TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 02/10 pp458–470
DOI: 10.26599/TST.2019.9010018
Volume 25, Number 4, August 2020

Survey of Pedestrian Action Recognition Techniques for Autonomous Driving

Li Chen, Nan Ma*, Patrick Wang, Jiahong Li, Pengfei Wang, Guilin Pang, and Xiaojun Shi

Abstract: The development of autonomous driving has brought with it requirements for intelligence, safety, and stability. One example of this is the need to construct effective forms of interactive cognition between pedestrians and vehicles in dynamic, complex, and uncertain environments. Pedestrian action detection is a form of interactive cognition that is fundamental to the success of autonomous driving technologies. Specifically, vehicles need to detect pedestrians, recognize their limb movements, and understand the meaning of their actions before making appropriate decisions in response. In this survey, we present a detailed description of the architecture for pedestrian action recognition in autonomous driving, and compare the existing mainstream pedestrian action recognition techniques. We also introduce several commonly used datasets used in pedestrian motion recognition. Finally, we present several suggestions for future research directions.

Key words: autonomous driving; pedestrian action recognition; action datasets; two-stream network

1 Introduction

With the development of Artificial Intelligence (AI) over recent years, more and more research has applied this technology to autonomous driving, and autonomous driving has become a typical application scenario of AI. Self-driving vehicle comprehensively utilizes intelligent devices, such as multi-sensors, cameras, and laser radars, distributed in different parts of the car body to sense the surrounding environment, and the intelligent driving assistance software to form a driver-centered driving model, which controls the car behavior through planning, decision-making, and control[1].

An autonomous driving system must be able to perceive moving obstacles, such as other vehicles and pedestrians, ahead of the vehicle. Sensing of the surrounding environment is achieved intelligently through traditional or deep learning algorithms, and serves the process of planning and decision-making that allows the vehicle to decelerate or stop appropriately.

Pedestrian action recognition in autonomous driving provides an important guarantee for the safe and stable operation of intelligent driving systems in complex and uncertain environments. It requires not only that vehicles detect pedestrians, but also that they recognize the movements of pedestrians, accurately judge their intentions and make decisions accordingly.

2 Pedestrian Action Recognition Process in Autonomous Driving

The main steps of pedestrian movement recognition in autonomous driving are pedestrian detection and action recognition. The system must detect the area

• Li Chen, Nan Ma, and Jiahong Li are with Beijing Key Laboratory of Information Service Engineering, College of Robotics, Beijing Union University, Beijing 100101, China. E-mail: chenli6450@163.com; xxtmanan@buu.edu.cn; jqrjiahong@buu.edu.
• Patrick Wang is with Northeastern University, Boston, MA 02115, USA. E-mail: patwang@ieee.org.
• Pengfei Wang is with the Communication and Information Center of Ministry of Emergency Management of the People's Republic of China, Beijing 100013, China, E-mail: feipengwang767@163.com.
• Guilin Pang and Xiaojun Shi are with the College of Robotics, Beijing Union University, Beijing 100101, China. E-mail: 15866864406@163.com; 1720665483@qq.com.
* To whom correspondence should be addressed.
Manuscript received: 2019-03-08; revised: 2019-04-26; accepted: 2019-05-05

where pedestrians may exist in the scene, and judge whether there are pedestrians present. A Region Of Interest (ROI) that may contain pedestrians is extracted, human bodies are detected within this ROI, and then the features of any detected pedestrians are extracted. A specific action recognition procedure, which identifies the acquired action according to agreed content, is performed on the result of the extraction, so that the computer can understand the pedestrian's intention and make an appropriate decision in response.

In accordance with this two-step process, this paper summarizes the relevant research in the field and analyzes a variety of studies. Pedestrian detection is the basis of this research, as whether pedestrians can be detected efficiently and features extracted accurately greatly affects the overall accuracy and robustness of action recognition experiments. Action recognition requires that smart cars can understand the meaning of pedestrian actions, analyze their behavioral intentions, output the result, and achieve an appropriate human–car interaction by making decisions in response.

2.1 Pedestrian detection in autonomous driving

Pedestrian detection in autonomous driving can be regarded as a branch of object detection, with a pedestrian or person representing the tag for detection. Pedestrian detection methods can be divided into two major types: those using traditional algorithms and those using algorithms based on deep learning. Typical examples of traditional algorithms include the Histogram of Oriented Gradients (HOG) features and Support Vector Machines (SVM) approach, the Deformable Parts Model (DPM) and pedestrian detection based on action features. Pedestrian detection algorithms based on deep learning include the standard Regional Convolutional Neural Network (R-CNN) detection framework enhanced with region proposals and CNN classification, as in Faster R-CNN. Another detection method is to convert the problem into an end-to-end target detection framework for regression problems; this method is typified by You-Only-Look-Once (YOLO).

2.1.1 Traditional algorithms

HOG feature and SVM algorithm. Illumination can affect the robustness and stability of algorithms used for pedestrian detection in autonomous driving in real-world scenarios. The HOG proposed by Dalal and Triggs[2] is an image descriptor for human detection that greatly improves the robustness of pedestrian detection in regard to lighting. The edge and gradient features extracted by HOG can describe the local shape well and effectively reduce the influence of small illumination offsets on the detection of human bodies. Therefore, HOG has been widely used for human detection and gesture recognition, as well as other application.

The HOG feature extraction process can be divided into the following steps:

- Input image;
- Perform gamma color correction to reduce the influence of lighting and background on the target;
- Calculate the gradient of each pixel, capture contour information, and further weaken the influence of illumination;
- Calculate the gradient histogram of cell units by dividing the whole target window into non-overlapping cell units of the same size, then calculate the gradient information for each cell separately;
- Combine several cells into a block (e.g., 3×3) and normalize the gradient histograms in these blocks to reduce the influence of contrast between foreground and background and local light on the gradient amplitude; and
- Combine all block histogram vectors in the image into an overall HOG feature vector.

In summary, the detector window is tiled with a grid of overlapping blocks, each containing a grid of spatial cells. For each cell, the weighted vote of image gradients in orientation histograms is calculated. These are locally normalized and collected into a single feature vector[3].

After extracting the characteristics of pedestrians using HOG algorithm, an SVM is used for training and classification to achieve pedestrian detection. SVM was first proposed by Cortes and Vapnik in 1995[4] and is widely used in machine learning due to its simple structure and strong generalization ability[5,6]. SVM is used to solve the binary classification problem. Researchers use SVM to solve the non-linear problem by introducing the kernel method: by replacing the core, different separation surfaces can be obtained, thereby achieving classification.

Since HOG features can reduce the influence of illumination and offset on detection, Dalal[3] first attempted to apply HOG to human detection. Since then, many scholars have fruitfully expanded on the HOG and SVM method for pedestrian detection purposes. For example, Liu et al.[7] combined the advantages of AdaBoost and SVM by proposing a cascaded human detection framework based on the Related HOG (RHOG) features of both classifiers, and achieved good detection

results. Although it is a successful pedestrian detection algorithm, it takes a long time to execute. Pang et al.[8] proposed two ways to improve on this: one is to reuse the features in the block to construct the HOG feature of the cross-detection window, another is to use an interpolation based on sub-blocks to efficiently calculate the HOG features of each block. Thereby, speed is improved while guaranteeing the accuracy of detection. Han et al.[9] proposed a robust detection method for humans and vehicles in static images based on an extended HOG and SVM approach, realizing robust detection by a two-stage approach, the first stage focuses on attention generation and the second stage fulfills hypothesis verification.

DPM algorithm. Pedestrian detection based on HOG and SVM has achieved good results in many experiments. However, when the human body is occluded, the accuracy of pedestrian detection based on HOG and SVM decreases. Some scholars have proposed using the DPM for pedestrian detection[10], which has improved the accuracy in cases where a pedestrian is partially obscured. The DPM algorithm can be seen as an extension of HOG and the processing flow is basically the same. The difference is that the humanoid model trained by HOG is a single model, which is effective for vertical front and rear pedestrian detection, but less effective when pedestrians only have side information. The DPM algorithm improves performance in this situation by using multiple models for detection, introducing a root model and component models (head, shoulder, arm, etc.) approach to improve the efficiency of the algorithm in detecting moving pedestrians. The DPM process is roughly shown in Fig. 1[10].

In relation to DPM, Yan et al.[11] proposed a multi-task model for pedestrian detection in response to the problem of low image resolution affecting pedestrian detection efficiency. This model includes a resolution-sensitive transformation, which places pedestrians with

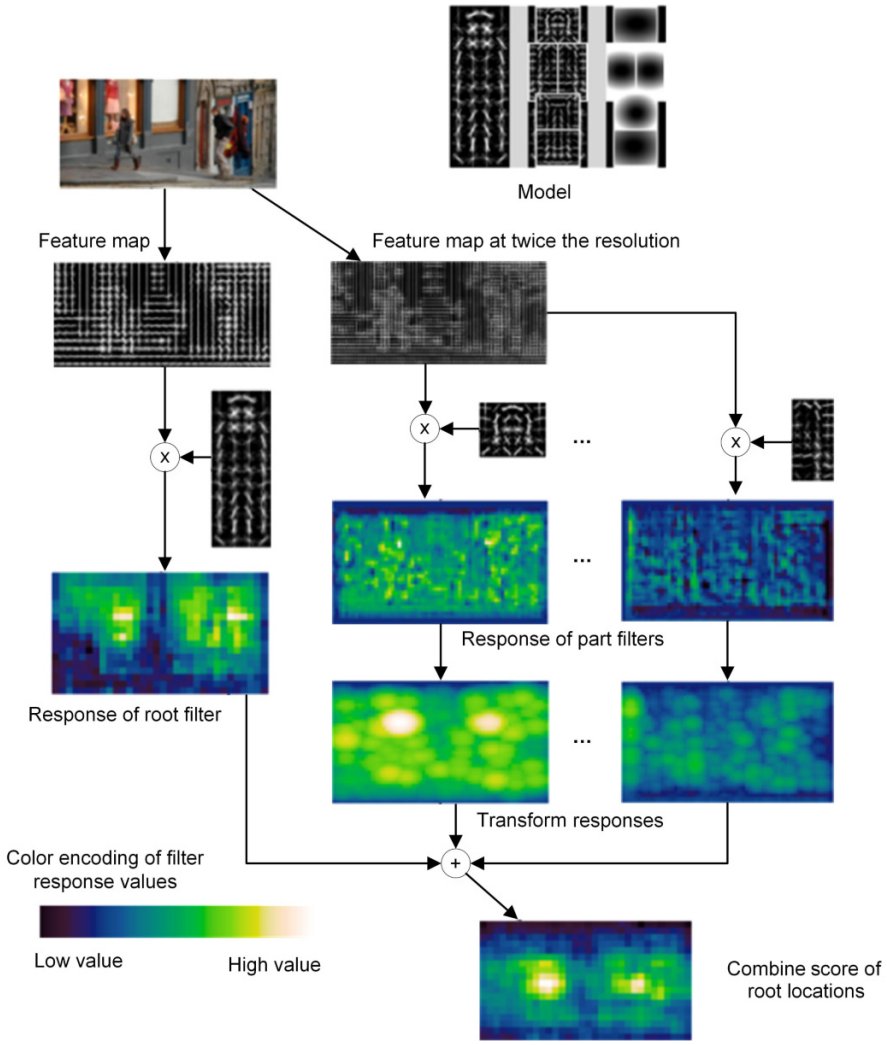


Fig. 1 Matching process at one scale.

different resolutions in the same space and constructs a general-purpose detector to distinguish pedestrians from the background. For model learning, Yan et al.[11] proposed a coordinate reduction method to iteratively learn a resolution-aware transformer and detector based on DPM, which has achieved good detection results. Zeng and Xiao[12] proposed a pedestrian detection method for traffic scenes, combining the single and double DPM models. This method first extracts the DPM features of the training samples from pedestrian datasets, such as INRIA and ETH, and trains the single and double DPM models through the Latent SVM method. Pedestrians in traffic scenes are then classified into two categories: separately distributed and mixed. Pedestrians can be effectively detected even when they are occluded. Furthermore, Yan et al.[13] solved the speed bottleneck problem of the DPM while maintaining the detection accuracy for complex datasets, and Tian et al.[14] proposed a spatiotemporal DPM, which studied the generalization of the DPM from 2D images to 3D spatiotemporal volumes and showed a strong robustness on several video datasets.

Pedestrian detection based on motion features. In response to the fact that most pedestrians are in motion, some researchers have recently sought to detect pedestrians by identifying motion features. For static scenes, a representative algorithm was proposed by Viola et al.[15] in 2005, which puts forward a pedestrian detection system based on the calculation of Haar-like features in different images and the combination of gradation information with motion information. This method can detect human bodies in low-resolution images featuring rain and snow, but it has poor applicability to occluded pedestrians. For dynamic scenes, Dalal and Triggs[2] proposed to construct a pedestrian detector by combining appearance description and optical flow characteristics to realize pedestrian detection in motion video. However, this method can only detect pedestrians in a single window and the detection effect over the entire image is relatively poor.

2.1.2 Deep learning algorithms

Two-stage method. With the development of deep learning over recent years, an increasing number of researchers are applying deep learning techniques to pedestrian detection in autonomous driving, thereby greatly improving detection accuracy compared with the traditional methods. R-CNN was the first attempt to use the deep learning method for pedestrian detection, proposed by Girshick et al.[16], it transforms the target detection problem into a classification problem. This method first generates a number of region proposals in the image, uses the convolutional network to extract the target features for each region proposal, and then uses SVM to train a classifier. Finally, it obtains a target boundary according to each region classification score through the non-maximum suppression optimized algorithm.

A CNN is used to extract features, which requires the size of the input picture to be fixed at 224×224. Therefore, after selecting the candidate box using the selective search method, the image needs to be preprocessed with a warp or crop, which takes some time. In order to improve the accuracy of pedestrian detection, SPP-net[17] was introduced so that the CNN can process pictures of any size. The core of Spatial Pyramid Polling (SPP) uses multiple sliding window pooling (refer to the blue 4×4, blue 2×2, and grey 1×1 windows in Fig. 2) to sample feature maps obtained from the upper layer (convolution layer). Max-pooling is used, and the fixed-length output can be obtained by combining the results obtained separately. The SPP layer performs three pooling operations on feature maps obtained from the previous convolution layer, such that each feature map becomes a fixed feature map. The SPP layer generally solves the problem that the input image must be a fixed size. By pooling and aggregating a feature map of different sizes, the accuracy and robustness of the algorithm are improved and the slowness of R-CNN is also solved to a certain extent.

In addition, the convolution network for R-CNN feature extraction and the network used for classification need to be trained separately, which demands some time and storage space for the training process. Meanwhile, the training of the classifier is not related to the feature extraction network, which affects the accuracy of object detection. Therefore, Girshick proposed a Fast Regional Convolutional Neural Network (Fast R-CNN) model[18], which integrates feature extraction and classification into a classification framework and improves the speed of training the model and the accuracy of object detection.

However, Fast R-CNN uses the selective research algorithm[19] to generate region proposals separately, which is very time consuming and makes the algorithm not executable in real time. Ren et al.[20] added a Region Proposal Network (RPN) to Fast R-CNN to generate region proposals, and constructed an end-to-end Faster R-CNN, which greatly improved on the speed of the operation. RPN adds a sliding window and two

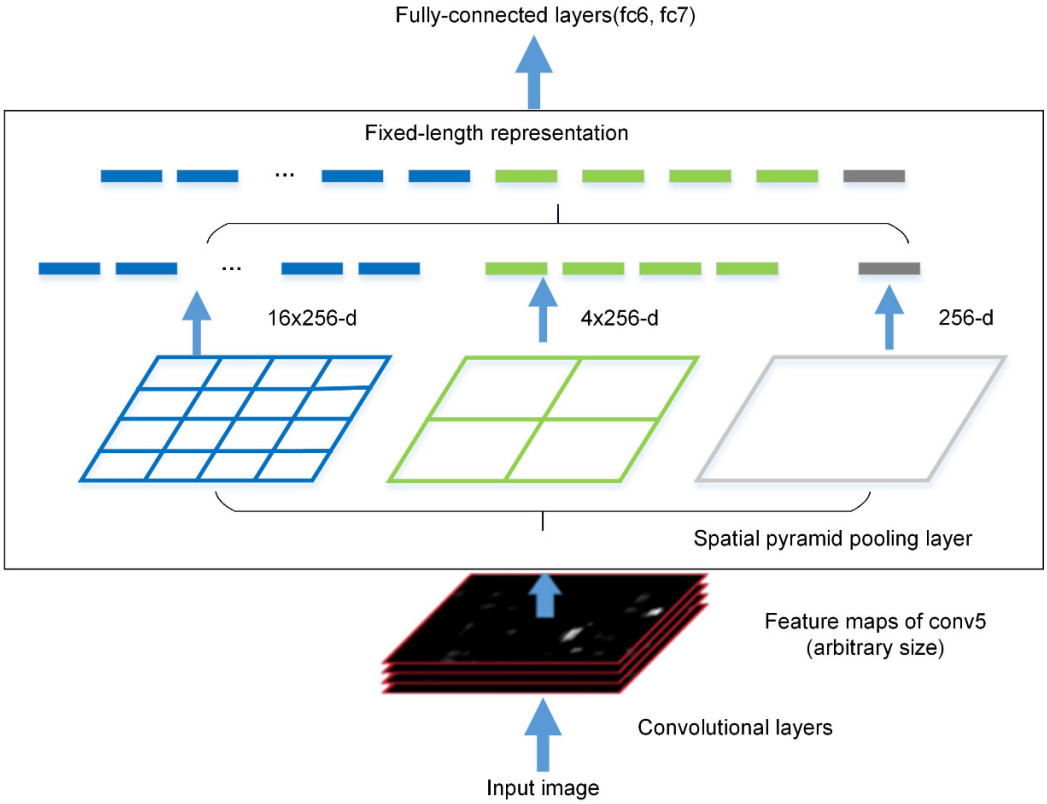


Fig. 2 SPP layer, here 256 is the filters number of the conv5 layer, and conv5 is the last convolutional layer.

convolutional layers after the CNN convolutional layer to obtain the region proposals. The first convolutional layer encodes each sliding window position of the feature map into a feature vector. The second convolutional layer corresponds to each sliding window. It outputs k region scores and k region proposals after regression for each sliding window position and a score of Top-N (where $N = 300$ in the study) after the non-maximum suppression of the score region. The second layer also tells the detection network which areas it should pay attention to, and essentially realizes the functions of selective research, EdgeBoxes and other methods.

In Ref. [21], Faster R-CNN is used to realize pedestrian detection in night infrared images. RPN is used to generate region proposals and Fast R-CNN is used to extract features and to classify and refine positions. Due to the convolutional layer of RPN and Fast R-CNN in this framework adopting a parameter sharing mechanism, the whole framework is end-to-end, which improves the speed of pedestrian detection and achieves real-time pedestrian detection at night. In Ref. [22], the author proposed a pedestrian detection model with scale perception based on Fast R-CNN. The model is divided into two sub-networks of different sizes to detect pedestrians of different scales. In Ref. [23], RPN is combined with cascaded random forest classifiers, which achieved the best detection results at that time. In Ref. [24], on the basis of the Faster R-CNN framework, PRN was improved to make it suitable for pedestrian detection. A multi-level feature extraction and fusion method was proposed to improve on the detection of small-scale pedestrians.

One-stage algorithm. Part of the deep learning pedestrian detection network is a one-stage detection algorithm, which does not need to look for region proposals and directly outputs the category probability and position coordinates of the image. The final detection results can be obtained directly from a single detection step. A typical one-stage detection framework is YOLO[25], proposed by Redmon et al. in 2015 as the first object detection system based on a single neural network. The framework of YOLO is shown in Fig. 3.

YOLO regards the whole image as input, and directly regresses the position and category of the bounding box in the output layer, which directly transforms the problem of object detection into a regression problem. Each image needs to be looked at only once to obtain the categories and corresponding positions of the objects. Therefore, YOLO detection is very fast, although it is not good at detecting small objects or objects that are

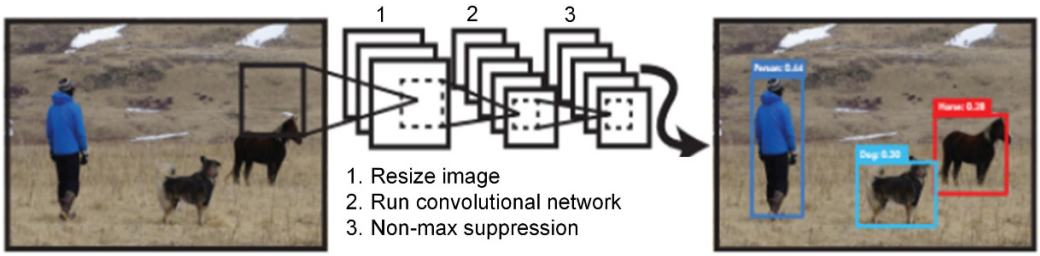


Fig. 3 YOLO detection system.

close to each other.

A pedestrian detection algorithm combining the Gaussian Mixture Model (GMM) and YOLO is proposed in Ref. [26]. The background is modeled by GMM, and pedestrians are initially detected with the results combined with the YOLO object detection network with different weights to improve the detection accuracy. In Ref. [27], a pedestrian detection and recognition method based on the depth residual network and YOLO method is proposed. Through analyzing the representation and distribution characteristics of pedestrian images, a 9:19 rectangular input CNN classification model is proposed to enhance the expression of pedestrian features. In addition, a YOLO pedestrian detection method based on a 50-layer pre-activated depth residual network is used to better characterize pedestrians. YOLO is also the basis for Ref. [28], a pedestrian detection and location method with high robustness in traffic environment is proposed according to vehicle vision and the characteristics of frequency resolution and target pedestrian size in a video, by optimizing network input size and using additional pedestrian data with a data augmentation strategy.

2.2 Action recognition

The ability of driverless vehicles to recognize pedestrian actions greatly affects the safety of autonomous driving. After detecting the presence of a pedestrian, it is necessary to identify the pedestrian's movements and to analyze the meaning of the pedestrian's actions. Related research on motion recognition and analysis can be dated back to the moving spot experiment of Johansson[29], conducted in 1973. By observing the information of light connected to a human body to recognize human motion, a 12-bone node human model was proposed. This point model method for describing human behavior plays an important role in guiding the subsequent behavior description algorithm based on the human skeleton.

In a simple and static scene, the background subtraction method can be used to segment pedestrian movements, which can then be matched and assigned according to feature descriptors. However, the scenario for autonomous driving is dynamic, complex, and uncertain, which increases the difficulty of human action recognition. In order to better recognize pedestrian actions in autonomous driving, it is necessary to extract information on pedestrian action characteristics. Feature extraction is therefore at the core of human action recognition.

Feature extraction methods can be divided into two types: traditional feature extraction and extraction using deep learning. Apparent feature extraction uses traditional algorithms to extract the shallow features of actions on the surface of images and obtain the expected feature information. Traditional representation extraction algorithms include optical flow and gradient. Feature extraction using deep learning methods automatically extracts action features at multiple levels by using the framework of deep learning to obtain unexpected feature information. A two-stream action recognition network based on CNN is a representative deep learning feature extraction method.

2.2.1 Traditional method

Optical flow characteristics. The earliest implementation of optical flow calculation was the HS optical flow algorithm proposed by Horn and Schunck[30]. The assumption of motion smoothness was proposed based on the assumption of optical flow invariance. From this, the objective function is constructed and the optical flow estimation is solved. Subsequently, optical flow computing entered a stage of rapid development, and many classical algorithms were proposed by researchers. For example, the LK optical flow algorithm proposed by Lucas and Kanade[31] is a two-frame differential optical flow estimation algorithm. In order to solve the problems of the non-conservation of brightness and discontinuity of motion in practical scenarios, Black and Anandan proposed the BA optical flow algorithm[32], and the optical flow method proposed by Brox and Malik sought to solve large displacements[33]. In recent years, new optical flow estimation algorithms have also been

proposed. For example, the DeepFlow method proposed by Weinzaepfel et al.[34] combines the matching algorithm with the differential method of optical flow, and constructs a six-layer matching framework through convolution and max-pooling, which is similar to a deep convolutional network.

Optical flow can effectively characterize human motion information, and much research has been conducted into extracting features by calculating optical flow and then identifying a pedestrian action. However, relying solely on optical flow characteristics does not guarantee a high accuracy of motion recognition, so optical flow characteristics are often combined with other motion features, as described in Refs. [35–37].

Gradient characteristics. Gradient feature is the most widely used action feature. Although the calculation of the gradient only needs to consider the difference between a pixel and the brightness of its adjacent region, and it thus takes a single form, an action can be effectively characterized after combination, transformation, and statistics are applied. Therefore, many action recognition methods based on gradient characteristics have been proposed.

HOG is the most widely used method of feature extraction, as presented at the CVPR conference by Dalal and Triggs[2]. The method is to characterize the partial information of the input image and to count the local information, which can characterize the features very effectively. Scale-Invariant Feature Transform (SIFT)[38] was also frequently used in early research. SIFT extracts the partial features of the image, finds the extreme points in the scale space, and extracts the position, scale, and direction information. It maintains the invariance of rotation, scale, and brightness, and also has a certain stability of perspective change and noise. The PCA-SIFTT method[39] improved on SIFT by using principal components analysis for dimensionality reduction to reduce the use of memory and improve the matching speed.

The fusion of gradient feature and optical flow feature through a histogram is the most efficient method to realize action representation. This idea was first put forward by Laptev et al.[40], who fused a gradient histogram with an optical flow histogram. First, the extreme points of the gradient feature in a three-dimensional neighborhood are extracted, and the optical flow between two adjacent pixels is calculated. Finally, the gradient and optical flow distribution are counted by histogram, and a normalization operation is carried out to realize the representation of an action.

HOG is insensitive to illumination changes and offsets, and can effectively characterize the edge features of pedestrians to identify pedestrian actions. However, HOG also has its shortcomings: the high feature dimension, large overlap, and a slow feature calculation speed because of using histogram statistics, which affects real-time performance and provides a poor processing ability for occluded pedestrians.

DT algorithm. The Dense Trajectories (DT) algorithm was proposed by Wang et al.[41] It uses dense sampling, feature point tracking, and feature point extraction, the procedure is as follows:

- Sample dense feature points on the original image;
- Filter feature points to preserve useful features and form spatial information;
- Track the feature points, forming the trajectory sequence of the feature points on the time axis, and forming the time domain information;
- Sample the feature points corresponding to each time slice and code the sequence to obtain the total feature; and
- Use an SVM classifier to classify.

The DT algorithm has been tested on 8 datasets, including KTH, YouTube, and UIUC. While the experimental results are superior than previous detection techniques, its performance is limited by the quality of the available optical flow. Therefore, Wang and Schmid[42] improved the algorithm in proposing iDT, which eliminates the background effect by estimating the motion model of the camera in orbit and optimizes the flow image. As the best action recognition algorithm prior to the application of deep learning, the iDT algorithm is highly effective and features good robustness. Many of its features have proven to be worth retaining, such as eliminating the background light flow caused by camera motion and extracting features along the trajectory. For example, Ref. [43] uses a CNN to extract features along the trajectory and thus makes further improvement.

2.2.2 Deep learning method based on two-stream network

Human action recognition in images and video has long been an important academic research field. In recent years, CNN has reached maturity for image recognition tasks. Since the publication of the winning model of the ImageNet Challenge in 2012, CNN-based methods have achieved outstanding performance in recognition in the image and video fields. For human action recognition

in video, Simonyan and Zisserman[44] proposed a two-stream CNN network, which uses two independent CNNs to process temporal and spatial information in the video separately. This framework is shown in Fig. 4.

The input of the space network is a single frame of video. Experiments on the input of the time network indicate that a two-way optical flow stacked over 10 consecutive frames is the best performing input setting. The two networks are trained separately and fused with SVM. This algorithm clearly improves performance over single-stream networks by capturing local temporal motion, but it has some shortcomings: the training clips are sampled uniformly from the video but the assumption that each clip is the same is not consistent with actual video action recognition tasks, and end-to-end training is difficult to achieve because the two-network training procedures are separated.

To solve the problem that two-stream networks cannot model long-term time structures, Wang et al.[45] proposed Temporal Segment Networks (TSN). Unlike the basic two-stream network using a single frame or single frame heap, TSN sparsely samples a series of short clips from the whole video. TSN also improves on the two-stream fusion algorithm by obtaining the fusion score using the weighted average of the final temporal and spatial scores.

To improve the performance of two-stream networks, Feichtenhofer et al.[46] proposed a two-stream convolutional network fusion for video action recognition. This algorithm uses CNN to fuse the temporal and spatial streams. In the fifth convolutional layer, convolution and pooling are merged, and a further fusion is carried out in the last (eighth) layer. The final output of this fusion is used for spatial and temporal loss evaluation. Ng et al.[47] proposed a two-stream and Long Short-Term Memory (LSTM) model operating on the outflow of basic dual-stream networks. The main improvement offered by this method is the use of an LSTM network in RNN to fuse the time and space streams.

Yan et al.[48] proposed Spatial-Temporal Graph Convolutional Networks (ST-GCN), which automatically learns both the spatial and temporal patterns from skeleton-based data. This framework not only provides great expressive power, but also has a stronger generalization capability for action recognition. Zhu et al.[49] proposed a hidden two-stream convolutional network for action recognition, which achieves end-to-end training and is a major breakthrough compared with previous algorithms. The use of optical flow by two-stream network architectures makes it necessary to estimate the optical flow before each sampling frame, which adversely affects storage space and speed. An unsupervised architecture is proposed to generate optical flow for the frame stack: the model MotionNet, which generates motion information from video frames, is trained and then added to the temporal stream to achieve end-to-end action prediction. Prediction speed and related deficiencies are improved with the automatic generation of optical flow, as the author is not reliant on the slower traditional methods of generating optical streams. This framework is shown in Fig. 5.

3 Commonly Used Public Action Datasets

Action datasets are an indispensable tool for action recognition research. An open and comprehensive dataset can not only improve the efficiency of the action recognition process and reduce the time and cost of data collection, but also provide a standard and unified test platform for the merits and demerits of various action recognition algorithms, so as to promote the development of the field. Typically used action datasets are KTH, the UCF series, Hollywood2, and Google AVA.

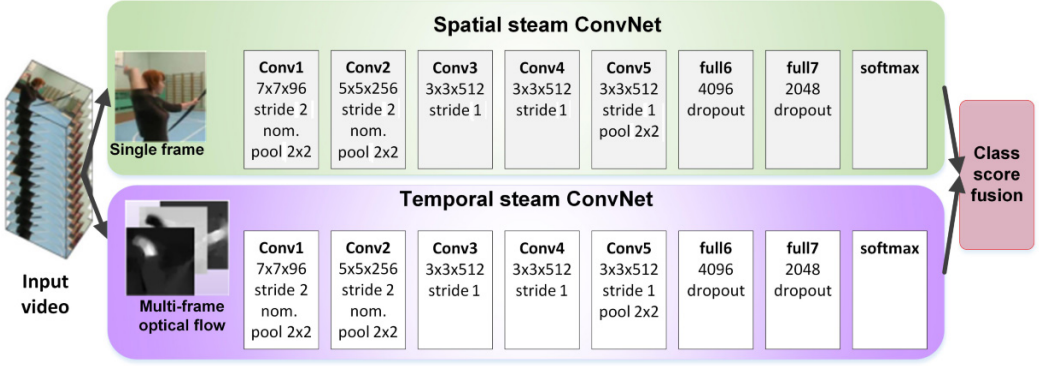


Fig. 4 Two-stream architecture for video classification.

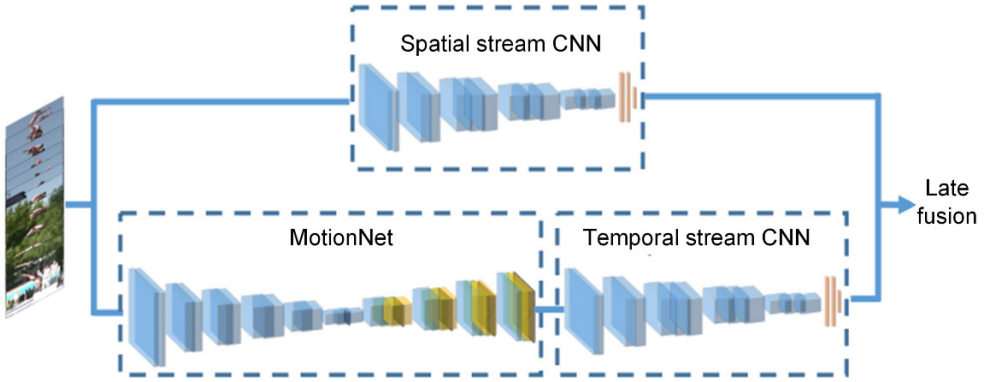


Fig. 5 Ilustration of proposed hidden two-stream network.

Samples from each of these datasets are provided in Fig. 6. A comparison of several commonly used datasets is shown in Table 1, which divides them into single action and video datasets.

4 Summary

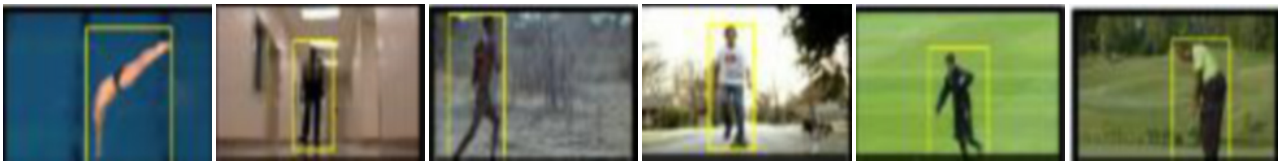
Pedestrian action recognition is a major research focus in the field of image detection and classification. Detection (judging whether there are pedestrians present), location (extracting pedestrians from a complex environment), and recognition (accurately understanding the meaning of specific actions) constitute the research content of pedestrian action recognition. This paper summarizes the related research, and draws from it the following three challenges involved in pedestrian action recognition.

(1) Complex environment. The complexity of the environment and the degree of noise aggregation in the background interfere with the accuracy and

KTH dataset[50]

(a) Walking (b) Jogging (c) Running (d) Boxing (e) Hand waving (f) Hand clapping

UCF Sports dataset[51]



(g) Diving (h) Walking (i) Running (j) Skateboarding (k) Kicking (l) Golf Swing

Hollywood2 dataset[52]

(m) Hugging (n) Running (o) Driving (p) Getting out of the car (q) Eating (r) Answering phone

Google AVA[53]

(s) Walking (t) Bowing (u) Jogging (v) Dancing (w) Grouching (x) Jumping

Fig. 6 Representative human action datasets.

Table 1 Human action recognition datasets.

Dataset name	Year	Number of classes	Number of actors	Number of videos	Attribute
KTH	2004	6	25	2391	Single action
UCF Sports	2008	9	Many	200	Single action
UCF 50[54]	2013	50	Many	6676	Single action
UCF101[55]	2012	101	Many	13 320	Single action
UCF YouTube[56]	2009	11	Many	$\geqslant$1100	Continuous action
Hollywood2	2009	12	Many	3669	Continuous action
Google AVA	2017	80	Many	$\geqslant$57 600	Continuous action

efficiency of pedestrian action recognition. Researchers have effectively segmented pedestrians from complex backgrounds by studying the fixed relationship between objects and pedestrians in the image. By combining the inherent characteristics of pedestrians, the distinctiveness of pedestrians can be enhanced so as to get a more complete and accurate separation of pedestrian models.

(2) Sheltered pedestrians. In addition to the influence of complex backgrounds on pedestrian detection, the problems of mutual occlusion between pedestrians and the occlusion of pedestrians by other objects also have a certain impact on accurate detection and recognition. The efficiency will decrease if images are detected only from a single perspective, so researchers have proposed multi-view object recognition methods[57]. Data from different perspectives be combined to improve the efficiency of pedestrian detection and action recognition in complex fields.

(3) Definition of action. Currently there are semantic definitions of actions, but there is limited pedestrian action data for autonomous driving. Meanwhile, there are no standard action benchmarks. By classifying and defining autonomous driving actions in more detail, more intelligent and anthropopathic recognition results can be produced.

Pedestrian action recognition will play an important role in autonomous driving, safety monitoring, environmental management in urban human settlements and other fields. To solve the problem that pedestrians are obscured in a single view, researchers have proposed the use of multi-view data for pedestrian action recognition. In the future, pedestrian detection and action recognition based on stereo vision and multi-view pedestrian action definitions will become a new focus of pedestrian action recognition research.

Acknowledgment

We are grateful to the anonymous reviewers for their constructive suggestions. This study was partially funded by the National Natural Science Foundation of China (Nos. 61871038, 61803034, and 61672178), Beijing Natural Science Foundation (No. 4182022), and Beijing Union University Graduate Funding Project.

References

[1] N. Ma, Y. Gao, J. H. Li, and D. Y. Li, Interactive cognition in self-driving, (in Chinese), *Sci. Sin. Inform.*, vol. 48, no. 8, pp. 125–138, 2018.

[2] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893.

[3] N. Dalal, Finding people in images and videos, Ph.D. dissertation, Institute National Polytechnology de Grenoble-INPG, Grenoble, France, 2006.

[4] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] D. Singh, M. A. Khan, A. Bansal, and N. Bansal, An application of SVM in character recognition with chain code communication, in *Proceedings of Control and Intelligent Systems (CCIS)*, Mathura, India, 2015, pp.167–171.

[6] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah, H. A. Rahman, F. Hussin, H. Abdullah, and R. Saidur, A review on applications of ANN and SVM for building electrical energy consumption forecasting, *Renewable and Sustainable Energy Reviews*, vol. 33, no. 1, pp. 102–109, 2014.

[7] H. Liu, T. Xu, X. Wang, and Y. Qian, Related HOG features for human detection using cascaded AdaBoost and SVM classifiers, in *Proceedings of Advances in Multimedia Modeling*, Berlin, Germany, 2013, pp. 345–355.

[8] Y. Pang, Y. Yuan, X. Li, and J. Pan, Efficient HOG human detection, *Signal Processing*, vol. 91, no. 4, pp. 773–781, 2011.

[9] F. Han, Y. Shan, R. Cekander, H. S. Sawhney, and R. Kumar, A two-stage approach to people and vehicle detection with HOG-based SVM, in *Proceedings of Performance Metrics for Intelligent Systems 2006 Workshop*, Gaithersburg, MD, USA, 2006, pp. 133–140.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, Object detection with discriminatory trained part-based models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[11] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, Robust multi-resolution pedestrian detection in traffic scenes, in *Proceedings of the IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR), Portland, OR, USA, 2013, pp. 3033–3040.

[12] J. X. Zeng and C. Xiao, Pedestrian detection combined with single and couple pedestrian DPM models in traffic scene, *Acta Electronica Sinica*, vol. 44, no. 11, pp. 2668–2675, 2016.

[13] J. Yan, Z. Lei, L. Wen, and S. Z. Li, The fastest deformable part model for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014, pp. 2497–2504.

[14] Y. Tian, R. Sukthankar, and M. Shah, Spatiotemporal deformable part models for action detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, 2013, pp. 2642–2649.

[15] P. Viola, M. J. Jones, and D. Snow, Detecting pedestrians using patterns of motion and appearance, *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014, pp. 580–587.

[17] K. He, X. Zhang, S. Ren, and J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[18] R. Girshick, Fast-RCNN, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440–1448.

[19] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gever, and A. W. M. Smeulders, Selective search for object recognition, *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in *Proceedings of Neural Information Processing Systems 28 (NIPS 2015)*, Montreal, Canada, 2015, pp. 91–99.

[21] L. G. Ye, S. Y. Sun, K. J. Gao, and H. T. Zhao, Nighttime pedestrian detection based on faster region convolution neural network, (in Chinese), *Progress in Laser and Optoelectronics*, vol. 54, no. 8, pp. 123–129, 2017.

[22] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, Scale-aware fast R-CNN for pedestrian detection, *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[23] L. Zhang, L. Lin, X. Liang, and K. He, Is faster R-CNN doing well for pedestrian detection?, in *European Conference on Computer Vision (ECCV)*, Amsterdam, the Netherlands, 2016, pp. 443–457.

[24] A. X. Guo, B. Q. Li, and Y. Li, Pedestrian detection based on deep convolutional neural network, (in Chinese), *Computer Engineering and Applications*, vol. 52, no. 13, pp. 162–166, 2016.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 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, 2016, pp. 779–788.

[26] Z. Gao, S. Li, J. Chen, and Z. J. Li, Pedestrian detection method based on YOLO network, *Computer Engineering*, vol. 44, no. 5, pp. 215–219, 2018.

[27] X. Z. Hao and Z. G. Chai, Improved pedestrian detection method based on depth residual network, (in Chinese), *Application Research of Computers*, vol. 36, no. 6, pp. 1–3, 2019.

[28] X. Z. Hao and L. Q. Huang, Pedestrian detection based on deep neural network in traffic environment, (in Chinese), *Information & Comunications*, vol. 185, no. 5, pp. 74–77, 2018.

[29] G. Johansson, Visual perception of biological motion and a model for its analysis, *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[30] B. K. P. Horn and B. G. Schunck, Determining optical flow, *Artificial Intelligence*, vol. 17, nos. 1–3, pp. 185–203, 1981.

[31] B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, in *Proc. 7th Int. Conf. on Artificial Intelligence (IJCAI)*, Vancouver, Canada, 1981, pp. 674–679.

[32] M. J. Black and P. Anandan, The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields, *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[33] T. Brox and J. Malik, Large displacement optical flow: descriptor matching in variational motion estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, 2011.

[34] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, DeepFlow: Large displacement optical flow with deep matching, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013, pp. 1385–1392.

[35] H. Wang and C. Schmid, Action recognition with improved trajectories, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013, pp. 3551–3558.

[36] X. Peng, C. Zou, Y. Qiao, and Q. Peng, Action recognition with stacked Fisher vectors, in *European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014, pp. 581–595.

[37] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on non-linear dynamic systems for the recognition of human actions, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 1932–1939.

[38] D. G. Lowe, Object recognition from local scale-invariant features, in *ICCV*, vol. 99, no. 2, pp. 1150–1157, 1999.

[39] Y. Ke and R. Sukthankar, PCA-SIFT: A more distinctive representation for local image descriptors, In *CVPR* , vol. 4, no. 2, pp. 506–513, 2004.

[40] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, Learning realistic human actions from movies, in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, Anchorage, AK, USA, 2008, pp. 1–8.

[41] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, Dense trajectories and motion boundary descriptors for action recognition, *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[42] H. Wang and C. Schmid, Action recognition with improved trajectories, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013, pp. 3551–3558.

[43] L. Wang, Y. Qiao, and X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 4305–4314.

[44] K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos, in *Proceedings of Neural Information Processing Systems 27 (NIPS 2014)*, Montreal, Canada, 2014, pp. 568–576.

[45] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in *European Conference on Computer Vision*, Amsterdam, the Netherlands, 2016, pp. 20–36.

[46] C. Feichtenhofer, A. Pinz, and A. Zisserman, Convolutional two-stream network fusion for video action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2016, pp. 1933–1941.

[47] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, Beyond short snippets: Deep networks for video classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4694–4702.

[48] S. Yan, Y Xiong, and D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018.

[49] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, Hidden two-stream convolutional networks for action recognition, in *Asian Conference on Computer Vision*, Perth Weatern, Australia, 2018, pp. 363–378.

[50] C. Schuldt, I. Laptev, and B. Caputo, Recognizing human actions: A local SVM approach, in *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004*, Cambridge, UK, 2004, pp. 32–36.

[51] M. D. Rodriguez, J. Ahmed, and M. Shah, Action MACH: A spatio-temporal maximum average association height filter for action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, USA, 2008, pp. 1–6.

[52] M. Marszalek, I. Laptev, and C. Schmid, Actions in context, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 2929–2936.

[53] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al., AVA: A video dataset of spatio-temporally localized atomic visual actions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6047–6056.

[54] K. Reddy and M. Shah, Recognizing 50 human action categories of Web videos, *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2017.

[55] K. Soomro, A. R. Zamir, and M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv: 1212.0402, 2012.

[56] J. Liu, Y. Yang, and M. Shah, Learning semantic visual vocabularies using diffusion distance, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 461–468.

[57] H. Su, S. Maji, E. Kalogeraki, and E. Learned-Miller, Multi-view convolutional neural networks for 3D shape recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 945–953.

Li Chen received the BS degree from Beijing Union University, Beijing, China. She is currently a master student in Beijing Union University, Beijing, China. Her current research interest is pedestrian action recognition.

Nan Ma received the PhD degree from the University of Science & Technology Beijing, Beijing, China, in 2013. She is now an associate professor in the College of Robotics, Beijing Union University, Beijing, China. Her current research interests cover interactive cognition, multimedia content analysis, and knowledge discovery. She has published more than 30 papers in international journals and conferences.

Patrick Wang received the PhD degree from Oregon State University, USA. He is currently a professor in Northeastern University, Boston, USA. His current research interests include intelligent pattern recognition and interactive machine learning.

Jiahong Li received the PhD degree from Beijing Institute of Technology, Beijing, in 2017. Currently, he is a lecturer at the College of Robotics, Beijing Union University. His research interests include interactive cognition, distributed estimation fusion, and multi-agent decision theory with a focus on applications for autonomous vehicles. He is a member of IEEE.

Pengfei Wang received the MS degree from Beijing Union University, Beijing, in 2015. He is now a software engineer in the Communication and Information Center of Ministry of Emergency Management of the People's Republic of China, Beijing, China. His research interests include internet of things and digital image processing.

Guilin Pang is an undergraduate student in Beijing Union University. He won the First Prize of the 2nd World Intelligent Driving Challenge Virtual Scene Group. His main research interest is lane detection.

Xiaojun Shi is an undergraduate student in Beijing Union University. He won the Second Prize of the 2nd World Intelligent Driving Challenge Virtual Scene Group. His main research interest is digital image processing.