We use SGD as an optimizer. We set learning rate 0.1, and every 30 epochs the rate is multiplied by 0.1. Momentum is set as 0.9, weight decay is set 0.0001.

We set an output value of bounding box $[0, 224]$, but output value order of ResNet is one digit. We consider optimization is difficult in such order, so we add multiplying a constant layer before the output layer. This layer multiplies the input value by 100. It was better accuracy than doesn't add the output layer.

Our model is implemented on the Torch7 [14].

6 Experiments

To find the best approach, we compare ours with some representative approaches. We use AlexNet [3], VGG-16 [15], Network-in-Network(NIN) [16], ResNet-50, ResNet-101 and ResNet-152 [12, 13] as a base model for comparison. These well-organized models achieved high-level results on the ILSVRC [4]. We define the accuracy calculation as follow:

$$accuracy = \frac{TP_{detection} \cap TP_{prediction}}{GT} \tag{2}$$

where TP is number of true positive of detection box and label prediction, and GT is a number of ground truth (total of frames). We decided a success of pedestrian detection when an intersection-over-union (IoU) between ground truth and network prediction is lager than 0.5. We calculate IoU using following equation:

$$IoU = \frac{box_{ground-truth} \cap box_{prediction}}{box_{ground-truth} \cup box_{prediction}} \tag{3}$$

where *box* indicates the region (number of pixels) of bounding boxes.

As a result, accuracy in test set of all models except ResNet-152 was mostly same as random prediction. The optimization does not work well. Against the other models, the ResNet-152 successfully optimized the joint task with the high non-linearlity from residual network.

We reached to the 68.8% accuracy and process speed 58.8 fps on NVIDIA GeForce GTX 1080 using ResNet-152. The detailed results are shown in Figure 3. In the false positive, our system predicts risk low (ground-truth is risk zero). In the true positive's visualization image, more parts become red than false positive's visualization image. It means that our system doesn't get an evidence to predict risk low. In fact, the confidence score is mostly the same value in 3 classes. We consider it is caused by overfitting. We define the image without human as a risk zero, but PNM dataset have only near-miss situation, so there are few frames where no one exist. As a result, our model gets overfitting and it became difficult to estimate a risk zero.

7 Conclusion

In this paper, we presented pedestrian near-miss dataset (PNM dataset). The purpose of this dataset is to analysis near-miss scenes and thereby to improve safety system for autonomous vehicles. And we also proposed an end-to-end pedestrian detection and risk prediction model on the PNM dataset. Our proposed method runs 58.6fps on a GPU and its accuracy reached to 68.8%.

References

[1] Navneet Dalal, et al. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition*, 2005.

[2] Piotr Dollar, et al. "Pedestrian detection: A benchmark." *Computer Vision and Pattern Recognition*, 2009

[3] Alex Krizhevsky, et al. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, 2012

[4] Jia Deng, et al. "Imagenet: A large-scale hierarchical image database." *Computer Vision and Pattern Recognition*, 2009.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." *Computer Vision and Pattern Recognition*, 2012.

[6] Ross Girshick. "Fast r-cnn." *International Conference on Computer Vision*, 2015.

[7] Jianan Li , et al. "Scale-aware Fast R-CNN for Pedestrian Detection." *arXiv preprint arXiv*:1510.08160 (2015).

[8] Shaoqing Ren, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*, 2015.

[9] Liliang Zhang, et al. "Is Faster R-CNN Doing Well for Pedestrian Detection?." *European Conference on Computer Vision*, 2016.

[10] Heng Wang, et al. "Action recognition by dense trajectories." *Computer Vision and Pattern Recognition*, 2011.

[11] Karen Simonyan, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos."

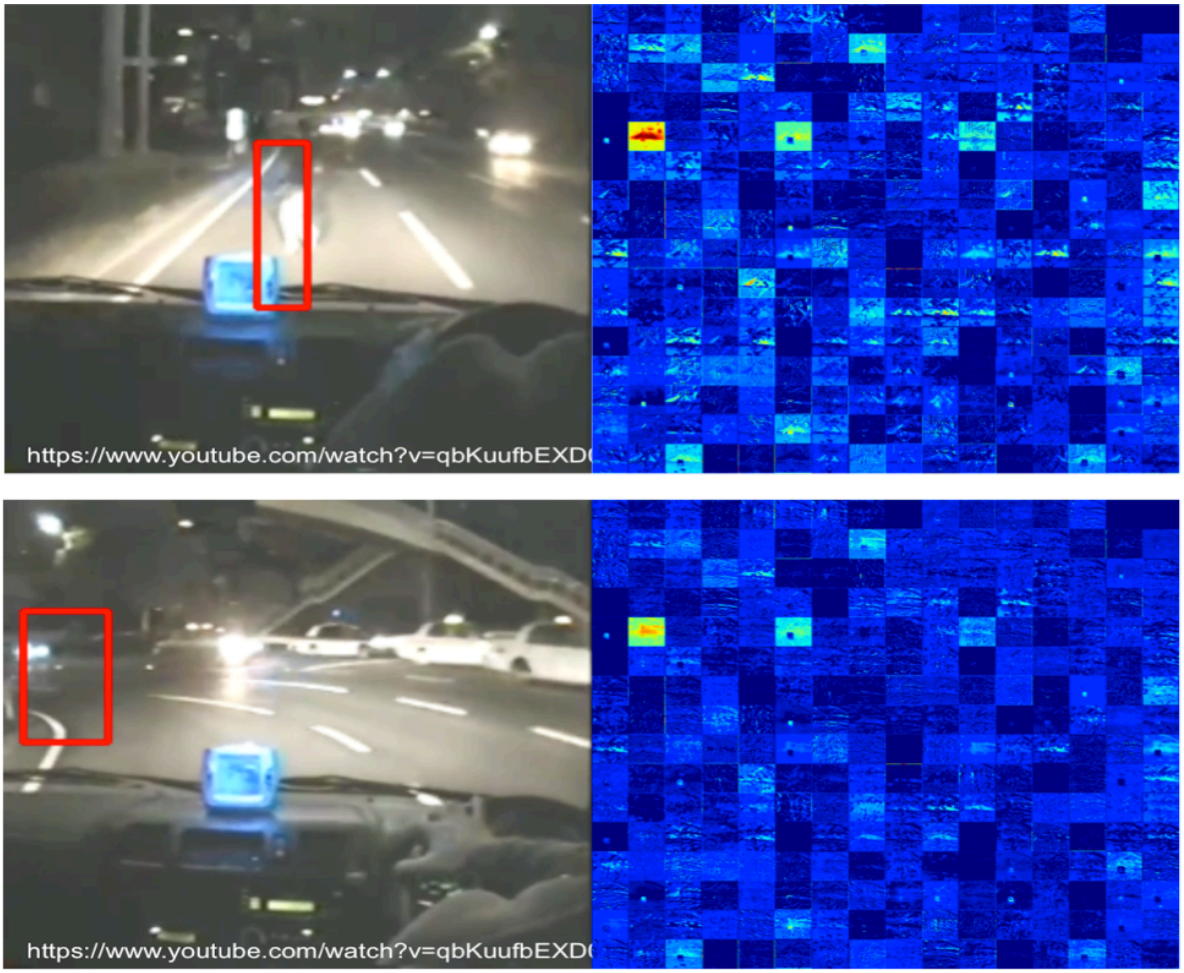

Figure 3. The results of pedestrian detection/risk prediction and convmap visualization. The top image is true positive example, the bottom image is false positive example. The right figure at each image is visualized result of a hidden layer.

Advances in Neural Information Processing Systems, 2014.

[12] Kaiming He, et al. "Deep residual learning for image recognition." *arXiv preprint arXiv*:1512.03385, 2015.

[13] Kaiming He, et al. "Identity mappings in deep residual networks." *arXiv preprint arXiv*:1603.05027, 2016.

[14] Onan Collobert, et al. "Torch7: A matlab-like environment for machine learning." *NIPS Workshop.*, 2011.

[15] Karen Simonyan, et al. "Very deep convolutional networks for large-scale image recognition." *International Conference on Learning Representations*, 2015.

[16] Min Lin, et al. "Network in network." *arXiv preprint arXiv*:1312.4400, 2013.