sensors

MDPI

Article

# Temporal and Fine-Grained Pedestrian Action Recognition on Driving Recorder Database

**Hirokatsu Kataoka [1,*], Yutaka Satoh [1], Yoshimitsu Aoki [2], Shoko Oikawa [3] and Yasuhiro Matsui [4]**

1 National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8560, Japan; yu.satou@aist.go.jp
2 Department of Electronics and Electrical Engineering, Keio University, Yokohama 223-8522, Japan; aoki@elec.keio.ac.jp
3 Tokyo Metropolitan University, Tokyo 192-0364, Japan; shoko_o@hotmail.com
4 National Traffic Safety and Environment Laboratory, Tokyo 182-0012, Japan; ymatsui@ntsel.go.jp
* Correspondence: hirokatsu.kataoka@aist.go.jp; Tel.: +81-29-861-2267

Received: 5 January 2018; Accepted: 8 February 2018; Published: 20 February 2018

**Abstract:** The paper presents an emerging issue of fine-grained pedestrian action recognition that induces an advanced pre-crush safety to estimate a pedestrian intention in advance. The fine-grained pedestrian actions include visually slight differences (e.g., walking straight and crossing), which are difficult to distinguish from each other. It is believed that the fine-grained action recognition induces a pedestrian intention estimation for a helpful advanced driver-assistance systems (ADAS). The following difficulties have been studied to achieve a fine-grained and accurate pedestrian action recognition: (i) In order to analyze the fine-grained motion of a pedestrian appearance in the vehicle-mounted drive recorder, a method to describe subtle change of motion characteristics occurring in a short time is necessary; (ii) even when the background moves greatly due to the driving of the vehicle, it is necessary to detect changes in subtle motion of the pedestrian; (iii) the collection of large-scale fine-grained actions is very difficult, and therefore a relatively small database should be focused. We find out how to learn an effective recognition model with only a small-scale database. Here, we have thoroughly evaluated several types of configurations to explore an effective approach in fine-grained pedestrian action recognition without a large-scale database. Moreover, two different datasets have been collected in order to raise the issue. Finally, our proposal attained 91.01% on National Traffic Science and Environment Laboratory database (NTSEL) and 53.23% on the near-miss driving recorder database (NDRDB). The paper has improved +8.28% and +6.53% from baseline two-stream fusion convnets.

**Keywords:** fine-grained pedestrian action recognition; two-stream convnets; driving recorder; advanced driver-assistance systems (ADAS)

## 1. Introduction

Understanding a pedestrian intention is an important work in the recent ADAS and self-driving cars. In an urgent situation, a couple of seconds that is generated by a pedestrian intention estimation could be critical to avoid a collision.

At first, the research of pedestrian analysis was studied in vision-based localization in traffic safety systems [1,2]. The research topic has focused on feature descriptors, classifiers and training strategies from an Red, Green and Blue (RGB)-image input. The performance of pedestrian localization has been rapidly increased due to the burden of experiments about pedestrian localization methods. The noteworthy stream was so-called deep neural networks (DNN), which can automatically learn an effective feature at each recognition task. The DNN performs better than conventional approaches [3–5] in recognition tasks such as image recognition [6,7]. The DNN methods have been applied into intelligent transport

systems (ITS) studies, which include the pedestrian localization [8,9]. We believe that the pedestrian study should be shifted to the next step. Therefore, the paper poses a new problem to the pedestrian study in the ITS field, namely "fine-grained pedestrian action recognition", which is to distinguish different actions between subtle changes. The fine-grained pedestrian action recognition, such as walking straight into turning (see Figure 1), is quite important to estimate an intention for more advanced safety systems.


**Figure 1.** Fine-grained pedestrian actions on the self-collected databases: (**a**) crossing; (**b**) walking straight; (**c**) turning; and (**d**) riding a bicycle. Fine-grained pedestrian action recognition should be an issue in safety systems that have a recognition problem with distinguishing different actions between subtle changes. To improve the recent safety systems such as advanced driver assistance systems (ADAS) and self-driving cars, the concept is very important because a pedestrian intention can be estimated in advance.

There are three difficulties with achieving the concept of fine-grained pedestrian action recognition. The first difficulty is to capture a small but meaningful change in walking pedestrian actions. The problem should be considered with a sophisticated feature representation. In the second, a vehicle-mounted camera is always moving and the background is cluttered. We should extract a suitable feature in a focused area by excluding moving/cluttered background areas. The third difficulty is data collection. The collection of intention change is very difficult due to the rarity of turning action in actual driving. We should construct a feature extraction with relatively small video data.

The paper proposes a novel concept of fine-grained pedestrian action recognition for intention estimation in a safety system. Two databases have been collected by dividing two scenarios into experimental and practical scenes. The experimental dataset contains fine-grained pedestrian actions in a static background, and the practical dataset includes fine-grained actions from a driving recorder. The paper also proposes a convnet-based descriptor that contains a couple of modifications to adapt problem-specified difficulties. In addition to the baseline architecture, a feature enhancement and parameter adaptation has been implemented. The experimental section shows the effectiveness of database and system configuration, and demonstrates pedestrian intention recognition with fine-grained pedestrian action recognition.

The paper is organized as follows. Section 2 summarizes related work. The databases and proposed approach are presented in Sections 3 and 4, respectively. The experimental results are presented and discussed in Section 5. Finally, Section 6 summarizes the paper.

## 2. Related Work

### *2.1. Pedestrian Detection*

Since Dalal et al. presented the histograms of oriented gradients (HOG) [3], pedestrian detection has been an active topic in computer vision. The HOG method has been improved into the Co-occurrence HOG (CoHOG) [10] and Extended CoHOG (ECoHOG) [11]. The CoHOG is known as a high-standard detection approach for pedestrian detection by representing edge pair [10]. Moreover, the ECoHOG replaced edge-pair by gradient magnitude in order to represent strength of curves and lines [11].

Dollar et al. followed the gradient-based features; they confirmed that their integral channel feature (ICF) [12] has resulted in faster and more accurate descriptions of pedestrians.

Due to the great success of convolutional neural networks (CNN), image classification has greatly improved since the era of codeword vectors [13–15]. An outstanding result was obtained by AlexNet [16] in the ImageNet large-scale visual recognition challenge 2012 (ILSVRC 2012), and this is a large impact on the use of deep neural networks in computer vision. Recent models, such as the Visual Geometry Group Net (VGGNet) [6], GoogLeNet [17] and residual networks (ResNet) [7], are known as deeper architectures. Above the line of CNN, we have obtained a sophisticated detection algorithm, Region-based CNN (R-CNN) [18]. The R-CNN is constructed of two phases, namely generating object proposal and category classification. Although the detection algorithm performs better than the conventional detection methods such as HOG and integral channel features (ICF), the original R-CNN is absolutely slow in terms of processing speed (47 s/image). Therefore, the faster algorithm is proposed in Fast/Faster R-CNN [19,20]. The recent algorithms (e.g., Single-shot multibox detector (SSD) [21], you only look once (YOLO) [22,23]) have been improved toward a real-time processing. Zhang et al. [8] provided both a sophisticated model and dataset in pedestrian detection. They claimed that a more sophisticated annotation was required to improve a detection. The well-organized works [24,25] have been extensively studied and compared with human-level detection, and it is considered to be the state-of-the-art method for pedestrian detection. Unlike the pedestrian detection with RGB-input, Dalal et al. [26] and González et al. [27] have applied a temporal channel with optical flows.

The performance of pedestrian detection is being closer to the human-level performance. According to the line, we believe that the pedestrian study should be shifted to the next step. Our proposal is to conduct a pedestrian intention estimation based on the fine-grained action recognition. At first, space-time representation is described including a couple of works for fine-grained action recognition. Then, the recent traffic databases are listed.

### 2.2. Space-Time Representation

Space-time interest points (STIPs) have been the primary focus for action recognition [28]. In a STIP, the time $t$ space is added to the $x, y$ spatial domain. The most important approach is that of dense trajectories (DT) [29], which track densely sampled feature points. In addition, Wang et al. proposed the IDT [30], which estimates the camera motion in order to remove the detection-based noise; it also incorporates a Fisher vector [15].

Recently, temporal models with CNN have been proposed [31–33]. Tran [31] proposed a convolution model for xyt maps that is based on the RGB sequence. The 3D convolutional networks (C3D) approach directly captures the temporal features in an image sequence. Another approach, two-stream CNN, is a well-organized algorithm that captures the temporal feature of an image sequence [32]. The integration of the spatial and temporal streams allows us to effectively enhance the representation of motion. We thus better understand how the spatial information relates to the temporal feature.

### 2.3. Traffic Database

Several practical databases for pedestrian detection and autonomous driving have been proposed in the past decade. Representatives include the INRIA person dataset [3], the Caltech pedestrian dataset [34], and the KITTI dataset [35]. Dalal et al. [3] created the INRIA person dataset. These were important contributions to solving the problem of pedestrian detection. Dollar et al. followed their work, and pursued the problem of pedestrian detection by using the Caltech pedestrian dataset [34,36]. Their detailed analysis was beneficial for improving the descriptors, classifier, and model, removing several difficulties regarding analysis of the benchmark.

The KITTI has been used to set meaningful vision problems for autonomous vehicles [35]; these include problems in stereo vision, optical flow, visual odometry, semantic segmentation, 2D/3D object detection, and 2D/3D tracking. For stereo and optical flow, the problems were updated in 2015 [37]. Thanks to sophisticated approaches, such as fully convolutional networks (FCN) [38] and R-CNN,

there has been improved performance for solving these problems using the KITTI benchmark dataset. In addition, a manner of geometry allows us to improve the rate of object detection [39] and the optical flow [40] not only in stereo [41]. In semantic segmentation, the method extracts knowledge about dense connections, and this can be used with a graphical model [8,42].

Against the databases, our proposed database provides a problem toward a pedestrian intention recognition based on the fine-grained pedestrian action recognition. There is an urgent need for a collection of pedestrian videos to implement and evaluate the significant recognition work.

## 3. Self-Collected Databases

The section presents two self-collected databases about fine-grained pedestrian action recognition, National Traffic Science and Environment Laboratory database (NTSEL) and near-miss driving recorder database (NDRDB). Based on the definition of fine-grained recognition (Fine-grained categorization lies in the continuum between basic level categorization (frog vs. piano) and identification of individuals (face recognition, biometrics). The visual distinctions between similar categories are often quite subtle. See [43]), the four pedestrian actions are defined; *walking*, *crossing*, *turning*, and *riding a bicycle* (Figure 1). Intuitively, the pedestrian action categories are divided into different walking directions seen from an in-vehicle camera. Moreover, *riding a bicycle*, which is a confusing category by appearance, is added in the databases. We considered the two different scenarios with (i) a simple and static background for a pure performance test with a space-time representation (NTSEL), and (ii) dynamic and cluttered backgrounds to evaluate a performance in a practical situation (NDRDB). The summary of databases is shown in Table 1. The database description is described as follows.

**Table 1.** Summary of databases.

| DB | NTSEL Database; NTSEL | Near-Miss Driving Database; NDRDB |
|---|---|---|
| | #Video (#Frame) | #Video (#Frame) |
| #*Walking* | 25 (2648) | 15 (515) |
| #*Crossing* | 25 (2726) | 43 (1773) |
| #*Turning* | 25 (923) | 13 (593) |
| #*Riding a Bicycle* | 25 (1632) | 11 (457) |
| #Total | 100 (7929) | 82 (3338) |

### 3.1. NTSEL Database (NTSEL)

To evaluate a performance of space-time representation by excluding a background effect, we have experimentally collected traffic videos with four different fine-grained actions in the outside of the laboratory. The video database contains 100 videos in total and these are equally divided (25 videos) per category. The video database is captured in a simple and static background; therefore, we basically analyze how effective the space-time representation is on the NTSEL. The driving recorder is attached and the videos are captured from a static vehicle. In the database, three actors walked and rode a bicycle in front of the driving recorder. The distance is from 3.0 to 20.0 m. The *turning* is occurred at the 5.0, 10.0, 15.0 and 20.0 m distances.

### 3.2. Near-Miss Driving Recorder Database (NDRDB)

The society of automotive engineering of Japan (JSAE) is providing the Hiyari–Hatto database, which includes near-miss incidents [44]. The database contains video, GPS (global positioning system) and CAN (controller area network) data. We focused on pedestrian actions to analyze their fine-grained categorization in practical situations. Although the near-miss videos are difficult to collect, 82 videos have been collected in the database. The driving recorders are attached on a moving vehicle; therefore, four fine-grained actions are set—*walking*, *crossing*, *standing* and *riding a bicycle*. The moving camera makes the computer vision difficult with problems such as motion blur, and relative motion

between vehicle and pedestrian. The database contains 15 (walking), 43 (crossing), 13 (standing) and 11 (riding a bicycle) videos, respectively.

### *3.3. Difficulties of Self-Collected Databases*

By using the collected databases, there are three main difficulties with achieving the concept of fine-grained pedestrian action recognition:

- We should notice small changes in pedestrian actions, namely a meaningful change such as walking straight into turning should be recognized. In many cases, the change occurs in a moment. Therefore, a sophisticated descriptor is preferable to catch a subtle difference.
- An in-vehicle driving recorder is moving depending on the vehicle ego-motion (The NTSEL database does not include moving background. However, the database contains difficulties coming from a cluttered background and fine-grained pedestrian actions.). The fine-grained pedestrian action recognition should be done in a cluttered scene that contains complicated and moving backgrounds.
- The collection of pedestrian fine-grained action is difficult. A large amount of fine-grained pedestrian action data allows us to significantly understand fine-grained pedestrian actions from a vehicle-mounted driving recorder. Although the collection of such data is very difficult due to the rarity of action change such as turning in actual driving, we should treat a feature extraction and learning in a small-scale walking action database, with the aim of improving the avoidance of accidental situations. Therefore, we should consider how to learn about a strong model with a small-scale database.

## 4. Proposed Approach

Based on the two-stream fusion convnets [45], we have constructed an improved classifier in fine-grained pedestrian action recognition. The section shows the proposed architecture and a couple of improvements as follows:

**Architecture.** The proposed approach is shown in Figure 2. The basemodel, two-stream fusion convnets [45], supplies a better representation than conventional two-stream convnets [32] with fused convolutional maps and additional convolutions. In the fusion layer ("Fusion" in Figure 2), we can get a more sophisticated representation through several additional convolutions after layer-fusion between RGB-input and flow-input. Intuitively, the multiple modality analysis allows us to extract important features such as small changes of moving area in a pedestrian's sequence.

The basic architecture consists of two-stream fusion convnets [45] (convolutional maps; $m$), deep convolutional activation features (DeCAF) [46] (vector; $v$), and SVM (category; $c$) from inputs from RGB ($I_{rgb}$) and optical flow ($I_{flow}$).

We begin by calculating convolutional maps $m$ for a given videos $I_{rgb}$ and $I_{flow}$ in Equation (1):

$$m = f(I_{rgb}, I_{flow}; w), \tag{1}$$

where the function $f$ outputs convolutional maps, which are parameterized by convolutional kernels $w$. We then convert DeCAF $v$ from convolutional maps $m$ with linear function in fully-connected layer $g$ in Equation (2):

$$v = g(m). \tag{2}$$

Finally, a category $c$ is trained with an SVM model.

**DeCAF with fine-tuning architecture.** DeCAF is employed with the first fully-connected (fc) layer, which has 4096 dimensions, since we should train with small-scale databases. The DeCAF is known as an effective technique when there is no large-scale database. In our pre-experiment, a training could not converge an end-to-end training with two-stream fusion convnets. Moreover, the activation feature is improved with self-collected databases from an initial pre-trained UCF101 parameters [47,48]

through a fine-tuning training. We verified the effectiveness of fine-tuning in the experimental section. Table 2 shows the with or without fine-tuning with NTSEL and NDRDB databases.

**Unit adaptation.** An fc layer is verified with {128, 256, 512, 1024, 2048, and 4096} units. Basically, the original 4096 units based on an ImageNet model are assigned, but the number of dimensions should be fixed depending on the recognition problem. The above-mentioned fc units are compared in the experimental section. Table 3 describes performance rates with various fc units on NTSEL and NDRDB databases.

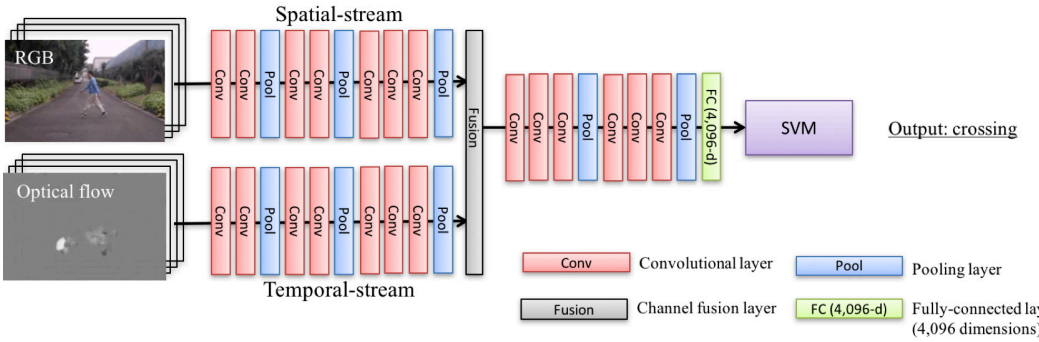


**Figure 2.** Flowchart of our proposed approach: Proposed architecture for fine-grained pedestrian action recognition. We assign two-stream fusion convnets [45] originally proposed by Feichtenhofer et al. The conventional work operates channel-sum with two different convolutional maps in an intermediate layer of spatial- and temporal-stream. After the channel fusion layer ("fusion" in the architecture), we add several convolutional and pooling layers (conv and pool) in order to generate a strong feature, e.g., subtle difference in walking pedestrian. In the classification step, we employ deep convolutional activation features (DeCAF; the first fully-connected layer (FC) with 4096-d vector) to converge the small-scale database by combining with support vector machines (SVM) [46]. Two-stream fusion convnets and DeCAF + SVM are trained with a training-set on self-collected databases.

**Table 2.** With or without fine-tuning.

| | NTSEL (%) | NDRDB (%) |
|---|---|---|
| End-to-End | N/A | N/A |
| Without fine-tuning (DeCAF) | 82.73 | 46.70 |
| With fine-tuning (DeCAF) | **88.73** | **51.77** |

**Table 3.** Various fc units on the self-collected databases.

| #Fc-Unit | NTSEL (%) | NDRDB (%) |
|---|---|---|
| 128 | 88.58 | 51.01 |
| 256 | 88.73 | 48.47 |
| 512 | 89.30 | 51.49 |
| 1024 | 86.30 | **53.23** |
| 2048 | 89.01 | 49.87 |
| 4096 | **91.01** | **53.23** |

## 5. Experiment

The section describes the experimental settings, results and discussion.

### *5.1. Implementation*

In the spatial-stream, the input was 224 pixels × 224 pixels × 3 channels. In the temporal-stream, a basic stacked flow [32] was implemented in order to create an input of 224 pixels × 224 pixels × 20 channels.

All initial learning rates were set to 0.001, and updating was set to a factor of 0.1 at each 1/2 and 3/4 of total learning epochs. The training is terminated maximum 50 epochs (We assigned a model which achieves the best rate). A high dropout ratio is set in each of the fully-connected layers (0.8–0.9 at each fc connection). The mini-batch size is 32 in the experiment.

We split training/testing sets into NTSEL and NDRDB databases. In NTSEL, we set 15 videos for training and 10 videos for testing. In NDRDB, we set 2/3 for training and others for testing. The training and testing splits are fixed for a fair evaluation.

### *5.2. Exploration Study*

We carry out a couple of tunings to improve fine-grained pedestrian action recognition. The fine-tuning, fc units and SVM parameter are adjusted in the section.

**With or without fine-tuning (see Table 2).** The results with fine-tuning on both databases are listed in Table 2. Starting from UCF101 pre-trained model by Wang [47], we fit into our NTSEL and NDRDB databases with fine-tuning training. In the situation, we use general purpose video features without fine-tuning, or the traffic video feature with fine-tuning after extracting vectors of DeCAF. As a result of fine-tuning, with fine-tuning is better than without one (+6.00% on NTSEL, +5.07%). The value describes that the traffic specified feature is effective for the problem of fine-grained pedestrian action recognition. A fine-tuned architecture, DeCAF and SVM make an effective configuration. Hereafter, the configuration is used in this experiment.

**Various fc units (see Table 3).** We confirmed that there is a suitable number of units at each database. Table 3 shows the relationship between #fc-unit and its performance rate. According to the table, it is suitable to use 4096-d (91.01%) on NTSEL and 1024/4096-d (53.23%) on NDRDB. 4096-d on NTSEL and 1024-d on NDRDB are assigned as a tuned parameter.

**SVM parameter (see Figure 3).** To fix a so-called "c-parameter" in SVM, we tuned several parameters at each database. We adopted the parameters as {0.01, 0.1, 1.0, 10, 100} in Figure 3. In Figure 3, we simultaneously decided the SVM parameters with representative approaches. The parameter depends on the problem, but it is effective for improving the performance rate. Finally, the performance rates have been increased +8.28% on NTSEL (82.73 to 91.01) and 6.53% on NDRDB (46.70 to 53.23).

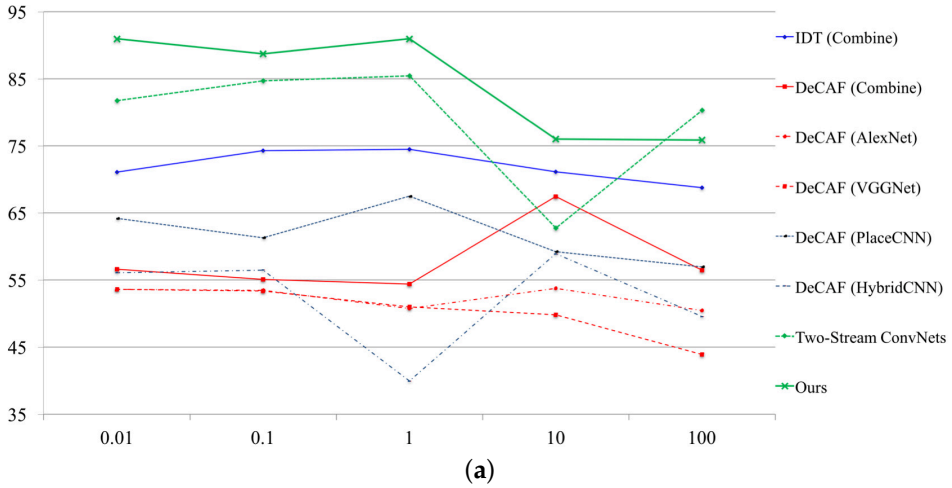


**Figure 3.** *Cont.*

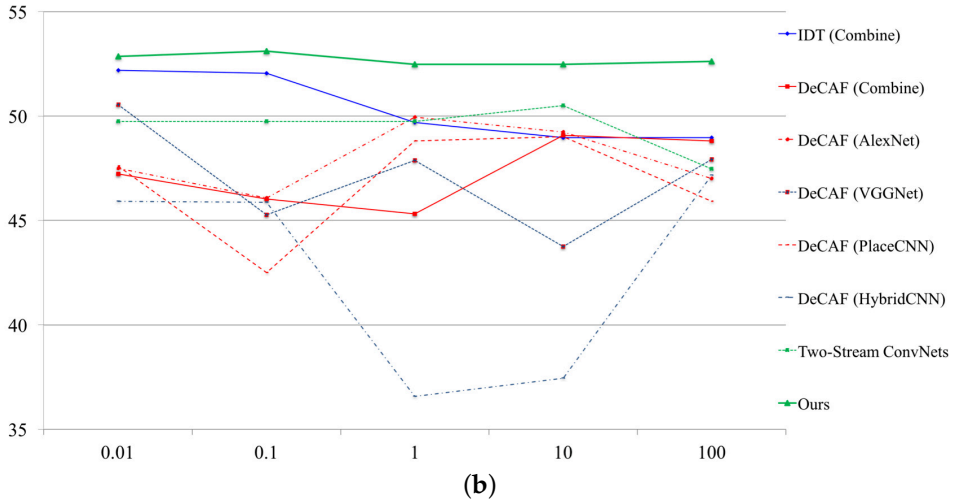


**Figure 3.** SVM parameter tuning. (**a**) relationship between performance rate and SVM parameter on NTSEL; (**b**) relationship between performance rate and SVM parameter on NDRDB.

### *5.3. Comparison of Representative Approaches*

We investigated the effectiveness of some motion representations. Here, IDT [30], DeCAF [46] and two-stream convnets [32] are employed. The abstract of the approaches is listed below:

**IDT.** IDT is the de-facto-standard hand-crafted model for video representation. The setting is based on the original implementation. To generate a codeword vector, motion boundary histograms (MBH) (192-d), histograms of optical flow (HOF) (108-d), and HOG (96-d) are captured each time a trajectory is sampled; they are incorporated into a feature vector.

**DeCAF.** Activation features were extracted based on the AlexNet/VGG-16. In the paper, we set fc6 for each CNN architecture. We used ImageNet pre-trained model (ImageNet, ImageNet with VGG-16) [6,16], Places205 pre-trained model (Places205) [49], and ImageNet + Places205 pre-trained model (HybridCNN) [49]. One more model, all combined (ImageNet, ImageNet with VGG-16, Places205, HybridCNN), is applied.

**Two-stream convnets.** We used two-stream CNN based on VGG-16 [47] by Wang et al. The pre-trained model is trained with UCF101, and additional training was performed on self-collected databases.

**Trajectory-pooled deep-convolutional descriptors (TDD) [33].** The TDD is at the intersection of the IDT and two-stream convnets. A large number of trajectories are sampled to access convolutional maps. Here, spatial convolutional maps are used.

The comparison with representative approaches is listed in Table 4. Our proposal achieved the top rate on NTSEL (91.01%) and a competitive rate on NDRDB (52.23%). We have improved two-stream fusion convnets with well-organized parameters. The model assigns both channels of RGB and flow to extract a feature from a minor change in moving area.

Undoubtedly, the approach is effective in NTSEL, which has a static background at each video. In the NTSEL, a major moving area is only pedestrian's walking; therefore, convnet-based methods with optical flows (ours and two-stream convnets) tend to have high-accuracy on NTSEL. The approaches such as DeCAF (67.48% with Places205) and spatial-stream (69.04%) struggle to enhance the feature for fine-grained pedestrian action recognition. Accuracy was not achieved even if we use a combined DeCAF (67.48%). The flow-based models including IDT can easily catch a suitable feature for classification. The combined IDT with HOG/HOF/MBH achieved 74.52% on NTSEL that is better rate than spatial-convnet; however, the noisy flows on background area are disturbing.

**Table 4.** The performance rates on the NTSEL & near-miss DR dataset.

| Approach | NTSEL (%) | NDRDB (%) |
|---|---|---|
| IDT (HOG) | 70.18 | 50.43 |
| IDT (HOF) | 64.76 | 52.05 |
| IDT (MBH) | 65.38 | 49.12 |
| IDT [30] | 74.52 | 52.19 |
| DeCAF (ImageNet) [16] | 53.78 | 49.94 |
| DeCAF (ImageNet with VGG-16) [6] | 53.63 | 50.54 |
| DeCAF (Places205) [49] | 67.48 | 49.02 |
| DeCAF (Hybrid) [49] | 58.91 | 47.17 |
| DeCAF (Combined) [46] | 67.44 | 49.07 |
| Two-stream ConvNets (Spatial) | 69.04 | 48.47 |
| Two-stream ConvNets (Temporal) | 64.05 | 45.93 |
| Two-stream ConvNets [32] | **85.44** | 50.50 |
| TDD [33] | 68.39 | **54.66** |
| Ours | **91.01** | **53.23** |

On one hand, fine-grained pedestrian action recognition on NDRDB is difficult due to the moving background, cluttered background and relatively small changes of pedestrian intention. Despite the difficult situation of video recognition, our proposal achieved the second best accuracy on NDRDB. Our two-stream fusion convnets significantly extracts good features by excluding the effects of moving background. We consider the effect is coming from complementation between RGB (texture) and flow (moving area) each other. Hand-crafted IDT recorded a good accuracy with trajectory-based representation. In NDRDB, TDD had the best accuracy with the combination of trajectory-based feature and deep convolutional representation. The method can receive both merits from hand-crafted features and deeply learned representations. The trajectory-based approach picked up background flows, but there are informative features such as ego-motion by a vehicle.

**Discussion.** According to the results on the experiments, the combined streams with spatial (RGB input) and temporal (optical flow input) channels should be implemented in fine-grained pedestrian action recognition. The combined representation simultaneously works with an effective feature extraction with both streams in additional convolutional layers (after fusion-layer in Figure 2), and noisy areas' exclusion. Moreover, we have improved a performance rate by fine-tuning and DeCAF. In the situation, an end-to-end is not converged with only a small-scale database. There are only $10^2$ ($10^3$)-order videos (frames) against the pre-training with UCF101 [48], which has 13,220 videos on the database. The other parameter tuning contributed to increasing the performance rates. Finally, we have improved +8.28% and +6.53% from a baseline.

The self-collected databases present an important issue of "fine-grained pedestrian action recognition" in addition to the feature improvement by fine-tuning. Towards a practical-level performance, we should try to improve the accuracy of the problem. The solution will be data collection and model updates from the current configurations.

In the database construction, the limitation of data variation should be treated in the future. For examples, we would like to extend our databases to deploy the pedestrian intention recognition at night. Recently, a couple of databases [50,51] are proposed with various sensors such as far infrared (FIR) and multispectral camera.

To improve the current two-stream fusion convnets, we would like to use hand-crafted knowledge like TDD. Many sampling points allow the model to recover a more sophisticated representation. Moreover, a possible combination is coming from a skeleton-based feature. For example, Fang et al. estimated and quantized pedestrian's skeletons to extract a stably spatiotemporal vector [52].

### *5.4. Visual Results*

We list several visual results on NTSEL in Figure 4. There are four different sequences in various actions from a couple of pedestrians. For all pedestrian sequences (the first three rows), the pedestrians are relatively small, but our proposal successfully recognized the action category. Especially in the turning actions (second and third rows), we can estimate a moment in advance. The recognition of turning allows us to predict the next action, e.g., crossing a street and walking straight.

However, in the last row, our proposal corrected riding a bicycle as walking/turning. The second half of the sequence contains a turning action while riding a bicycle, and this is a partially correct answer.

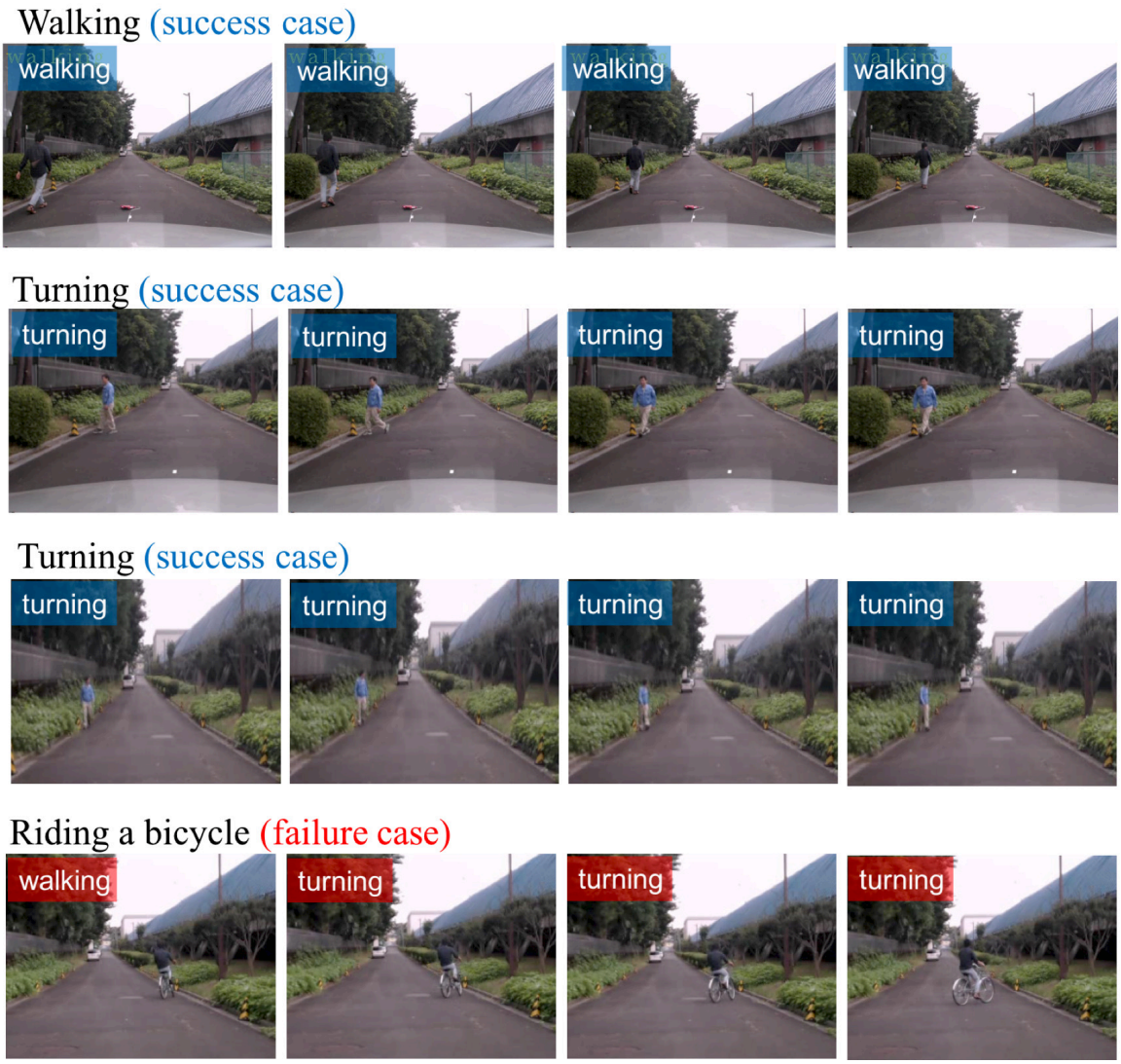


**Figure 4.** Visual results on NTSEL dataset: the first three lines, there are three success cases as the examples of walking and turnings. The last row shows the failure case in a sequence of a person is riding a bicycle. Especially in the second row, we succeeded with an estimation of pedestrian intention in advance. The turning walking action is important for a safety system.

## 6. Conclusions

The paper proposed an important issue, fine-grained pedestrian action recognition for an advanced safety system. The fine-grained pedestrian action recognition allows us to estimate a pedestrian's intention. We have collected two different databases: NTSEL and near-miss driving recorder database (NDRDB) to provide the recognition problem. The videos included in both databases are captured

by vehicle mounted driving recorders. The NTSEL is the simple setting with static background, but there are difficulties such as cluttered background and relatively small pedestrians in a video sequence. The NDRDB is more practical video data by moving in the background. We also proposed a improved convnet-based descriptor as a baseline with two-stream fusion convnets. We added a couple of modifications such as fine-tuning with self-collected databases and a fully-connected unit adaptation. Moreover, we assigned the deep convolutional activation features (DeCAF) since the databases have only $10^2$-order size in terms of videos. Due to the results of fine-grained pedestrian action recognition, our proposal achieved the top rate on NTSEL (91.01%) and the second best rate on NDRDB (53.23%). We revealed that the combined streams with RGB and flow inputs are important in order to be highly accurate recognition. Finally, we have increased +8.28% on NTSEL and +6.53% on NDRDB, respectively.

In the future, we need to improve the accuracy of practical settings (e.g., usage of NDRDB) by database collection and model updates.

**Acknowledgments:** This research has been supported by the AIST funding.

**Author Contributions:** Hirokatsu Kataoka has implemented the approaches and constructed the databases. He also wrote the paper, together with Yoshimitsu Aoki, Yutaka Satoh, Shoko Oikawa, Yasuhiro Matsui.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

| **Important Acronyms** | |
|---|---|
| DNN | Deep Neural Networks |
| CNN, convnet(s) | Convolutional Neural Networks |
| NTSEL | National Traffic Science and Environment Laboratory Database |
| NDRDB | Near-Miss Driving Recorder Database |
| Conv | Convolutional Layer |
| Pool | Pooling Layer |
| Fusion | Fusion Layer |
| FC | Fully-Connected Layer |
| DeCAF | Deep Convolutional Activation Features |
| SVM | Support Vector Machines |

## References

1. Geronimo, D.; Lopez, A.M.; Sappa, A.D.; Graf, T. Survey of Pedestrian Detection for Advanced Driver Assistance Systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1239–1258.
2. Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. Ten years of pedestrian detection, what have we learned? In Proceedings of the European Conference on Computer Vision Workshop (ECCVW), Zurich, Switzerland, 6–12 September 2014.
3. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
4. Viola, P.; Jones, M. Rapid Object Detection using a Boosted Cascaded of Simple Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, USA, 8–14 December 2001; pp. 511–518.
5. Felzenszwalb, P.; Girshick, R.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645.
6. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representation (ICLR), San Diego, CA, USA, 7–9 May 2015.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

8. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. How Far are We from Solving Pedestrian Detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
9. Zhang, L.; Lin, L.; Liang, X.; He, K. Is Faster R-CNN Doing Well for Pedestrian Detection? In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
10. Watanabe, T.; Ito, S.; Yokoi, K. Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection. In Proceedings of the 3rd Pacific-Rim Symposium on Image and Video Technology (PSIVT), Tokyo, Japan, 13–16 January 2009.
11. Kataoka, H.; Tamura, K.; Iwata, K.; Satoh, Y.; Matsui, Y.; Aoki, Y. Extended Feature Descriptor and Vehicle Motion Model with Tracking-by-detection for Pedestrian Active Safety. *IEICE Trans. Inf. Syst.* **2014**, 296–304, doi:10.1587/transinf.E97.D.296.
12. Dollar, P.; Tu, Z.; Perona, P.; Belongie, S. Integral Channel Features. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 7–10 September 2009.
13. Csurka, G.; Dance, C.R.; Fan, L.; Willamowski, J.; Bray, C. Visual Categorization with Bags of Keypoints. In Proceedings of the Workshop on Statistical Learning in Computer Vision (ECCVW), Prague, Czech Republic, 11–14 May 2004.
14. Jegou, H.; Douze, M.; Schmid, C.; Perez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010.
15. Perronnin, F.; Sanchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the European Conference on Computer Vision (ECCV), Heraklion, Greece, 5–11 September 2010.
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012.
17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
18. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
19. Girshick, R. Fast R-CNN. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015.
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
23. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
24. Du, X.; El-Khamy, M.; Lee, J.; Davis, L.S. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017.
25. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion Loss: Detecting Pedestrians in a Crowd. *arXiv* **2017**, arXiv:1711.07752.
26. Dalal, N.; Triggs, B.; Schmid, C. Human Detection using Oriented Histograms of Flow and Appearance. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006.
27. González, A.; Vázquez, D.; Ramos, S.; López, A.; Amores, J. Spatiotemporal Stacked Sequential Learning for Pedestrian Detection. In Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis, Santiago de Compostela, Spain, 17–19 June 2015.
28. Laptev, I. On Space-Time Interest Points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123.

29. Wang, H.; Klaser, A.; Schmid, C. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79.
30. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013.
31. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the International Conference on Computer Vision (ICCV), Los Alamitos, CA, USA, 7–13 December 2015.
32. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition. In Proceedings of the Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014.
33. Wang, L.; Qiao, Y.; Tang, X. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
34. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761.
35. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012.
36. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009.
37. Menze, M.; Geiger, A. Object Scene Flow for Autonomous Vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
38. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
39. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
40. Bai, M.; Luo, W.; Kundu, K.; Urtasun, R. Exploiting Semantic Information and Deep Matching for Optical Flow. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.
41. Luo, W.; Schwing, A.; Urtasun, R. Efficient Deep Learning for Stereo Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
42. Kundu, A.; Vineet, V.; Koltun, V. Feature Space Optimization for Semantic Video Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
43. First Workshop on Fine-Grained Visual Categorization. Available online: https://sites.google.com/site/cvprfgvc/ (accessed on 9 February 2018)
44. Hiyari-Hatto Database. Available online: http://web.tuat.ac.jp/~smrc/drcenter.html (accessed on 9 February 2018).
45. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
46. Donahue, J.; Jia, Y.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. DeCAF: A deep convolutional activation feature for generic visual recognition. In Proceedings of the International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014.
47. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y. Towards Good Practices for Very Deep Two-Stream ConvNets. *arXiv* **2015**, arXiv:1507.02159.
48. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Action Classes From Videos in the Wild. *arXiv* **2012**, arXiv:1212.0402.
49. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning Deep Features for Scene Recognition using Places Database. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014.

50. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
51. González, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; López, A.M. Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison. *Sensors* **2016**, *16*, 820, doi:10.3390/s16060820.
52. Fang, Z.; Vázquez, D.; López, A.M. On-Board Detection of Pedestrian Intentions. *Sensors* **2017**, *17*, 2193, doi:10.3390/s17102193.