applied sciences

MDPI

Article

Binary Dense SIFT Flow Based Position-Information Added Two-Stream CNN for Pedestrian Action Recognition

Sang Kyoo Park [1], Jun Ho Chung [1], Dong Sung Pae [2,*] and Myo Taeg Lim [1,*]

[1] School of Electrical Engineering, Korea University, Seoul 02841, Korea
[2] Department of Software, Sangmyung University, Cheonan 31066, Korea
* Correspondence: paeds915@smu.ac.kr (D.S.P.); mlim@korea.ac.kr (M.T.L.)

Abstract: Pedestrian behavior recognition in the driving environment is an important technology to prevent pedestrian accidents by predicting the next movement. It is necessary to recognize current pedestrian behavior to predict future pedestrian behavior. However, many studies have recognized human visible characteristics such as face, body parts or clothes, but few have recognized pedestrian behavior. It is challenging to recognize pedestrian behavior in the driving environment due to the changes in the camera field of view due to the illumination conditions in outdoor environments and vehicle movement. In this paper, to predict pedestrian behavior, we introduce a position-information added two-stream convolutional neural network (CNN) with multi task learning that is robust to the limited conditions of the outdoor driving environment. The conventional two-stream CNN is the most widely used model for human-action recognition. However, the conventional two-stream CNN based on optical flow has limitations regarding pedestrian behavior recognition in a moving vehicle because of the assumptions of brightness constancy and piecewise smoothness. To solve this problem for a moving vehicle, the binary descriptor dense scale-invariant feature transform (SIFT) flow, a feature-based matching algorithm, is robust in moving-pedestrian behavior recognition, such as walking and standing, in a moving vehicle. However, recognizing cross attributes, such as crossing or not crossing the street, is challenging using the binary descriptor dense SIFT flow because people who cross the road or not act the same walking action, but their location on the image is different. Therefore, pedestrian position information should be added to the conventional binary descriptor dense SIFT flow two-stream CNN. Thus, learning biased toward action attributes is evenly learned across action and cross attributes. In addition, YOLO detection and the Siamese tracker are used instead of the ground-truth boundary box to prove the robustness in the action- and cross-attribute recognition from a moving vehicle. The JAAD and PIE datasets were used for training, and only the JAAD dataset was used as a testing dataset for comparison with other state-of-the-art research on multitask and single-task learning.

Keywords: pedestrian-action recognition; two-stream convolutional neural network (CNN); binary descriptor dense scale-invariant feature transform (SIFT) flow; position-information feature

Citation: Park, S.K.; Chung, J.H.; Pae, D.S.; Lim, M.T. Binary Dense SIFT Flow Based Position-Information Added Two-Stream CNN for Pedestrian Action Recognition. *Appl. Sci.* **2022**, *12*, 10445. https://doi.org/10.3390/app122010445

Academic Editor: Alexandros A. Lavdas

Received: 26 September 2022
Accepted: 13 October 2022
Published: 17 October 2022

1. Introduction

In the past decade, autonomous vehicles [1–4] have been widely studied in terms of recognizing obstacles, such as road lines, other vehicles, or pedestrians, and planning the optimal path in the road environment. Although many studies have been conducted on path planning after line detection and on vehicle recognition [5–7], pedestrian recognition research has scarcely been conducted except regarding pedestrian detection [8–13]. The general pedestrian detection models on a typical road can cause an accident due to unexpected pedestrian behavior.

Autonomous vehicles must predict future pedestrian behavior to reduce pedestrian accidents. Pedestrian behavior cannot be predicted using the speed and position-based dynamic system but can be predicted by previous pedestrian behavior [14]. Thus, to predict

future pedestrian behavior, current pedestrian behavior must be recognized. However, according to a recent survey on recognizing pedestrian behavior [15], few studies have addressed pedestrian behavior recognition.

There has been considerable research on recognizing human-action datasets [16–18]. The OpenPose based convolutional neural network (CNN) has been used for spatial features [19,20] and long short-term memory (LSTM) has been used for temporal features [21]. In recent years, the flow-image-based two-stream CNN has been widely used to deal with spatial and temporal features [22–25]. Similar to human behavior, pedestrian behavior is also recognized by the OpenPose and LSTM [26–28]. The existing algorithms that recognize human behavior are difficult to apply in the driving environment with many obstacles such as buildings, trees or traffic poles due to the problematical conditions in the driving environment. Although there has been some research on the CNN and recurrent neural network (RNN), there is still no research on the flow-based two-stream CNN for pedestrian behavior recognition that is robust in moving vehicles in an outside environment.

The most widely used two-stream CNN is the optical flow-based algorithm [22]. However, the optical flow has limitations concerning the brightness constancy and piecewise smoothness assumptions, making recognizing pedestrian behavior using optical flow challenging in moving vehicles and outside environments [29,30]. Because of these limitations, most research on pedestrian behavior does not recognize pedestrian-action attributes (walking or standing) but recognizes cross attributes (crossing or not crossing) on the JAAD dataset. To overcome this problem, Park [25] proposed the binary descriptor dense SIFT flow two-stream CNN instead of optical flow.

Another problem in pedestrian behavior recognition is that the pedestrian has multiple attribute tasks, such as action attributes, cross attributes, and reaction attributes. There has been research on multitask learning on the JAAD dataset [31], but it is tough to recognize multitask attributes. Because cross-attribute recognition has high accuracy in single-task learning, it also has high accuracy in multitask learning. However, for action attributes, the accuracy is not high even in multitask learning because few studies have addressed single-task learning. Thus, to increase the performance of multitask learning, action-attribute recognition models with high accuracy in single-task learning should be used in multitask learning.

The main contributions of this paper are introduced as follows.

- Not only walking action attribute but also crossing attribute are recognized by using position-information added two stream CNN. The position-information stream increases performance of crossing attribute recognition especially.
- Multi-task learning which trains both walking and crossing attributes at once is adopted. Then performance of proposed method is compared by other single-task learning network.
- Although proposed method uses the 2D detecting and tracking algorithm than ground-truth boundary-box, the performance of behavior recognition is still high. Thus, it can be applied in a real-world environment.

In this paper, we propose related work to compare human-action recognition and pedestrian behavior recognition algorithms in Section 2. Then, the three contributions of this paper are introduced in Section 3. Section 4 evaluates the performance of the action and cross-attribute recognition with the position-information feature-image added two-stream CNN. Finally, the conclusions are presented in Section 5.

2. Related Work

2.1. Human-Action Recognition between the Ground-Truth Boundary Box and Detection Algorithm

Most researchers have used the ground-truth boundary box because the person is in the center of the image to focus on human actions. Thus, most researchers who use deep learning train models using the ground-truth boundary-box input images [26,32].

The model trained with the ground-truth boundary box exhibits high performance when evaluated with the testing dataset, which also consists of a ground-truth boundary box.

However, there is a limitation to applying a network model to a testing dataset obtained in a real-world environment. The image dataset in the real-world environment should be obtained using a detection algorithm, and the real-world dataset is different from the ground-truth boundary-box dataset because the boundary becomes blurred when the dynamically moving human is in a place where there is a change in light. To increase the evaluation accuracy of the dataset obtained in a real-world environment, some researchers have tried to fuse a model with a detection algorithm without the ground-truth boundary-box dataset [31]. However, there is still a problem that the accuracy of human-action recognition algorithm fused with the detection algorithm in previous research is low. In addition, human motion can be captured through the detection algorithm for an instant, but it cannot capture human action because actions are a set of motions over time. Thus, the tracker must be adopted after a detection algorithm to solve the problem of human-action recognition. In addition, an algorithm is needed that recognizes human behavior and is robust to changes in the background and illumination.

2.2. Human-Action Recognition with the Two-Stream CNN

Human action comprises a set of sequence motions according to time. Thus, motion-feature and time-flow information should be extracted to recognize human action. However, the general CNN is suitable for extracting motion-feature information but has difficulty extracting time-flow information. Moreover , the general RNN is suitable for extracting time-flow information but has difficulty extracting motion-feature information. To solve these problems, Simonyan and Zisserman [22] introduced the two-stream CNN, which is suitable for extracting motion-feature and time-flow information. In this algorithm, the flow information is a flow image obtained through the optical flow representing a change in human movement in continuous time sequence. The network model of the two-stream CNN uses optical flow images accumulated between N frames. The model is divided into spatial and temporal streams, each dealing with motion-feature and time-flow information concerning human actions through five convolutional layers and three fully connected layers [33].

Binary Descriptor Dense SIFT Flow-Based Two-Stream CNN

Optical flow has difficulty detecting human actions when the target moves drastically in an outdoor environment because of the optical flow theory formulation. The main assumptions of optical flow are brightness constancy and piecewise smoothness [29,30]. More specifically, brightness constancy assumes that the light in the image should be constant with little change, and piecewise smoothness assumes that a moving target should fit within the neighboring pixels with a small width. This limitation of optical flow makes it difficult to recognize action when detecting a pedestrian from a moving vehicle in an outdoor environment. To solve this problem, SIFT flow [34] (a feature-based matching algorithm) and the binary descriptor dense SIFT flow [25] were developed to reduce the computational cost of the SIFT flow. Figure 1 is a JAAD dataset sample of a moving pedestrian from a moving vehicle. Figure 1a depicts the RGB image of the moving pedestrian over time. Because the pedestrian and background move when a moving vehicle detects a moving pedestrian, the optical flow image has the limitation that the distinction between the background and pedestrian is unclear, as illustrated in Figure 1b. However, the binary descriptor dense SIFT flow can clearly distinguish the pedestrian from the background, as displayed in Figure 1c. Research has been conducted on the two-stream CNN [25] using a binary descriptor dense SIFT flow image, which is robust in extracting pedestrian features in a moving-vehicle environment. Like the optical flow-based two-stream CNN [22], the binary descriptor dense SIFT flow-based two-stream CNN also is a CNN-M-2048 [33] model with five convolutional layers and three fully connected layers.

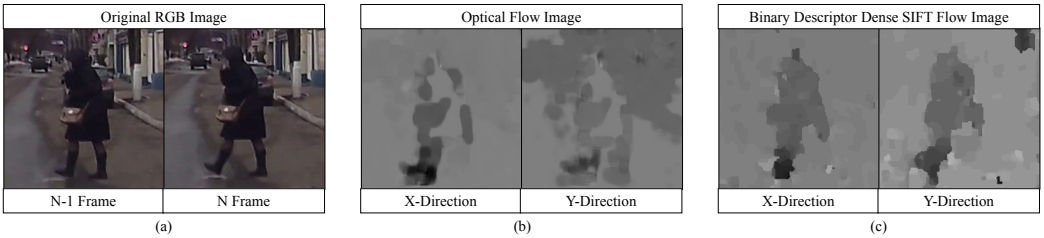

Figure 1. Comparison between optical flow and binary descriptor dense SIFT flow: (**a**) Serial original RGB images, (**b**) x- and y -direction optical flow images, and (**c**) x- and y-direction binary descriptor dense SIFT flow images.

3. Position-Information Feature Added Two-Stream CNN with Pedestrian Detector and Tracker

The proposed algorithm, the position-information feature added two-stream CNN with detection and tracking, is introduced in this section. The system flow configuration is presented in Figure 2. The inputs are consist of first frame for detector and N frames for tracker. Total $N + 1$ frames of video extract the dense SIFT flow image and position-information image. Then RGB, dense SIFT flow and position-information image are learned by position-information feature added two-stream CNN as an input. When pedestrian changes the behavior in the medium term, the input frames are chasing new behavior well before N frames, because the pedestrian must change the behavior with preparation action in a few frame. If the N is large, the proposed method cannot chase the new action when the pedestrian changes the action in the medium term. Additionally, if the N is small, the performance of pedestrian behavior detection would be low. To maximize the performance and the robustness to changes, the value of N is chosen as 10.

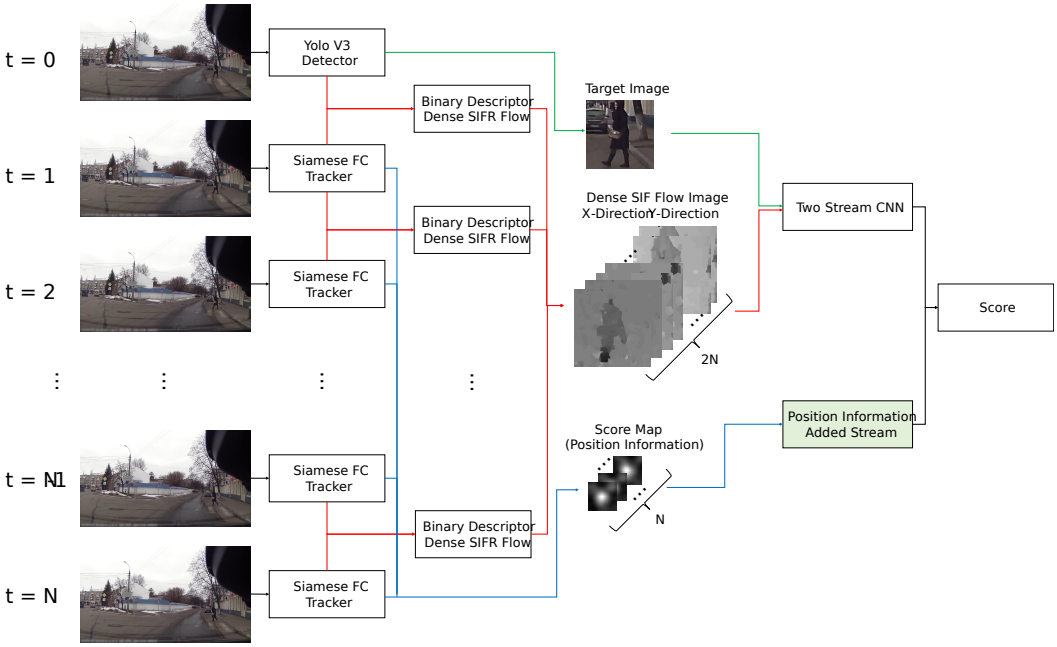

Figure 2. Position-information feature added two-stream CNN structure in a time sequence.

In the first section, the pedestrian-action dataset was applied to binary descriptor dense SIFT flow-based two-stream CNN to improve the performance of action-attribute recognition. As pedestrian actions comprise a set of motions over time, there is a limit to recognizing pedestrian actions with a simple current motion scene. The two-stream CNN is an algorithm that analyzes a set of flow images that indicate how pedestrian motion changes according to the time sequence [22]. In this paper, instead of optical flow, which is a flow image used in the conventional two-stream CNN, a binary descriptor dense SIFT flow, which is robust in complex motion and light changes, is used [25]. In the second section, the

position-information feature added two-stream CNN, which improved the conventional two-stream CNN, is used to increase the performance of cross-attribute recognition.

The cross attributes of a pedestrian are independent of the action attributes, such as the movement speed of the pedestrian, but are dependent on where the pedestrian is located [35]. Thus, the position-information feature of the pedestrian is expressed as a position-information feature image, and the position-information stream is fused with the conventional two-stream CNN. In the last section, we used the detector and tracker algorithm without the ground-truth boundary box to check whether the proposed algorithm applies to a moving vehicle in an actual outdoor environment. As most existing pedestrian behavior recognition studies operate algorithms based on ground-truth boundary boxes, there are limitations to their application in real-world vehicle environments.

3.1. Dense SIFT Flow-Based Two-Stream CNN for Action-Attribute Recognition

The two-stream CNN using a flow image with change features in pedestrian motion over time is applied to recognize the action attributes of a pedestrian. The network model structure of pedestrian-action recognition is the same as Figure 3 and different from human-action recognition in that the input is changed from a human-action dataset to a pedestrian-action dataset [25]. Similar to the human-action dataset, the pedestrian-action dataset also employs binary descriptor dense SIFT flow images using RGB images for N frames. Because the characteristics of the two datasets are similar, the two-stream CNN exhibits good performance in recognizing pedestrian action. However, the two-stream CNN has weak performance in cross-attribute recognition. Flow images extract pedestrian behavior features, whereas the two-stream CNN does not extract pedestrian cross-attribute features. Therefore, even if the pedestrian behavior is the same as walking action, the additional stream of extracting position features should be added to compare with whether the pedestrian is on the road or the sidewalks. Thus, this paper proposes the position-information feature-image added two-stream CNN, which has an additional stream compared to the conventional two-stream CNN.

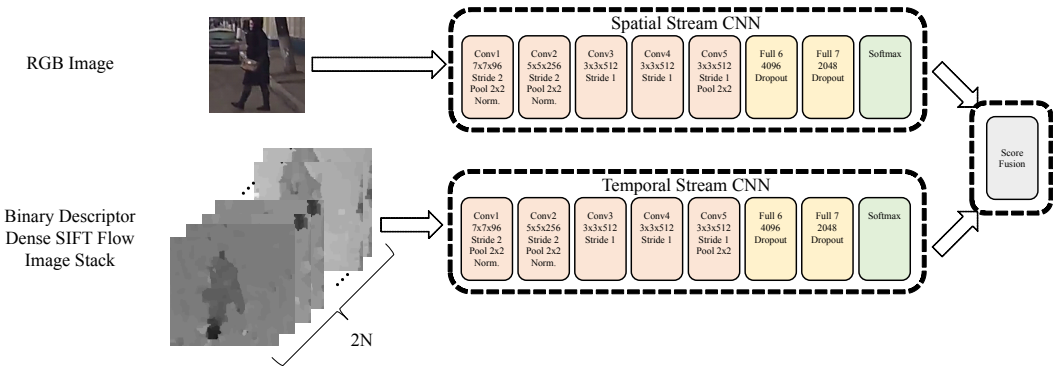

Figure 3. Binary descriptor dense SIFT flow-based two-stream CNN.

3.2. Position-Information Feature-Image Added Two-Stream CNN for Cross-Attribute Recognition

Whether pedestrians cross at road is independent of their behavior and depends on their position [35]. For example, if a walking pedestrian is located in the center of the image, the pedestrian is considered to be crossing the road because the pedestrian is walking on the crossroad. If the pedestrian is located on the horizon edge of the image, the pedestrian is not considered to be crossing the road because the pedestrian is walking on the sidewalk. Therefore, information on where the pedestrian is located in the image is required to recognize the cross attribute. This paper defines this information on pedestrian crossing as a position-information feature image. The position-information feature image is generated using a two-dimensional Gaussian filter applied to the area where the pedestrian is located in the RGB image, as illustrated in Figure 4.

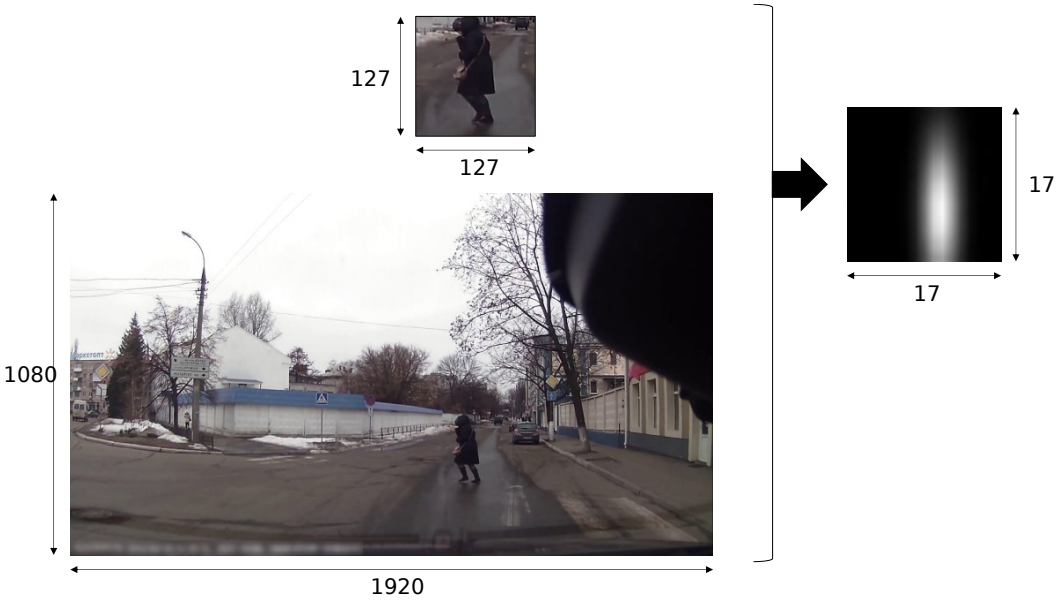

Figure 4. Position-information feature image representing the pedestrian position in the RGB image.

The improved network model of the conventional two-stream CNN is presented in Figure 5. A short stream called a position stream is added to apply the position-information feature image to the conventional two-stream CNN that does not degrade the performance of action-attribute recognition and increases the performance of cross-attribute recognition. As the input image of the network model, one RGB image enters the spatial stream, and N frames in the x- and y-direction of binary descriptor dense SIFT flow images enter the temporal stream similar to the conventional two-stream CNN. Then, the N frames of the position-information feature images also enter the position stream. The first convolutional layer is the same as the fifth convolutional layer of the other stream, and three fully connected layers are the same as the other fully connected layers to avoid increasing the number of parameters as much as possible.

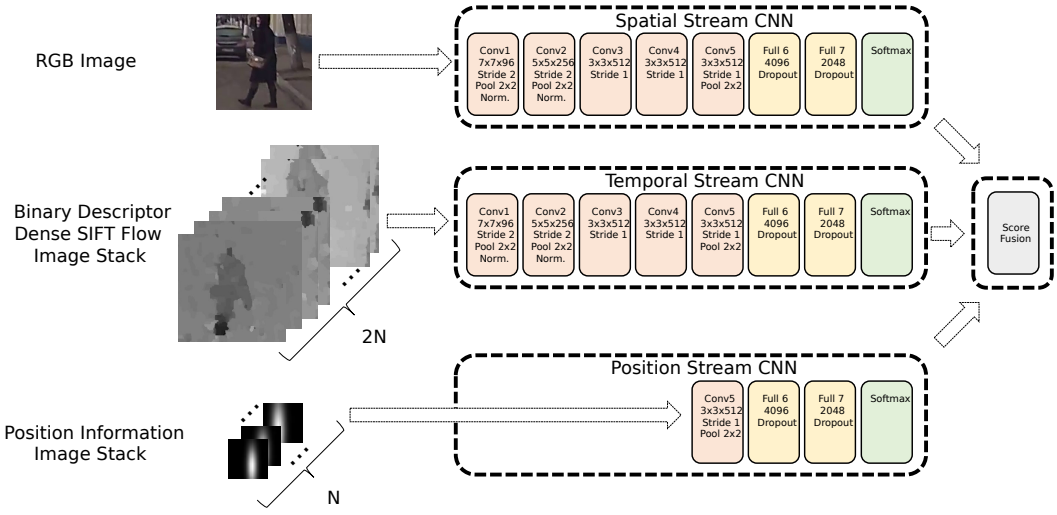

Figure 5. Position-information feature-image added two-stream CNN structure.

3.3. Fusion with a Detector and Tracker without a Ground-Truth Boundary Box

Recently, research on pedestrian-action recognition has trained the model by cropping the pedestrian image based on the ground-truth boundary box. However, the network model trained using the ground-truth boundary box has limitations in applying it to a vehicle driving in an actual outdoor environment. Research on the network model trained by cropping the image of the pedestrian through a detection algorithm without using the

ground-truth boundary box has been studied recently to verify the performance of the recognition algorithm even in an actual driving vehicle [31].

In this paper, YOLO v3 [36] was selected to confirm that the performance of the proposed algorithm is guaranteed when applied to a moving vehicle in a real-world environment. A pedestrian action is a set of motions over time; thus, the detection algorithm should be continuously executed for the serial time of N frames. However, because the time and memory cost are inefficient when detection is performed per frame, detection is used only in the first frame, and the tracking algorithm is applied to the other frames [37].

In this case, the Siamese fully convolutional tracker [38] was selected because the Siamese tracker has a score map representing the target position in the image that can be used as the position-information feature image, as depicted in Figure 6. The position-information feature image generated with the Siamese tracker is used as input to the position-information feature added binary descriptor dense SIFT flow two-stream CNN, as presented in Figure 7.

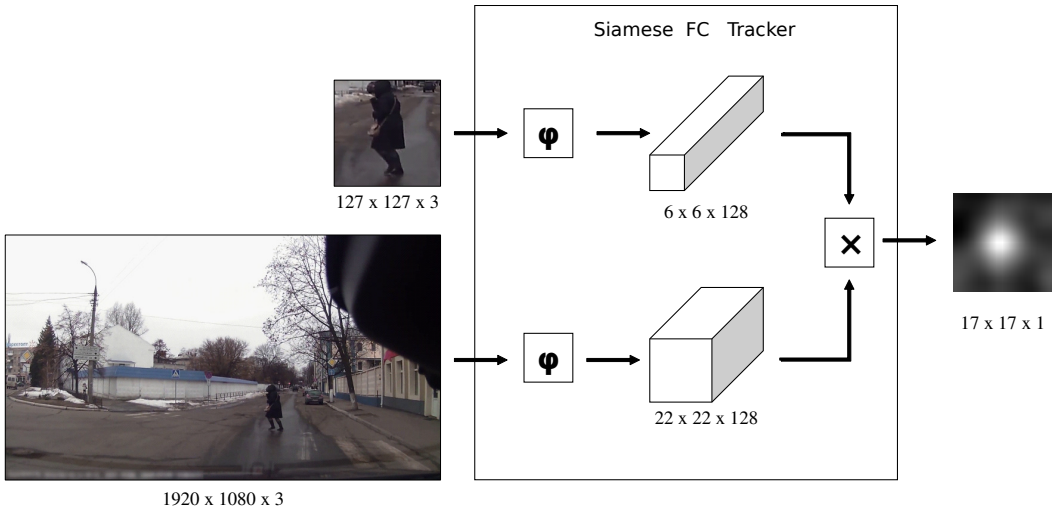

Figure 6. Position-information feature image obtained using a score map of the Siamese tracker output.

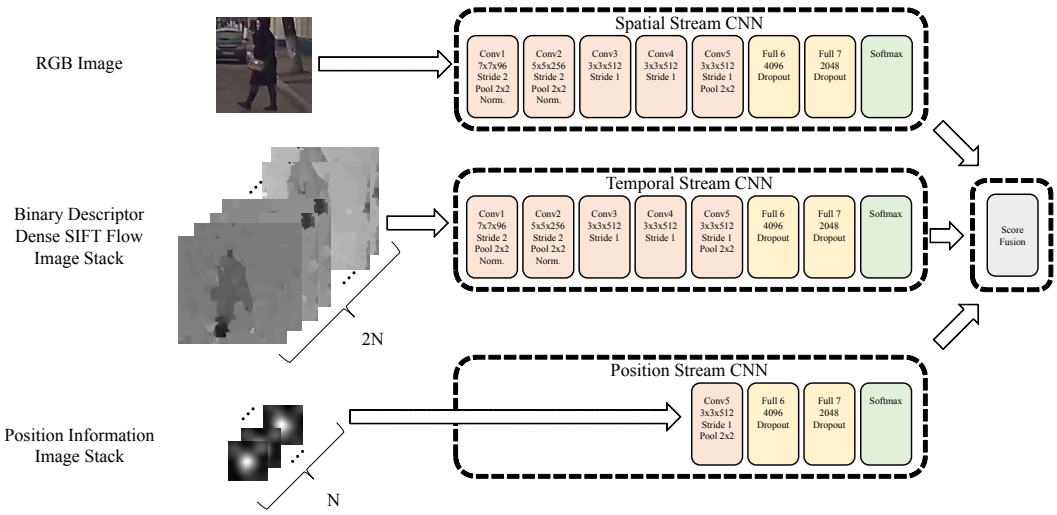

Figure 7. Position-information feature-image added two-stream CNN structure with a Siamese tracker score map.

JAAD and PIE Datasets

Datasets: Many datasets classify human behavior, but few datasets classify pedestrian behavior [15]. Although there are few datasets, the representative datasets of pedestrian behavior are listed in Table 1. The dataset with the greatest number of frames is the STIP dataset [39], but this dataset only has cross attributes and no action attributes. Therefore, the STIP dataset is inappropriate for evaluating the proposed method. The JAAD [40] dataset is currently the most widely used in pedestrian behavior recognition research and consists of five attributes. Each attribute is composed as follows: action attributes (walking and standing), cross attributes (crossing and not crossing), look attributes (looking and not looking), hand-gesture attributes (greet, yield, and right of way), and reaction attributes (clear path, speed up, and slow down). The numbers of training and testing samples expected from the JAAD dataset are provided in Table 2, where a sample is selected when the same attribute appears for 10 frames in a series.

As presented in Table 2, the look, hand-gesture, and reaction attributes are too biased toward one label. Therefore, only action and cross attributes are used for pedestrian behavior recognition. However, because the number of JAAD samples is few for training the proposed method, PIE [41], created in JAAD creator, was also used. Unlike the JAAD dataset, because PIE has no distinction between the training and testing datasets, the entire PIE dataset was used as the training dataset. Thus, the training dataset consists of the JAAD and PIE datasets, whereas only the JAAD dataset was used for evaluation with other research as the testing dataset. Finally, the dataset is listed in Table 3.

Multitask Learning: As presented in Table 3, the action and cross attributes were selected for pedestrian behavior recognition. Recent research has focused on single-task learning, which recognizes only action attributes [26] or only cross attributes [39,42–44]. In this paper, multitask learning was used to train two attributes simultaneously. All cases that can be composed of two attributes is defined as a class. Each attribute consists of two classes; thus, the entire class consists of four and is as follows: walking and crossing, walking and not crossing, standing and crossing, standing and not crossing. Afterward, multitask learning was implemented by dividing the four class outputs during evaluation. After arranging the dataset, the standing and crossing classes were unavailable because there are no people who is standing while crossing the road; therefore, only three classes were used to train the model.

Table 1. Datasets of pedestrian-action recognition.

Dataset	Action	Cross	Look	Hand Gesture	Reaction	No. of Frames
JAAD [40]	O	O	O	O	O	82,032
PIE [41]	O	O	O	O	X	909,480
STIP [39]	X	O	X	X	X	1,108,176

Table 2. Data sample per class in the JAAD dataset about 5 attributes.

Classes		Training Dataset (No. of Samples)	Testing Dataset (No. of Samples)
Action	Standing	529	339
	Walking	1211	1071
Cross	Not crossing	988	794
	Crossing	752	616
	Irrelevant	0	0

Table 2. *Cont.*

Classes		Training Dataset (No. of Samples)	Testing Dataset (No. of Samples)
Hand gesture	Undefined	1729	1405
	Greet	0	0
	Yield	0	0
	Right of way	0	0
	Other	11	5
Reaction	undefined	1655	1321
	Clear path	24	29
	Speed up	33	33
	Slow down	28	27
Look	Not looking	1460	1176
	Looking	280	234

Table 3. Data sample per class in the JAAD dataset and PIE dataset about action and cross attirubtes.

		Action		Cross	
		Walking	Standing	Crossing	Not Crossing
Training Dataset	JAAD	529	1211	988	752
	JAAD and PIE	30,255	27,317	12,490	45,082
Testing Dataset	JAAD	3796	561	2541	1816

4. Experimental Evaluations for Action Recognition

4.1. Experiment Environment

Implementation Detail

Pretraining: Before learning pedestrian actions using the JAAD and PIE datasets, pretraining was performed to recognize human actions using the UCF-101 dataset [45]. The spatial and temporal streams of the two-stream CNN were trained 40k times with UCF-101, and then the JAAD and PIE datasets were trained with fine-tuning.

Fine-tuning: Pretrained parameters were used for each of the spatial and temporal streams, and fine-tuning was performed with the JAAD and PIE datasets by adding the position stream to train the position-information feature added two-stream CNN. The momentum optimizer with a learning rate of 0.0001 and 0.9 momentum was used. For the input image, the 256 by 256 image was reduced to 224 by 224 using random cropping and flipping, and the input was received in a batch size of 128. Afterward, the network model learned through 50K iterations.

Learning Environment: The deep learning platform of the proposed method was Tensorflow. The graphics processing unit (GPU) was NVIDIA V100.

4.2. Ablation Analysis of Pedestrian-Action Recognition

4.2.1. Evaluation of the Binary Descriptor Dense SIFT Flow-Based Two-Stream CNN

Tables 4 and 5 compare using only the JAAD dataset and using both the JAAD and PIE datasets. The deep learning model used in this evaluation is the conventional two-stream CNN with binary descriptor dense SIFT flow, not the optical flow. First, as indicated in Table 4, due to the inequality of the number of samples in the JAAD dataset between the walking and standing classes, the accuracy of the standing class using only the JAAD dataset is low. Moreover, due to the fusion of the JAAD and PIE datasets, the imbalance in quantity between the walking and standing classes becomes more balanced; thus, it has

higher performance than using only the JAAD dataset. Second, as listed in Table 5, the overall accuracy of cross-attribute recognition improved.

Table 4. Test accuracy of pedestrian-action recognition using the conventional dense SIFT flow two-stream CNN.

	Walking Accuracy	Standing Accuracy	Average Accuracy
JAAD	92.44%	52.80%	82.91%
JAAD and PIE	91.01%	72.38%	88.70%

Table 5. Test accuracy of pedestrian cross-attribute recognition using the conventional dense SIFT flow two-stream CNN.

	Crossing Accuracy	Not Crossing Accuracy	Average Accuracy
JAAD	80.64%	60.68%	71.16%
JAAD and PIE	86.12%	65.64%	76.44%

Table 6 compares the accuracy of the position-information feature-image added two-stream CNN with the detector and tracker between the binary descriptor dense SIFT flow and optical flow. Because action-attribute recognition is affected by the quality of the flow image, there is a difference in the performance of action-attribute recognition. As depicted in Figure 1, the binary descriptor dense SIFT flow image, which is a feature-based matching algorithm, has a clearer distinction between the background and target than the optical flow, which assumes brightness constancy and piecewise smoothness. In contrast, cross-attribute recognition, which is affected by the position-information feature, has little difference in performance between the optical flow image and binary descriptor dense SIFT flow image.

Table 6. Test accuracy of the position-information feature-image added two-stream CNN with the detector and tracker between the optical flow and binary descriptor dense SIFT flow.

	Action			Cross		
	Walking	Standing	Average	Cross	Not Cross	Average
Optical Flow	86.90%	61.59%	84.68%	75.73%	63.22%	71.69%
Binary Descriptor Dense SIFT Flow	91.01%	72.38%	88.70%	86.12%	65.64%	76.44%

4.2.2. Evaluation of the Position-Information Feature-Image Added Two-Stream CNN

Table 7 compares the accuracy of the binary descriptor dense SIFT flow-based two-stream CNN with and without the position-information feature-image added stream. From the experimental results, the position-information feature-image added stream is adopted to improve the performance of cross-attribute recognition while maintaining the performance of action-attribute recognition.

Table 7. Comparison before and after adding the position-information feature-image stream.

	Action Attribute			Cross Attribute		
	Walking	Standing	Average	Cross	Not Cross	Average
Two-Stream CNN	91.01%	72.38%	88.70%	86.12%	65.64%	76.44%
Two-Stream CNN with Position stream	91.49%	72.35%	88.75%	86.43%	71.04%	79.13%

4.2.3. Evaluation of Fusion with a Detector and Tracking without the Ground-Truth Boundary Box

Table 8 compares the accuracy between using the ground-truth boundary box and detection and tracking algorithms without ground-truth boundary box. The proposed method trained using the ground-truth boundary box shows high accuracy for the action and cross attributes. In contrast, the proposed method using the detection and tracking algorithm without the ground-truth boundary box has inaccurate position-information features of the sample image. Therefore, there is no significant difference in accuracy for the action attributes, which are independent of the position-information feature. However, the cross attributes, which are dependent on the position-information feature, reveal a significant difference in accuracy which is lower in using detecting and tracking algorithm. Although the cross-attribute accuracy of the proposed method using the detection and tracking algorithms is lower than that with the ground-truth boundary box, the accuracy is guaranteed to be over 70%. The output image of the position-information added two-stream CNN is illustrated in Figure 8.

Table 8. Comparison of the model with and without the ground-truth boundary box.

	Action Attribute			Cross Attribute		
	Walking	**Standing**	**Average**	**Cross**	**Not Cross**	**Average**
Proposed Method with Ground-Truth Boundary Box	91.49%	72.35%	88.75%	86.43%	71.04%	79.13%
Proposed Method with Detecting and Tracking alogorithm	90.87%	72.23%	88.47%	74.50%	70.63%	72.90%

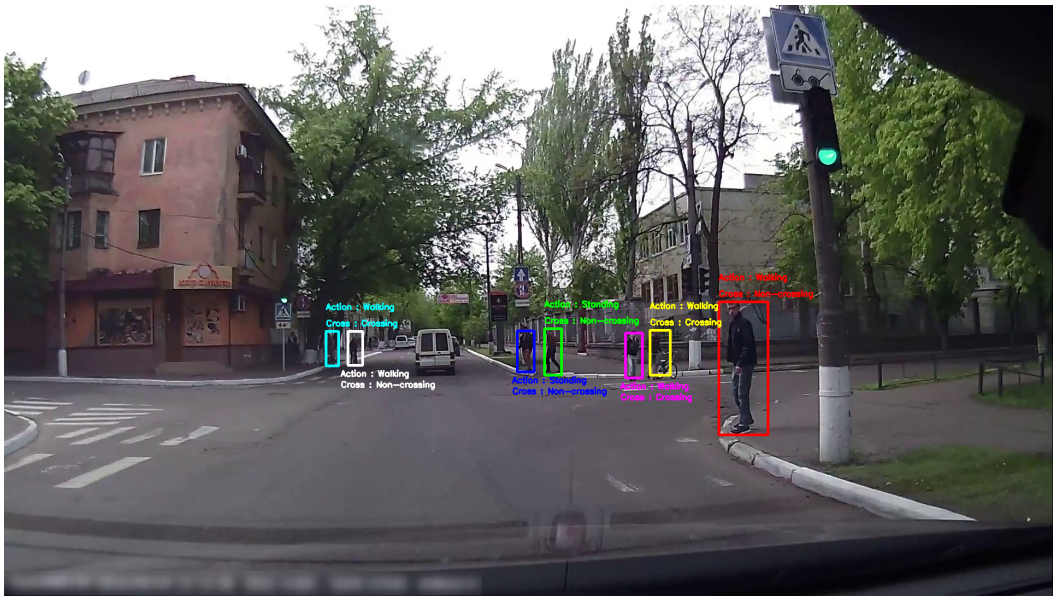

Figure 8. Image result of the position-information added two-stream CNN with the YOLO v3 detector and the Siamese fully convolutional tracker.

4.3. Comparison of the Experimental Results with State-of-the-Art Methods of Multitask Learning

Table 9 reveals the results of the comparison with other multitask learning research based on the testing dataset provided by the JAAD dataset. To the best of our knowledge, only one researcher [31] has explored multiattribute learning with detection, so we could only compare the model with Mordan's method. Mordan's method uses detection to determine the application results in a real-world environment without using the ground-

truth boundary box. However, because actions are a sequence of motions over time, there is a limitation to recognizing an action attribute through one frame. Moreover , Park's method [25] demonstrates high performance in action-attribute recognition because of the binary descriptor dense SIFT flow image accumulated along the time axis. The proposed method, the position-information feature-image added two-stream CNN, has higher performance than Park's method for cross-attribute recognition because of the position-stream network. Finally, the fusion of the position-information feature-image added two-stream CNN, YOLO v3 detector [36], and the Siamese tracker [38] demonstrates lower performance than before their merger. In particular, the model reveals substantial performance degradation for cross-attribute recognition because cross attributes are affected by the exact position of the position-information feature image. However, the proposed method still performs much better than Mordan's method.

Table 9. Comparison of state-of-the-art in multitask learning algorithm with the JAAD testing dataset.

Method	Ground-Truth Boundary Box	Action	Cross
Mordan [31]	X	29.90%	60.20%
Two-stream CNN (Park [25])	O	88.70%	76.44%
Proposed Method	O	88.75%	79.13%
Proposed Method with Detecting and Tracking alogorithm	X	88.47%	72.90%

4.4. Comparison of the Experimental Results with State-of-the-Art Methods for Single-Task Learning

Table 10 compares the results with other action-attribute recognition methods using single-task learning research based on the testing dataset (the JAAD dataset). To the best of our knowledge, only one study [26] has addressed action-attribute recognition with single-task learning. Because Park's method [25] is optimized for action-attribute recognition, it has higher performance on action-attribute recognition than Marginean's method [26], which is a single-frame recognition model. Even when the detector and tracker are added to the proposed algorithm, the model performs better than the state-of-the-art algorithm using the ground-truth boundary box.

Table 10. Comparison of pedestrian-action recognition for state-of-the-art in single-task learning algorithm with the JAAD testing dataset.

Name	Ground-Truth Boundary Box	Action Accuracy
Marginean [26]	O	79.73%
Two-Stream CNN (Park [25])	O	88.70%
Proposed Method	O	88.75%
Proposed Method with Detecting and Tracking alogorithm	X	88.47%

Table 11 presents the comparison with other cross-attribute recognition methods with single-task learning research based on the testing dataset (the JAAD dataset). Contrary to the recent papers introduced above, several studies on cross-attribute recognition have been conducted. Because cross attributes depend on the pedestrian position, most single-task learning and ground-truth boundary box–based research results have an accuracy of around 80%. Because the proposed algorithm focuses on pedestrian-action recognition, cross-attribute recognition is lower than for other research. However, the proposed method based on the detector and tracker exhibits a relatively high performance of 72.90%.

Table 11. Comparison of pedestrian cross-attribute recognition for state-of-the-art in single-task learning algorithm with the JAAD testing dataset.

Name	Ground-Truth Boundary Box	Cross Accuracy
Pop [42]	O	61.31%
Liu [39]	O	79.28%
Marginean [26]	O	81.00%
Wang [27]	O	81.23%
Chaabane [43]	O	86.70%
Fang [28]	O	88.00%
Singh [46]	O	84.89%
Two-Stream CNN (Park [25])	O	76.44%
Proposed Method	O	79.13%
Proposed Method with Detecting and Tracking alogorithm	X	72.90%

5. Conclusions

In this paper, we introduced the position-information feature-image added two-stream CNN for pedestrian behavior recognition with multitask learning and a detection and tracking algorithm. In multitask learning, to increase the performance of action-attribute recognition, the two-stream CNN is used based on the binary descriptor dense SIFT flow, not the optical flow. Then, to increase the performance of cross-attribute recognition, the position-information feature image generated by a Siamese tracker is added to the two-stream CNN. By increasing the action and cross attributes, the proposed method has higher performance than state-of-the-art methods on the JAAD dataset. Moreover, the proposed method guarantees the high accuracy of pedestrian-action recognition in vehicle-moving urban environments by using the YOLO detection and Siamese tracker instead of the ground-truth boundary box. Thus, the proposed algorithm can be adopted for various applications in autonomous vehicles.

In recent years, autonomous driving vehicles have begun to appear, and many studies have been conducted on pedestrian safety. Predicting pedestrian behavior has become an issue in saving pedestrian lives, but there is a problem because pedestrian behavior cannot be predicted with a dynamic model. It is necessary to understand and recognize the current pedestrian behaviors to predict future pedestrian behavior. In this paper, we show that the performance of pedestrian behavior recognition should be high with various attributes. Until now, the proposed method has limitation in that it can only recognize the walking and crossing road attributes not crossing intersection. Image can distinguish the pedestrian who crosses the road horizontally but does not cross the road vertically. Therefore, there is still room for improvement of recognizing the intersection crossing. It is necessary to study the vertical crossing attribute, then we have studied 3D detection by using point cloud which would be fused with proposed networks. The proposed method can facilitate saving pedestrians by adopting the advanced driver assistance system (ADAS)application. In addition, the proposed method can be adopted not only autonomous vehicle but also vehicle-to-infrastructure (V2I). For example, traffic sign pole alert warns the pedestrian who acts the strange behavior or sends the information of pedestrian behavior to the autonomous vehicle. This future technology can be applied at the urban planning policy or street design. We hope this research promotes further research on pedestrian behavior recognition for safety.

Author Contributions: Conceptualization, S.K.P. and M.T.L.; Data curation, S.K.P.; Formal analysis, S.K.P., D.S.P. and M.T.L.; Methodology, S.K.P., J.H.C., M.T.L. and D.S.P.; Software, S.K.P. and D.S.P.; Validation, M.T.L. and D.S.P.; Writing—original draft, S.K.P. and M.T.L.; Writing—review and editing, M.T.L. and D.S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT) (No. 2022R1F1A1073543).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The JAAD and PIE datasets presented in the study are openly available in the JAAD article [40] and PIE article [41].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Betz, J.; Zheng, H.; Liniger, A.; Rosolia, U.; Karle, P.; Behl, M.; Krovi, V.; Mangharam, R. Autonomous vehicles on the edge: A survey on autonomous vehicle racing. *IEEE Open J. Intell. Transp. Syst.* **2022**, *3*, 458–488. [CrossRef]
2. Badue, C.; Guidolini, R.; Carneiro, R.V.; Azevedo, P.; Cardoso, V.B.; Forechi, A.; Jesus, L.; Berriel, R.; Paixao, T.M.; Mutz, F.; et al. Self-driving cars: A survey. *Expert Syst. Appl.* **2021**, *165*, 113816. [CrossRef]
3. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A survey on 3d object detection methods for autonomous driving applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [CrossRef]
4. Marzbani, H.; Khayyam, H.; To, C.N.; Quoc, D.V.; Jazar, R.N. Autonomous vehicles: Autodriver algorithm and vehicle dynamics. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3201–3211. [CrossRef]
5. Wang, Z.; Zhan, J.; Duan, C.; Guan, X.; Lu, P.; Yang, K. A review of vehicle detection techniques for intelligent vehicles. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
6. Singhal, N.; Prasad, L. Sensor based vehicle detection and classification-a systematic review. *Int. J. Eng. Syst. Model. Simul.* **2022**, *13*, 38–60. [CrossRef]
7. Maity, S.; Bhattacharyya, A.; Singh, P.K.; Kumar, M.; Sarkar, R. Last Decade in Vehicle Detection and Classification: A Comprehensive Survey. *Arch. Comput. Methods Eng.* **2022**. [CrossRef]
8. Zhang, H.; Pop, D.O.; Rogozan, A.; Bensrhair, A. Accelerate High Resolution Image Pedestrian Detection with Non-Pedestrian Area Estimation. *IEEE Access* **2021**, *9*, 8625–8636. [CrossRef]
9. Ren, J.; Niu, C.; Han, J. An IF-RCNN Algorithm for Pedestrian Detection in Pedestrian Tunnels. *IEEE Access* **2020**, *8*, 165335–165343. [CrossRef]
10. Cai, J.; Lee, F.; Yang, S.; Lin, C.; Chen, H.; Kotani, K.; Chen, Q. Pedestrian as Points: An Improved Anchor-Free Method for Center-Based Pedestrian Detection. *IEEE Access* **2020**, *8*, 179666–179677. [CrossRef]
11. Wei, C.; Hui, F.; Yang, Z.; Jia, S.; Khattak, A.J. Fine-grained highway autonomous vehicle lane-changing trajectory prediction based on a heuristic attention-aided encoder-decoder model. *Transp. Res. Part Emerg. Technol.* **2022**, *140*, 103706. [CrossRef]
12. Claussmann, L.; Revilloud, M.; Gruyer, D.; Glaser, S. A review of motion planning for highway autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1826–1848. [CrossRef]
13. Liao, J.; Liu, T.; Tang, X.; Mu, X.; Huang, B.; Cao, D. Decision-making Strategy on Highway for Autonomous Vehicles using Deep Reinforcement Learning. *IEEE Access* **2020**, *8*, 177804–177814. [CrossRef]
14. Tsotsos, J.K.; Kotseruba, I.; Rasouli, A.; Solbach, M.D. Visual attention and its intimate links to spatial cognition. *Cogn. Process.* **2018**, *19*, 121–130. [CrossRef] [PubMed]
15. Chen, L.; Ma, N.; Wang, P.; Li, J.; Wang, P.; Pang, G.; Shi, X. Survey of pedestrian action recognition techniques for autonomous driving. *Tsinghua Sci. Technol.* **2020**, *25*, 458–470. [CrossRef]
16. Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; Luo, B. Pedestrian attribute recognition: A survey. *Pattern Recognit.* **2022**, *121*, 108220. [CrossRef]
17. Brehar, R.D.; Muresan, M.P.; Mariţa, T.; Vancea, C.C.; Negru, M.; Nedevschi, S. Pedestrian street-cross action recognition in monocular far infrared sequences. *IEEE Access* **2021**, *9*, 74302–74324. [CrossRef]
18. Yang, B.; Zhan, W.; Wang, P.; Chan, C.; Cai, Y.; Wang, N. Crossing or not? Context-based recognition of pedestrian crossing intention in the urban environment. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 5338–5349. [CrossRef]
19. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 914–927. [CrossRef] [PubMed]
20. Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; Del Bimbo, A. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. Cybern.* **2014**, *45*, 1340–1352. [CrossRef]
21. Pienaar, S.W.; Malekian, R. Human activity recognition using LSTM-RNN deep neural network architecture. In Proceedings of the 2019 IEEE 2nd Wireless Africa Conference (WAC), Pretoria, South Africa, 18–20 August 2019; pp. 1–5. [CrossRef]
22. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.

23. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with enhanced motion vector CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2718–2726. [CrossRef]
24. Zhao, Y.; Man, K.L.; Smith, J.; Siddique, K.; Guan, S.U. Improved two-stream model for human action recognition. *EURASIP J. Image Video Process.* **2020**, *2020*, 1–9. [CrossRef]
25. Park, S.K.; Chung, J.H.; Kang, T.K.; Lim, M.T. Binary Dense SIFT Flow Based Two Stream CNN for Human Action Recognition. *Multimed. Tools Appl.* **2021**, *80*, 35697–35720. [CrossRef]
26. Marginean, A.; Brehar, R.; Negru, M. Understanding pedestrian behaviour with pose estimation and recurrent networks. In Proceedings of the 2019 6th International Symposium on Electrical and Electronics Engineering (ISEEE), Galati, Romania, 18–20 October 2019; pp. 1–6. [CrossRef]
27. Wang, Z.; Papanikolopoulos, N. Estimating pedestrian crossing states based on single 2D body pose. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 2205–2210. [CrossRef]
28. Fang, Z.; López, A.M. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 4773–4783. [CrossRef]
29. Black, M.J.; Anandan, P. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comput. Vis. Image Underst.* **1996**, *63*, 75–104. [CrossRef]
30. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 25–36. [CrossRef]
31. Mordan, T.; Cord, M.; Pérez, P.; Alahi, A. Detecting 32 Pedestrian Attributes for Autonomous Vehicles. *arXiv* **2020**, arXiv:2012.02647.
32. Pop, D.O.; Rogozan, A.; Chatelain, C.; Nashashibi, F.; Bensrhair, A. Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction. *IEEE Access* **2019**, *7*, 149318–149327. [CrossRef]
33. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
34. Liu, C.; Yuen, J.; Torralba, A. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 978–994. [CrossRef] [PubMed]
35. Zhao, J.; Li, Y.; Xu, H.; Liu, H. Probabilistic prediction of pedestrian crossing intention using roadside LiDAR data. *IEEE Access* **2019**, *7*, 93781–93790. [CrossRef]
36. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
37. Luo, H.; Xie, W.; Wang, X.; Zeng, W. Detect or track: Towards cost-effective video object detection/tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Atlanta, Georgia, 8–12 October 2019; Volume 33, pp. 8803–8810. [CrossRef]
38. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 850–865. [CrossRef]
39. Liu, B.; Adeli, E.; Cao, Z.; Lee, K.H.; Shenoi, A.; Gaidon, A.; Niebles, J.C. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3485–3492. [CrossRef]
40. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 206–213. [CrossRef]
41. Rasouli, A.; Kotseruba, I.; Kunic, T.; Tsotsos, J.K. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6262–6271. [CrossRef]
42. Pop, D.O. Detection of pedestrian actions based on deep learning approach. *Stud. Univ. Babeş-Bolyai. Informatica.* **2019**, 64, 5–13. [CrossRef]
43. Chaabane, M.; Trabelsi, A.; Blanchard, N.; Beveridge, R. Looking ahead: Anticipating pedestrians crossing with future frames prediction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2297–2306. [CrossRef]
44. Rasouli, A.; Rohani, M.; Luo, J. Pedestrian Behavior Prediction via Multitask Learning and Categorical Interaction Modeling. *arXiv* **2020**, arXiv:2012.03298.
45. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
46. Singh, A.; Suddamalla, U. Multi-input fusion for practical pedestrian intention prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2304–2311. [CrossRef]