Hard Negative Mining from in-Vehicle Camera Images based on Multiple Observations of Background Patterns

Masashi Hontani[1], Haruya Kyutoku[1], David Wong[1], Daisuke Deguchi[2],
Yasutomo Kawanishi[1], Ichiro Ide[1] and Hiroshi Murase[1]
[1] *Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*
[2] *Information Strategy Office, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*

Keywords: Pedestrian Detection, Hard Negative Mining, Additional Learning.

Abstract: In recent years, the demand for highly accurate pedestrian detectors has increased due to the development of advanced driving support systems. For the training of an accurate pedestrian detector, it is important to collect a large number of training samples. To support this, this paper proposes a "hard negative" mining method to automatically extract background images which tend to be erroneously detected as pedestrians. Negative samples are selected based on the assumption that frequent patterns observed multiple times in the same location are most likely parts of the background scene. As a result of an evaluation using in-vehicle camera images captured along the same route, we confirmed that the proposed method can automatically collect false positive samples accurately. We also confirmed that a highly accurate detector can be constructed using the additional negative samples.

1 INTRODUCTION

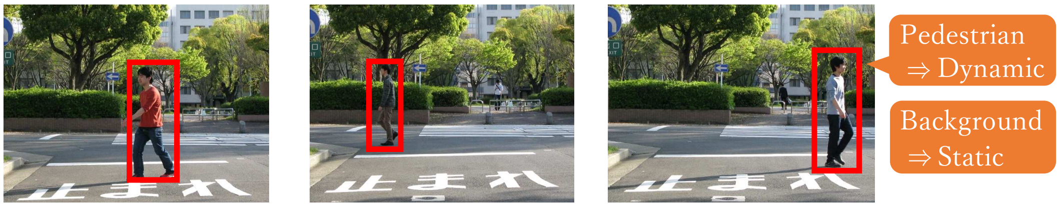
In recent years, there has been a decrease in the number of traffic accidents because of improvements in road and vehicle safety. However, even so, in Japan alone there were about five hundred thousand traffic accidents in 2016, and road safety remains an important social issue to be overcome (Statistics Bureau, 2017). Therefore, strategies for reducing traffic accidents remain an active area of development. For example, in the last few years, technologies for assisting drivers to drive safely have been put into practical use in production vehicles.

A fundamental technology for driving assistance systems is a detection method for pedestrians and obstacles around the vehicle (Redmon et al., 2016; Premebida et al., 2014). Modern vehicles may employ a variety of sensors for this purpose—for example, RADAR and cameras. In addition, using advanced sensors such as LIDAR for vehicle applications is an area of intensive research. However, since a monocular camera is inexpensive and compact, and can be easily fitted to existing vehicle designs, this paper aims to improve the accuracy of pedestrian detectors by employing a monocular camera.

When constructing a pedestrian detector, it is important for the classifier to learn the various appearances of pedestrians. Although in general thousands of training samples are necessary, manually collecting a large number of training samples is time consuming and expensive. Therefore, methods for constructing a detector from limited annotated samples have been proposed. In these methods, the initial pedestrian detector is constructed with annotated samples. Then, training samples are extracted from the detection results and re-annotated as positive or negative samples. Finally, a more accurate detector is reconstructed including these additional samples. For example, Yuan et al. proposed a method for constructing highly accurate traffic sign detectors by automatically processing the results of a base detector (Yuan et al., 2017). Traffic signs from in-vehicle camera images are characteristic in that they have a regular symmetrical shape and can be stably detected and tracked between sequential frames. Therefore, the initial detector's detection results can be labelled as positive or negative samples, and used for online learning to improve both detection and tracking results.

Another method proposed by Mitsugami et al. constructs a pedestrian detector for fixed cameras, adapted to capture location (Mitsugami et al., 2013). In this method, a detection is judged to be false when

Hontani, M., Kyutoku, H., Wong, D., Deguchi, D., Kawanishi, Y., Ide, I. and Murase, H.
Hard Negative Mining from in-Vehicle Camera Images based on Multiple Observations of Background Patterns.
DOI: 10.5220/0007376804350442
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 435-442
ISBN: 978-989-758-354-4



(a) t_1 (b) t_2 (c) t_3

Figure 1: Three images captured at the same location at different timings, showing detected pedestrians.

the same object is detected at the same position within sequential frames. Conversely, it is judged to be correct when the object moves smoothly, adhering to a pedestrian motion model. Finally, the detector is reconstructed with the additional information provided by the positive and negative samples.

These methods show that additional training using the detection results of the initial detector is useful for improving the overall detection accuracy. However, Yuan's method cannot be applied to detecting dynamic objects since it uses the characteristics of static objects. Also, Mitsugami's method assumes that images come from a fixed camera, so it cannot be applied to in-vehicle camera images captured by moving cameras. For driving assistance systems, it is necessary to automatically extract samples from the detection results of moving pedestrians from images captured by a moving, in-vehicle camera.

Meanwhile, hard negative mining is widely applied for constructing highly accurate object detectors (Felzenszwalb et al., 2010; Dalal et al., 2006; Shrivastava et al., 2016). In hard negative mining, false positives that are easily detected as incorrect are selected and used for re-training the detector. While negative samples can be supplied by using random background patches, it is considered efficient to automatically collect negative samples which caused the detector to trigger a false alarm, hence the name "hard negative".

From the above background, we consider a method of extracting samples from in-vehicle camera images and using them for additional training. Because it is difficult to accurately extract positive samples without manual annotation, we focus on hard negative mining by extracting negative samples that can be reliably extracted using prior scene knowledge. Therefore, we focus on the difference between the characteristics of pedestrian regions and background regions which may trigger a false detection in order to extract negative samples automatically. An example where pedestrians are detected in the same location at a different timing is shown in Fig. 1. As shown here, pedestrians are dynamic objects which are not typically captured at the same position over a long period of time. Conversely, background areas are static. We propose a method of extracting negative samples considering this observation.

The contributions of this paper can be summarized as follows:

1. The proposal of a novel framework of hard negative mining from non-annotated in-vehicle image sequences.
2. The combination of spatial and temporal alignments for the detection of frequently observed background patterns.
3. The use of the resulting hard negative samples for training state-of-the-art deep-learning-based detectors.

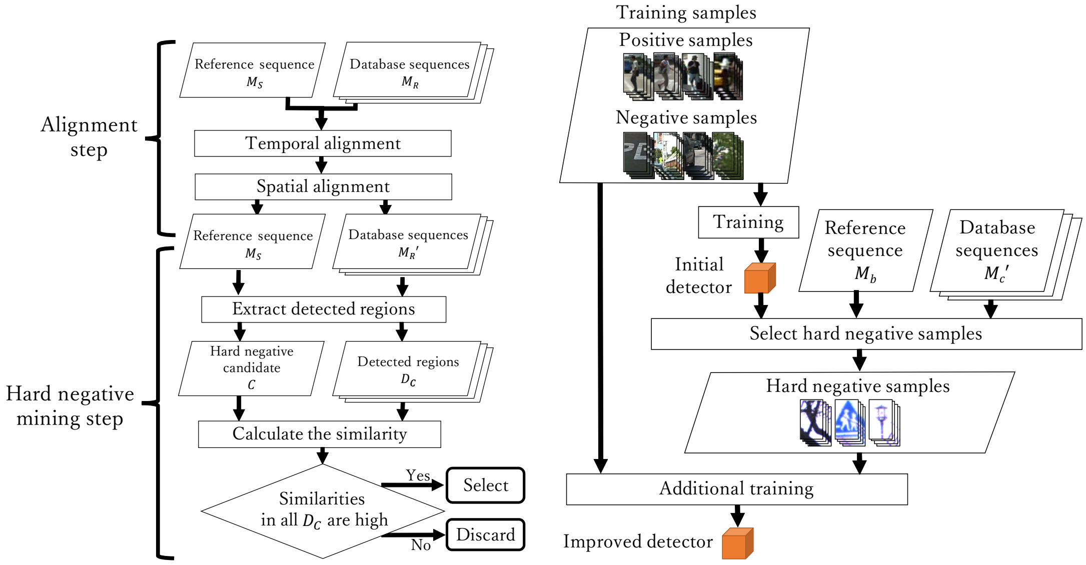
The rest of the paper is organized as follows: Section 2 describes our negative sample selection method. Then, we report evaluation experiments using images captured by an in-vehicle camera and discuss the results in Section 3. Finally, we conclude this paper in Section 4.

2 HARD NEGATIVE MINING FRAMEWORK

A pedestrian region detected by a pedestrian detector can be classified as either a correct detection (true positive) if the region truly contains a pedestrian, or a false detection (false positive) if the region consists of background. Since we wish to extract only the false detections as negative samples, the proposed method collects "hard negative" candidates. Here, a "hard negative" refers to a falsely detected pedestrian region, which is then used to create a negative sample, for use in further training of the pedestrian detector.

False detections made by pedestrian detections can be of two types (Mitsugami et al., 2013):

1. Detection of background areas with features similar to a pedestrian.
2. Detection of parts of a pedestrian.



(a) Alignment / hard negative mining steps

(b) Reconstruction step

Figure 2: System flow of the proposed method.

Of these two types, the first one is likely to occur in all images captured at a similar spatial location, as the background area containing a pattern visually similar to a pedestrian does not change. This may lead to a high rate of false detections in certain locations, which is undesirable for a driver assistance system. For the second type of false detections, although the region is incorrectly detected as a complete pedestrian, in reality, a pedestrian is likely present nearby—therefore, this type is less obtrusive. Because of this, and the fact that background regions tend to remain visually similar over time, in this research, hard negative candidates are extracted from false detections of the first type.

The proposed method prepares multiple image sequences captured along the same route. Among them, the sequence used for detecting negative samples is called the reference sequence. The other sequences used for determining whether or not the candidate is a false detection, are called the database sequences. The system flow of the proposed method is shown in Fig. 2. It consists of three steps: the alignment step (Fig. 2(a)) for associating a reference sequence with a database sequence, the hard negative mining step (Fig. 2(a)) for determining whether or not hard negative candidates are false detections, and the reconstruction step (Fig. 2(b)) for constructing a detector using additional training with negative samples.

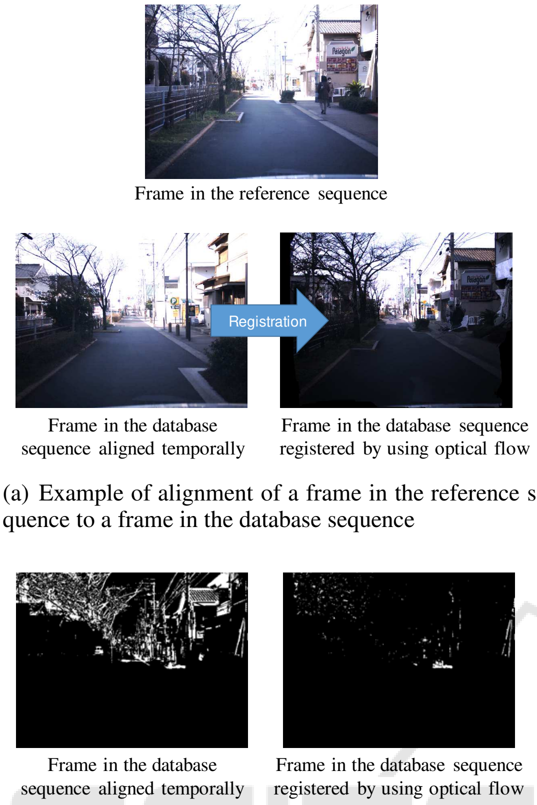
2.1 Alignment Step

In the alignment step, each of the multiple database sequences are aligned temporally and spatially in order to compare the reference sequence to the database sequences.

First, each database sequence is temporally aligned with the reference sequence. It is necessary to associate each frame within the database sequence with the frame within the reference sequence that was captured at the same location. Typically, temporal alignment can be performed referring to the visual similarity between the images in the sequences. However, there are differences of appearance between frames captured from the same location caused by different camera trajectories, small pose changes, and occlusion by obstacles. For this reason, it is difficult to associate frames accurately with a simple image matching method.

In this research, frames within database and reference sequences are temporally aligned using a method of associating frames proposed by Kyutoku et al. (Kyutoku et al., 2011). This method uses image features and epipolar geometry to determine a similarity measure between the capture locations of image frames, allowing temporal alignment to be performed.

Even after image sequences have been temporally aligned, capture positions and camera poses differ slightly between aligned frames. This leads to a positional gap between corresponding pixels of aligned frames, which makes direct visual comparison of aligned frames for background detection difficult. Therefore, temporally aligned frames are then aligned spatially using image registration. This is done using dense optical flow obtained by DeepFlow (Revaud et al., 2016). An example of spatial alignment of pixels in a temporally aligned image frame is shown


Frame in the reference sequence
Registration
Frame in the database sequence aligned temporally
Frame in the database sequence registered by using optical flow
Frame in the database sequence aligned temporally
Frame in the database sequence registered by using optical flow

(a) Example of alignment of a frame in the reference sequence to a frame in the database sequence

(b) Difference images for the frames in the database sequence

Figure 3: Examples of spatial alignment.

in Fig. 3. The image at the top of Fig. 3(a) is the frame in the reference sequence and the bottom left image is the frame in the database sequence selected by temporal alignment. On the bottom right is the image that has been spatially aligned using DeepFlow. The difference images of these frames are shown in Fig. 3(b). We can see that the gap between the reference and the database frames is reduced by the spatial alignment. This allows us to compare their visual features at corresponding positions directly in a common image coordinate system.

2.2 Hard Negative Mining Step

In the hard negative mining step, negative samples are extracted from hard negative candidates by comparing the reference sequence with the database sequences.

First, hard negative candidates are detected in the reference sequence by the initial pedestrian detector. All positive detection results are considered as hard negative candidates. In this paper, as one of the state-of-the-art deep learning based detectors, we employ YOLOv3 (Redmon et al., 2016) (hereafter referred to as the YOLO detector) as a baseline detector.

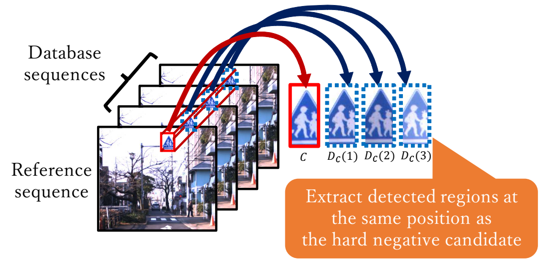

Database sequences
Reference sequence
C
$D_C(1)$
$D_C(2)$
$D_C(3)$
Extract detected regions at the same position as the hard negative candidate

Figure 4: Conceptual diagram of extraction of detected regions.

Next, detected regions $D_C(n)$ are extracted from the database sequences in order to determine whether or not each hard negative candidate C is a false detection, where $n = 1,2,..,n_{\mathrm{MAX}}$ represents the sequence number. A conceptual diagram of the extraction of detected areas is shown in Fig. 4. Detected regions $D_C(n)$ are extracted from each corresponding temporally and spatially aligned database frame at the same spatial position as the candidate in the reference frame. The similarities between C from the reference image and each of $D_C(n)$ from the database sequences are calculated. A CNN model (Ahmed et al., 2015) is used for calculating the similarity between C and $D_C(n)$. The result is a set of similarity values $S_C(n)$ $(n = 1,2,\ldots)$ for each C.

Finally, it is determined whether or not the candidate C is a false detection by using the calculated similarities $S_C(n)$. We consider that the corresponding $S_C(n)$ is high when $D_C(n)$ is captured from the same object or visual region. In this case, C is probably a false detection from the background, as the region does not change temporally between sequences. On the other hand, when the corresponding $S_C(n)$ is low, C is most likely a temporal object such as a pedestrian that appears in only one of $D_C(n)$. It is necessary to reliably select only C from background areas as false detections. Therefore, C is determined as a false detection only when the minimum of $S_C(n)$ exceeds a threshold as

$$\min_n S_C(n) > T, \tag{1}$$

in which case the detection is considered a hard negative sample. A conceptual diagram of hard negative sample determination using similarity is shown in Fig. 5.

2.3 Reconstruction Step

In the reconstruction step, a pedestrian detector is retrained including hard negative samples.

First, the hard negative candidates judged as a false positives in the hard negative mining step are

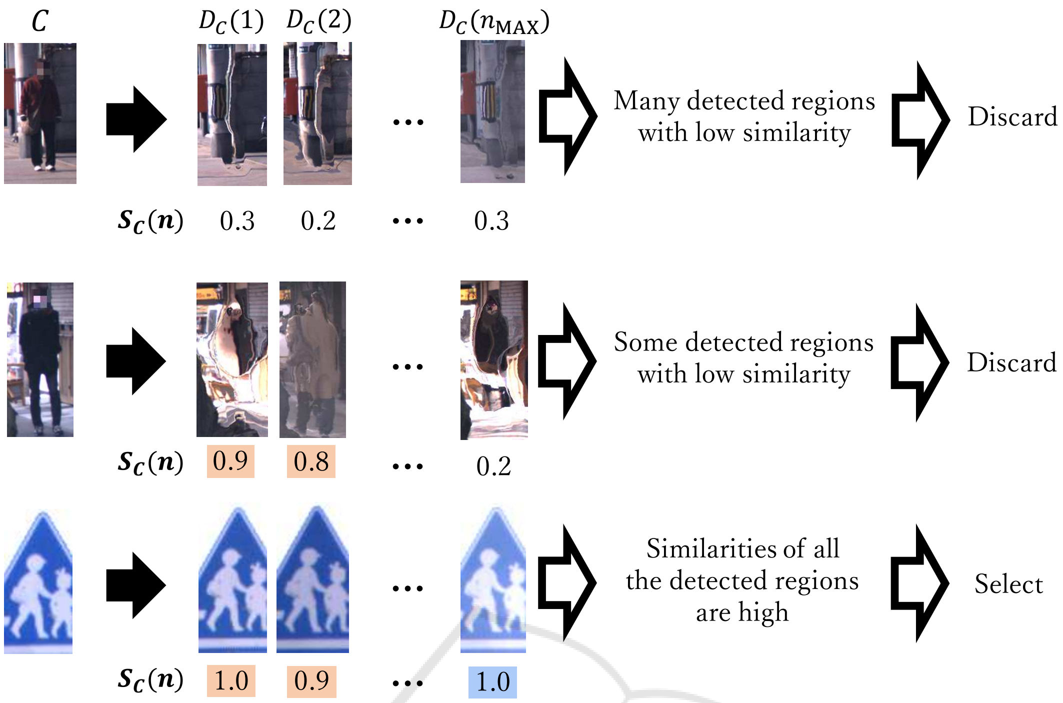

C
$D_C(1)$
$D_C(2)$
$D_C(n_{\mathrm{MAX}})$
Many detected regions with low similarity
Discard
$S_C(n)$
0.3
0.2
0.3
Some detected regions with low similarity
Discard
$S_C(n)$
0.9
0.8
0.2
Similarities of all the detected regions are high
Select
$S_C(n)$
1.0
0.9
1.0

Figure 5: Conceptual diagram of hard negative sample selection.

added to the original training samples used to train the initial detector. The training samples used in constructing the initial detector includes annotated pedestrians as positive samples, and negative samples are randomly sampled from the background region in the YOLO freamwork. In the reconsruction step, negative samples are randomly sampled from the hard negative samples and added to the initial samples. Finally, the pedestrian detector is re-trained with these training samples.

3 EXPERIMENT

Two evaluation experiments using in-vehicle camera images were conducted in order to confirm the effectiveness of the proposed method. Each will be described in detail below.

3.1 Experiment on Negative Samples Selection

An evaluation experiment using multiple in-vehicle camera images was conducted in order to confirm the accuracy of selecting negative samples by the proposed method.

3.1.1 Experimental Condition

For this experiment, four sequences composed of 2,000 to 2,600 frames each, were captured by an in-vehicle camera mounted on a car running along the same route travelling in the same direction, at different timings throughout the day. A Point Grey Research Grasshopper3 GS3-U3-28S4C-C and a Space VP-JHF8M-3MP 8 mm telephoto lens were used. Pedestrian bounding boxes were manually annotated in each camera image, resulting in 3,000 to 7,000 pedestrian frames per sequence.

One of the four in-vehicle camera sequences was used as the reference sequence and the remaining three were used as the database sequences. Next, hard negative candidates were detected from the reference sequence by the initial pedestrian detector in the reference sequence. The YOLO pedestrian detector described in Sec. 2.2 was trained using the CityPersons dataset (Zhang et al., 2017; Cordts et al., 2016) and the Caltech Pedestrian Detection Benchmark Dataset (Dollár et al., 2012; Dollár et al., 2009) (hereafter referred to as the Caltech dataset).

The negative sample selection method was applied to the hard negative candidates from the reference sequence for determining whether or not each candidate was a false detection. All candidates that were determined to be false detections were collected as hard negative samples.

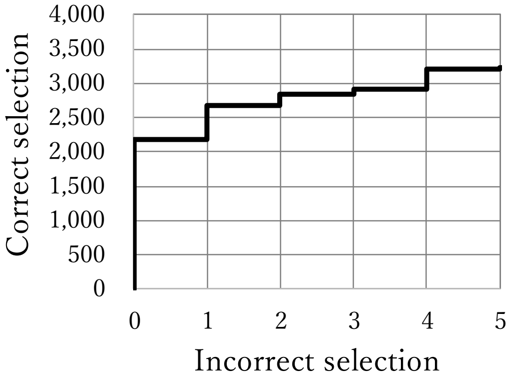


Figure 6: Extraction performance of negative samples by the proposed method.

3.1.2 Evaluation Metrics

The purpose of this research is to accurately select from the initial pedestrian detection results that do not contain a pedestrian. Therefore, correct and incorrect selections are defined in this experiment as follows:

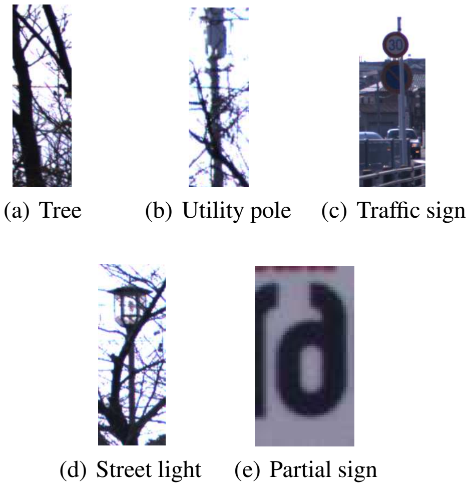
- **Correct Selection:** A detection result that only contains background is selected as a negative sample.
- **Incorrect Selection:** A detection result that contains a pedestrian is incorrectly selected as a negative sample.

The selected hard negative sample is determined to be a correct selection when the IoU (Intersection over Union) for the candidate and the annotated pedestrian bounding boxes is greater than 0.4. In other cases, it is determined to be an incorrect selection. Here, we use the correct and incorrect selection rates with respect to the total number of hard negative candidates as evaluation metrics.

3.1.3 Results and Discussion

The detected hard negative candidates were composed of 3,925 pedestrian images and 80,884 background images. The proposed method was applied to these hard negative candidates while varying the similarity threshold to judge whether or not each hard negative candidate is a false detection. As a result of this, a graph showing the number of correct selections and the number of incorrect selections is shown in Fig. 6. When the number of incorrect selection is zero, 2,150 background images are selected as hard negative samples.

When the similarity threshold was set to a value that yielded no incorrect negative sample selection, the negative samples were typically background images of five types: tree, utility pole, sign, light, and partial sign, as shown in Fig. 7. We consider that the types of the selected negative samples were fairly limited due to the relatively consistent environment covered in the dataset. Therefore, in the future, we will need to confirm that various types of hard negative samples can be extracted from various driving environments.


(a) Tree (b) Utility pole (c) Traffic sign

(d) Street light (e) Partial sign

Figure 7: Types of the selected hard negative samples.

An example in which spatial alignment failed due to differences in appearance of the entire image is shown in Fig. 8. This occurred because frames at different capture locations were sometimes aligned incorrectly during temporal alignment. If the pixel distance between an object in two aligned frames is too large, the optical flow cannot be correctly calculated. Therefore, it is necessary to improve the accuracy of temporal alignment between frames in order to avoid such failed alignments.

Also, an example where the alignment failed due to the influence of the shadow is shown in Fig. 9. The image registration using optical flow tends to fail when two images contain large difference in appearance. It is necessary to also improve the accuracy of spatial alignment since such failure can lead to incorrect extractions.

3.2 Experiment on Detector Construction

An evaluation experiment to construct the pedestrian detector by incorporating the negative samples selected in Sec. 3.1 was conducted in order to confirm their usefulness for training.

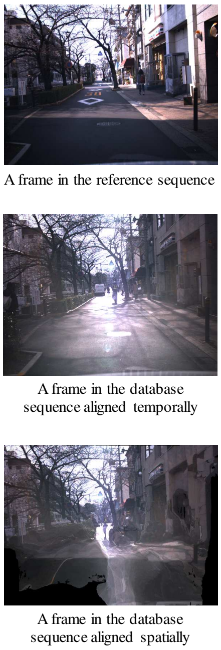

A frame in the reference sequence
A frame in the database sequence aligned temporally
A frame in the database sequence aligned spatially

Figure 8: Example of an image where spatial alignment failed.

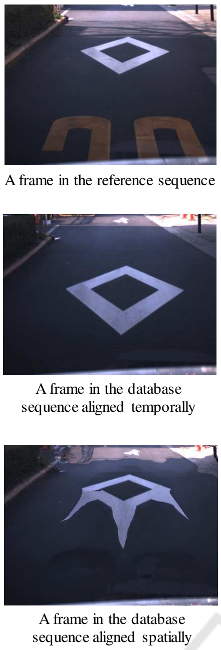

A frame in the reference sequence
A frame in the database sequence aligned temporally
A frame in the database sequence aligned spatially

Figure 9: Example of spatial alignment results including shadows.

3.2.1 Experimental Condition

The annotated pedestrians of the CityPersons and Caltech training datasets were used as the base training samples. The negative samples judged as false detections in Sec. 3.1 were also added in the training samples. The negative samples were selected using the threshold at which the number of incorrect selections was zero (see Fig. 6).

An image sequence (15,445 frames) captured in a city environment during the day using the same in-vehicle camera as the dataset used in Sec. 3.1.1 was used for the evaluation of the pedestrian detectors.

YOLO detectors were constructed using the following training methods:

- **Proposed:** Trained with the CityPersons dataset and the Caltech dataset up to 150,000 iterations. Afterward, trained with additional negative samples selected using the proposed method up to 50,000 iterations.
- **Comparative:** Trained with the CityPersons dataset and the Caltech dataset up to 200,000 iterations. No additional negative samples were included.

Pedestrian detection was then performed on the evaluation dataset using the proposed and the comparative YOLO detectors. The DET (Detection Error Tradeoff) curve was used as the evaluation matrix.

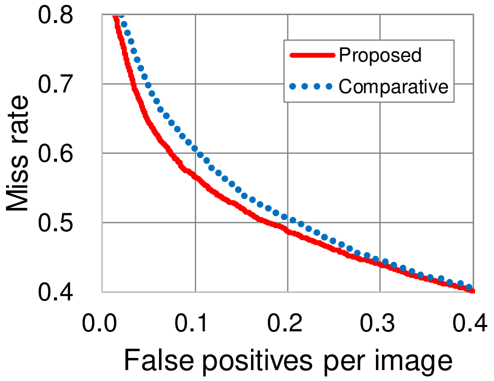


Figure 10: DET curves of each detector.

3.2.2 Result

The DET curves of each detector are shown in Fig. 10. We confirmed that the accuracy of the detector can be improved by training with additional negative samples extracted using the proposed method. For example, when we allowed one false positives in ten images (0.1), the miss rate improved by 4%.

4 CONCLUSION

In this paper, we proposed a method to improve the accuracy of pedestrian detectors for in-vehicle cameras. Specifically, we focused on the characteristic that patterns in background regions which are often incorrectly extracted as pedestrians are static. Therefore, negative samples are extracted automatically from in-vehicle camera image sequences captured multiple times along the same route, by identifying where false positive patterns are observed repeatedly between temporally and spatially aligned images.

In the evaluation experiments, the effectiveness of the proposed method in extracting negative samples and the usefulness of the hard negative samples in additional training were confirmed.

Future tasks related to this research include improving the accuracy of temporal and spatial alignment between image frames, and examining features used for calculation of the similarity of false positives between sequences.

ACKNOWLEDGEMENTS

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research.

REFERENCES

Ahmed, E., Jones, M. J., and Marks, T. K. (2015). An improved deep learning architecture for person reidentification. In *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3908–3916.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3213–3223.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Proc. 9th European Conf. on Computer Vision*, vol. 2, pages 428–441.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. In *Proc. 2009 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 304–311.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4):743–761.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Kyutoku, H., Takahashi, T., Mekada, Y., Ide, I., and Murase, H. (2011). On-road obstacle detection by comparing present and past in-vehicle camera images. In *Proc. 2011 IAPR Conf. on Machine Vision Applications*, pages 357–360.

Mitsugami, I., Hattori, H., and Minoh, M. (2013). Improving human detection by long-term observation. In *Proc. 2nd IAPR Asian Conf. on Pattern Recognition*, pages 662–666.

Premebida, C., Carreira, J., Batista, J., and Nunes, U. (2014). Pedestrian detection combining RGB and dense LIDAR data. In *Proc. 2014 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 4112–4117.

Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 779–788.

Revaud, J., Weinzaepfel, P., Harchaoui, Z., and Schmid, C. (2016). Deepmatching: Hierarchical deformable dense matching. *Int. J. of Computer Vision*, 120(3):300–323.

Shrivastava, A., Gupta, A., and Girshick, R. B. (2016). Training region-based object detectors with online hard example mining. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 761–769.

Statistics Bureau, Ministry of Internal Affairs and Communications, Japan (2017). Japan statistical yearbook 2018. http://www.stat.go.jp/english/data/nenkan/67nenkan/index.htm (accessed 2018/9/28).

Yuan, Y., Xiong, Z., and Wang, Q. (2017). An incremental framework for video-based traffic sign detection, tracking, and recognition. *IEEE Trans. on Intelligent Transportation Systems*, 18(7):1918–1929.

Zhang, S., Benenson, R., and Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4457–4465.