A Multimodal Detection and Tracking System Based on Deep-Learning for Traffic Monitoring

Xinxu Li[1]; Guizhen Yu[2]; Yunpeng Wand[3]; Xinkai Wu[4]; and Yalong Ma[5]

[1]Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang Univ., Xue Yuan Rd. No. 37, HaiDian District, Beijing 100191
[2]Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang Univ., Xue Yuan Rd. No. 37, HaiDian District, Beijing 100191 (corresponding author). E-mail: yugz@buaa.edu.cn
[3]Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang Univ., Xue Yuan Rd. No. 37, HaiDian District, Beijing 100191
[4]Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang Univ., Xue Yuan Rd. No. 37, HaiDian District, Beijing 100191
[5]Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang Univ., Xue Yuan Rd. No. 37, HaiDian District, Beijing 100191

ABSTRACT

Traffic surveillance videos can provide data of multimodal (car, bus, pedestrian, bicyclist, et al.) activities for transportation research in general. However, automated data extraction remains an issue due to illumination variation, size scaling, and angle changes. To tackle these challenges, this paper proposed a deep-learning based multimodal detection and tracking system. The proposed system consists of two modules: 1) a multimodal detection module which is based on the object detection framework of Faster R-CNN, and 2) a multiple object tracking module which is based on the KCF tracking algorithm. A comprehensive experiment based on traffic videos captured in different scenarios is conducted to evaluate the performance of the proposed system. Testing results demonstrate that the proposed system can achieve better performance for multimodal detection and tracking compared with existing methods. The proposed system is tested to be robust to illumination variation, size scaling, and angle changes.

INTRODUCTION

Traffic surveillance camera has obvious advantages over other sensors such as loop detector (Masher, 1975) or microwave radar (Clippard, 1996), dues to its simple installation process, less maintenance cost and wide applicability (Mimbela et al., 2000). Data of multimodal activities captured conveniently by traffic surveillance camera, such as volumes, speeds, and space-time trajectories are essential in the field of transportation research in general. However, due to illumination variation, size scaling, and angle changes, automated data extraction for multimodal from traffic videos remains a challenge.

In current researches, object detection techniques based on computer vision can be divided into three classes: traditional method, machine learning, and deep learning. Generally, in traditional methods, some simple rules are constructed to detect the objects in terms of features extracted manually, such as edge detection (Canny 1986), background subtraction (Alan, 2001),

and frame difference (Zhan et al., 2007). All of them are sensitive to background motion and scene complexity. Machine learning generates a better classification criteria after learning those manually extracted features. Support Vector Machine (Osuna et al., 1997) and AdaBoost Cascaded Classifier (Viola et al., 2001) are two commonly used machine learning methods, which are more reliable than traditional method. Deep learning is a branch of machine learning, containing several typical network architectures. Convolutional Neural Network (CNN) (Lécun et al., 1998), which is a special neural network architecture of Deep-Learning, has recently been applied to various computer vision tasks (Krizhevsky et al., 2012) such as image classification, semantic segmentation, object detection, and many others. Such great success of CNN is mostly attributed to their outstanding performance in feature extraction.

Object tracking is one of the most important components in our work. Given the initialized state of a target in a frame of a video, the goal of tracking is to estimate the states of the target in the subsequent frames. Firstly, searching the moving target directly from the video sequences is the core of the tracking (Lucas et al., 1981). Then tracking by detection method, such as Struck (Hare et al., 2011) and TLD (Kalal et al., 2012), and this method became widely used since it is more reliable. Object tracking has been studied for several decades, and much progress has been made in recent years.

Traffic monitoring based on surveillance camera, however, has been less affected by the latest technology. Considering of the great success of deep-learning in object detection, this paper attempts to apply it to traffic monitoring. The proposed system consists two modules: 1) a multimodal detection module which is based on the object detection framework of Faster R-CNN (Ren et al., 2016); and 2) a multimodal tracking module which is based on the KCF tracking algorithm (Henriques et al., 2015). With the proposed system, data of multimodal activities at signalized intersections for transportation research are derived.

RELATED WORK

For multimodal detection in traffic surveillance videos, Angel et al. (2003) proposed a background subtraction method for vehicle detection. Wang (2011) proposed a frame difference algorithm for moving vehicle detection in freeway, which only performed well under the relatively simple background. Both these methods can achieve multimodal detection, but they are invalid for stationary objects. Typical machine learning based object recognition algorithm is less sensitive to multimodal moving or stationary. Sotelo et al. (2006) used the Support Vector Machine classifier based on multi-features for pedestrian detection, while the multi-features are only suitable for pedestrian. Meng et al. (2015) used the Adaboost Cascaded Classifier based on Haar-like features for vehicle detection, while this method had difficulty in clutter scenarios in videos. Based on appropriate features, all these classifiers above can achieve a multimodal detection, although they are binary classifier. In that case, multiple classifiers are necessary and the time consumption is huge.

For multimodal tracking in traffic surveillance videos, Nagel et al. (1998) proposed an optical flow method for vehicles. While this method can only detect and track moving objects. A camshaft based multiple vehicles tracking for visual traffic surveillance was developed by Liu et al. (2009). Camshaft (Bradski et al., 1998) method cannot perform well for objects as dark as the road surface, which limits its application. Dong et al. (2013) tracked pedestrian based on the TLD (Kalal et al., 2012) framework. And Zou et al. (2013) used TLD (Kalal et al. 2012) to track vehicle. This Tracking-by-Detection method becomes time-consuming when it is used for multiple target tracking.

In this paper, a CNN based detector is implemented instead of using the method mentioned above. CNN is supported by the framework of GPU Parallel Computing, which can speed up the calculation process. Meanwhile, we use KCF to achieve a fast and accurate tracking.

SYSTEM FRAMEWORK

As mentioned above, the framework consists of a multimodal detection module and a multimodal tracking module, which is shown in Figure 1. The core component of first part is a detector based on Faster R-CNN, tasked with distinguishing between the multimodal and the surrounding environment. Although the performance of detector is satisfactory, the detected results in each frame are independent. To collect multimodal data with time series, the system must find each object's corresponding positional relation among frame sequences, which can be conducted by tracking. Consequently, the KCF tracking method is integrated into the system.

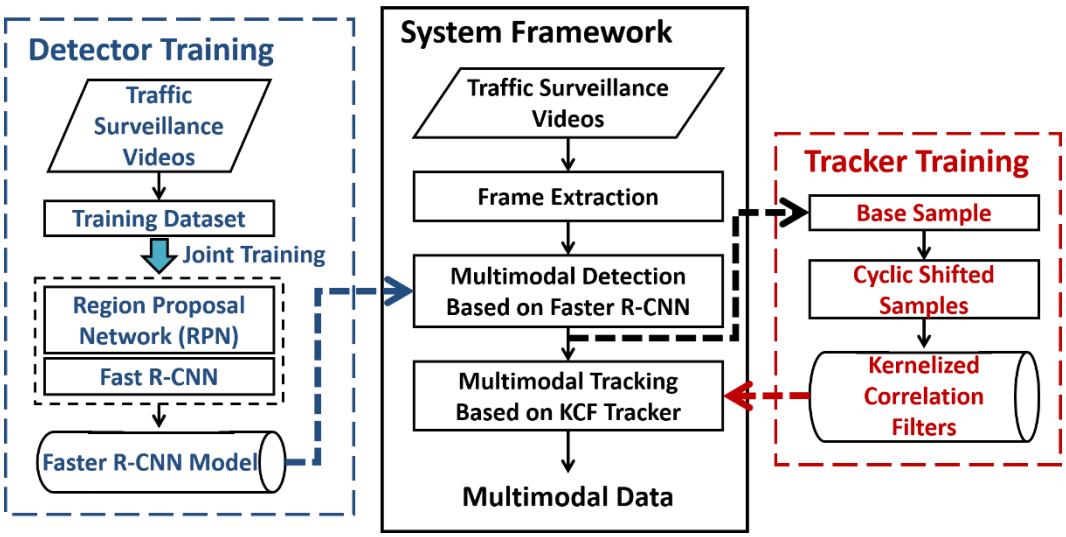

Figure 1. The framework of proposed system.

In the running process, the detector can perform multimodal detection simultaneously, while one tracker can only track one object in a moment. Therefore, under the tracking stage, the system deploys multiple trackers for each object correspondingly. To optimize the tracking process, the structure of tracker is designed as follows:

$$\begin{cases} Tracker_n : \{ID, Label, Point\_Pool, Frame_{in}, Frame_{out}\} \\ Point\_Pool : \{P_1, P_2, \ldots, P_i, \ldots, P_{final}\} \\ P_i : \{i.x, y\} \end{cases} \quad (1)$$

where, i is the sequence number of frame; n is the sequence number of this object, $ID = n$; $Label$ is the type of this object; P_i is the location of this object in frame_i; is the collection of all P_i of this object from appear to disappear; $Frame_{in}$ is the sequence number of the frame where this object first appear; $Frame_{out}$ is the sequence number of the frame where this object disappear.

Once a new object is detected, a tracker for it is established. We assume that $Tracker_n$ gives a $P_{i\_tracker}$ of $object_n$ in $frame_i$ in terms of P_{i-1}, and the detector provides all positions of

multimodal at the same time, we use symbol P to represent those positions here. The P_i can be determined by:

$$P_i = \begin{cases} P_x & if\ P_x \cap P_{i\_track} > 80\% \\ P_{i\_track} & else \end{cases} \quad (2)$$

Due to the complexity of the scenes in the surveillance video, the error of tracking will gradually accumulate. Since Faster R-CNN can always give the accurate positions of objects, errors is corrected every frame by (2). Besides, tracking stops if objects leave the designated region or some special situations. The criteria of terminating the tracking process is expressed by (3):

$$Stop = \begin{cases} 1 & if\ P_i\ is\ out\ of\ the\ designated\ region \\ 1 & if\ the\ number\ of\ "P_i = P_{i\_track}" > T_f \\ 1 & if\ Frame_i - Frame_{in} > T_n \\ 0 & else \end{cases} \quad (3)$$

where T_f and T_f are the thresholds set in terms of the videos' scenarios.

Situation 1 means a normal tracking process. Situation 2 indicates a failed tracking: Since detector has not correct the tracker for T_f consecutive frames, there is no object near the tracker. Situation 3 shows that the tracker is tracking false object. The above situations will lead to a tracking termination, after which data of tracked objects are saved and documented in the database.

METHODOLOGY

As mentioned earlier, to derive trajectory of each traffic participant in traffic surveillance videos, both a detector and a tracker are essential. Faster R-CNN achieves a better performance compared with existing methods. In order to balance the time consumption by computational complexity of deep neural network, a fast tracking method --KCF is adopted.

Multimodal Detection based on Faster R-CNN

In this part, the Faster R-CNN is employed to achieve multimodal detection. It is based on CNN, which can extract more expressive features than the traditional manual extraction in detection stage, therefore it is applicable to occlusion, size scaling and angle changes scenarios. Figure 2 shows the framework of Faster R-CNN (upper), which consists of CNN (left), Region Proposal Network (RPN) (right), and Fast R-CNN (Girshick 2015) (the right part of the upper).

CNN (Zeiler and Fergus model, etc.) (Zeiler et al., 2013) can turn an input image (size: 600×1000) into 256 feature maps (size: 40×60), which are shared by a RPN and a Detection Network. Then RPN generates region proposals, which are used by Detection Network for multimodal detection.

To generate region proposals, a 3×3 window slides over the feature map output by the last shared convolution layer. Each sliding window is mapped to a lower-dimensional vector (256-d for ZF, etc.) (Zeiler et al., 2013) and a point in input image. This vector is fed into two sibling fully-connected (FC) layers: a box-regression layer (reg) and a box-classification layer (cls). At each point location, anchors, which are 9 region proposals in image including 3 scales (128, 256, 512) and 3 aspect ratios (1:1, 1:2, 2:1), are simultaneously predicted Thus the reg layer has 4×9=36 outputs encoding the coordinates of 9 boxes. The cls layer outputs 2×9=18 scores that

estimate probability of object / not-object for each proposal. For a feature (size: 40×60), there are 40×60×256 anchors in total, each with an objectness score. The anchors whose objectness scores are in top three hundred will be input to detection network as region proposal

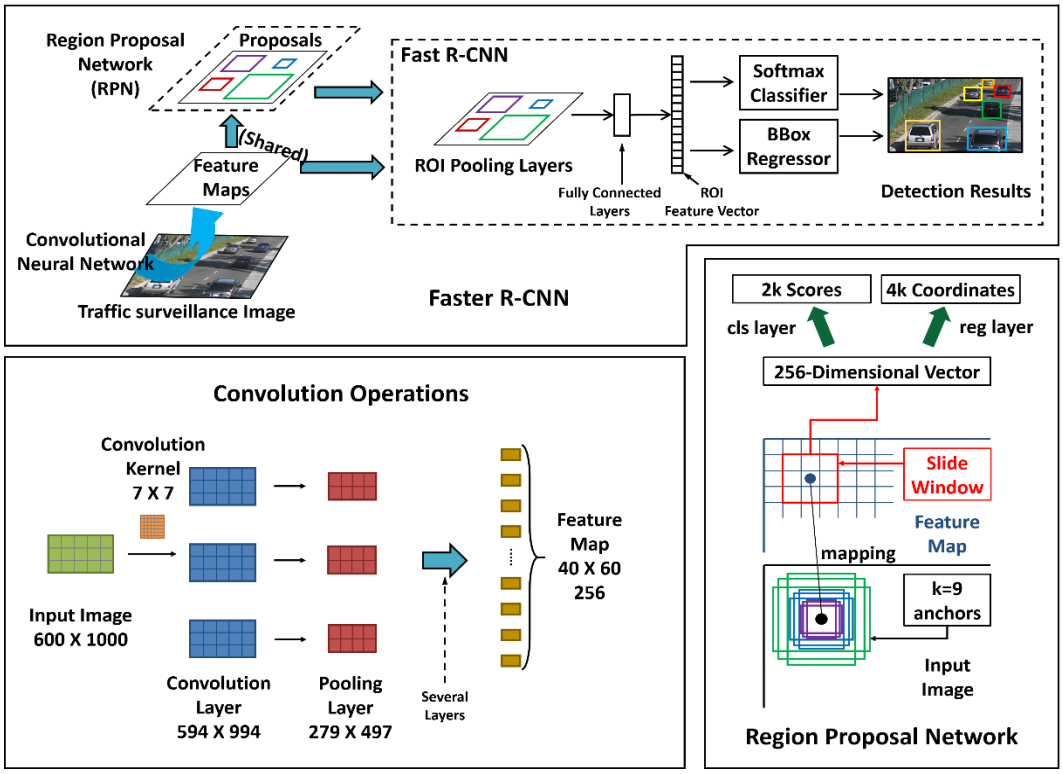

Figure 2. The framework of faster R-CNN

Then, each region proposal was put into a region of interest (ROI) pooling layer. Each fixed-length (4096-d for ZF, etc.) (Zeiler et al., 2013) feature vector is fed into a sequence of fully-connected (FC) layers that finally branch into two sibling output layers: a softmax classification layers and a bounding-box regression layer. The former produces softmax probability estimates over multiple classes including the "background" class. The other layer outputs 4 refined bounding-box positions values for each region proposal. All parameters of layers are calculated in the training. In the training process, only the parameters that make the loss function attains the minimum value can meet the requirements. The back propagation and the stochastic gradient descent are used to minimize the loss function.

Multiple Object Tracking based on KCF

To derive objects' trajectories, the KCF tracking method is employed. This method is a kind of Tracking-by-Detection method, which trains a detector and uses this detector to test on many candidate patches to search the most likely region. Due to computational complexity and detection accuracy are both related to the sample capacity in training process, it is hard to balance instantaneity and robustness for traditional Tracking-by-Detection method. However, by constructing cyclic shifted samples (see Figure 3), which can be diagonalized with the discrete Fourier transform (DFT), the KCF tracking method reduces both cache and computation by several orders of magnitude.

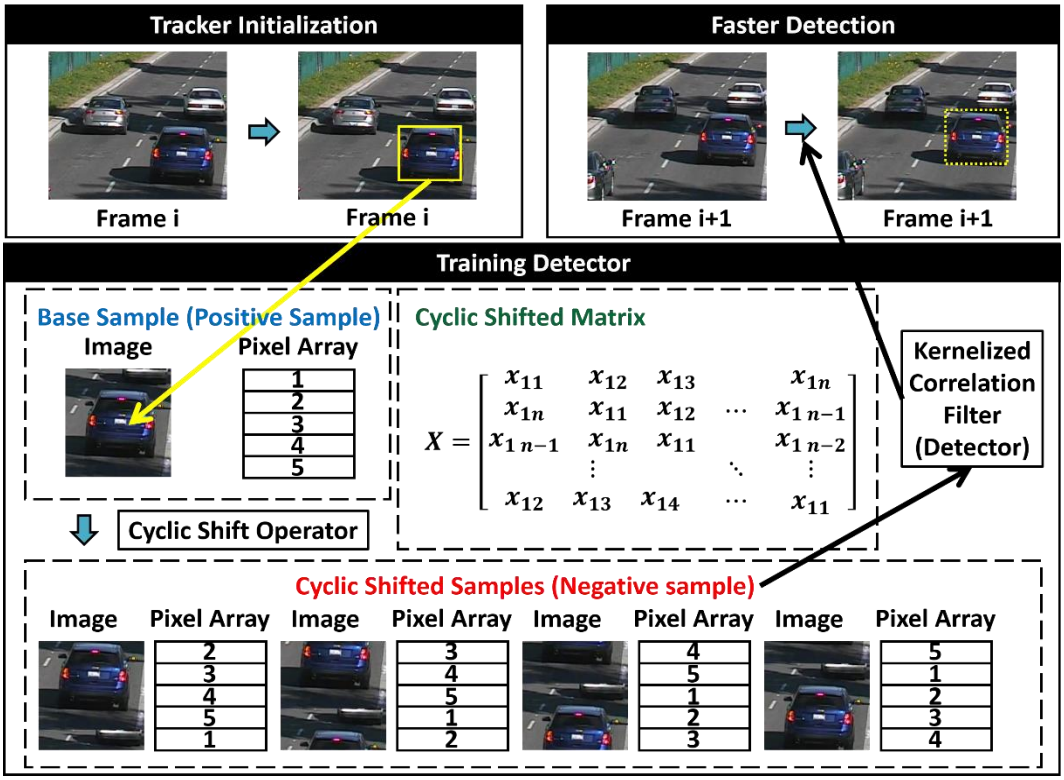

Figure 3. Using cyclic shifted samples to fast training and detection process of KCF tracker.

The goal of training process is to find a function $f(\boldsymbol{x}_i) = \boldsymbol{\omega}^T \boldsymbol{x}_i$ that minimizes the squared error over samples $\boldsymbol{x}_i$ and their regression targets y_i. The closed-form solution in the Fourier domain is (4):

$$\boldsymbol{\omega} = \left(\boldsymbol{X}^H \boldsymbol{X} + \lambda\right)^{-1} \boldsymbol{X}^H \tag{4}$$

where, $\boldsymbol{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\right]^T$ is the matrix of the samples' data; $\boldsymbol{y} = \left(y_1, y_2, \ldots, y_n\right)$ is the vector of the samples' value; $\boldsymbol{X}^H$ is the Hermitian transpose; and λ is a regularization parameter that controls over fitting.

In the process of computing the solution, solving a large system of linear equations is essential unless $\boldsymbol{X}$ is a cyclic shifted matrix as shown in Figure 3. As cyclic shifted matrices can be diagonalized by the discrete Fourier transform, and (4) can simplified to (5):

$$\boldsymbol{\omega} = \sqrt{n}\mathbf{F}\frac{\boldsymbol{x}_1 \odot \hat{\boldsymbol{y}}}{\boldsymbol{x}_1 \odot \boldsymbol{x}_1 + \boldsymbol{\lambda}} \tag{5}$$

where, $\boldsymbol{F}$ is a discrete Fourier matrix, which is a constant matrix that does not depend on $\boldsymbol{x}_1$; $\boldsymbol{x}_1 = \mathcal{F}\ \boldsymbol{x}_1\ = \sqrt{n}\boldsymbol{F}\ \boldsymbol{x}_1$ denotes the DFT of the generating vector; and $\odot$ represents the element-wise product.

As the matrix operation is simplified as the point product of vector, the computational complexity reduced. Therefore, the object need to be tracked in frame is regarded as the base sample. The other samples are obtained after a cyclic shift operator on the base sample,

meanwhile their sample value y_i are given according to the shifted level.

After that, a correlation filter is trained. To search the object of interest, regions $\sum_{i=1}^{n} \boldsymbol{x}_i$ are input to correlation filter. Since this is a filtering operation, it can be formulated more efficiently in the Fourier domain. Typically, the maximum response is the location of the tracked object. Cyclic shifted matrices and Fast Fourier Transform achieve the fast training and detection, and a multi-channel data calculation function make it is possible to use the HOG in tracking, which is a classic feature in computer vision. Thus, KCF tracker has a good performance on illumination variation and tracking speed, despite it is sensitive to scale variation, shape change and fast motion.

EXPERIMENT AND EVALUATION

In this section, comprehensive experiments based on traffic videos captured in different scenarios are conducted to evaluate the performance of the proposed system. The experiments consist of the test of multimodal detection and multimodal tracking. These evaluation results is explained in detail in the latter part of this section.

Experiment Platform and Device

In our work, the surveillance cameras are installed at 5~10 meters above the ground, monitoring the roads constantly. Therefore, the traffic videos collect a variety of data such as light and weather changes, different road conditions for the experiments. The experimental device with a spec of an Intel i7-5930k CPU and an NVidia Titan X GPU is used to run deep learning network.

Performance Evaluation of Multimodal Detection

The performance of multimodal detection is evaluated by recall and precision, which are in the range of [0, 1]. Recall and precision are defined as (6) and (7):

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

where, *TP* (True Positive) is the number of objects detected correctly; *FP* (False Positive) is the number of objects detected incorrectly; *FP* (False Negative) is the number of missed objects.

Recall reflects the proportion of the cases of being correctly classified in the total positive objects. Precision reflects the proportion of the cases of being correctly classified in the total detected objects. When the recall approaches to 1, it means that multimodal detection performances well. Similarly, precision also follows the same principle.

In this paper, 1000 frames from traffic videos are selected for detection experiment. The size of each frame is in the range of 1280×720 ~ 1280×960. Different scenarios, such as intersection and freeway in a variety of weather and illumination conditions, are included in those frames. Multiple objects appear on each frame. The size of each object is in the range of 0.5% ~ 15% of the frame. The average time for processing a frame is near 0.25s. The test dataset above is sufficient to prove the applicability of the proposed system. We make a statistic on those detected objects that are not far away from the camera and have not been seriously covered. In

the statistical process, the recall and precision of each type of multimodal are calculated one by one. For example, to calculate the recall and precision of car, the number of cars detected correctly or incorrectly and the number of missed cars are counted manually and brought into (6) and (7). Figure 4 shows the types of multimodal (upper) and the result of detection in an intersection (down). The upper left corner of Figure 4 is the detection time. The edges of the bounding boxes are determined based on objects' types. Table 1 shows the detection results, including object type, the amount of training samples, recall and precision.

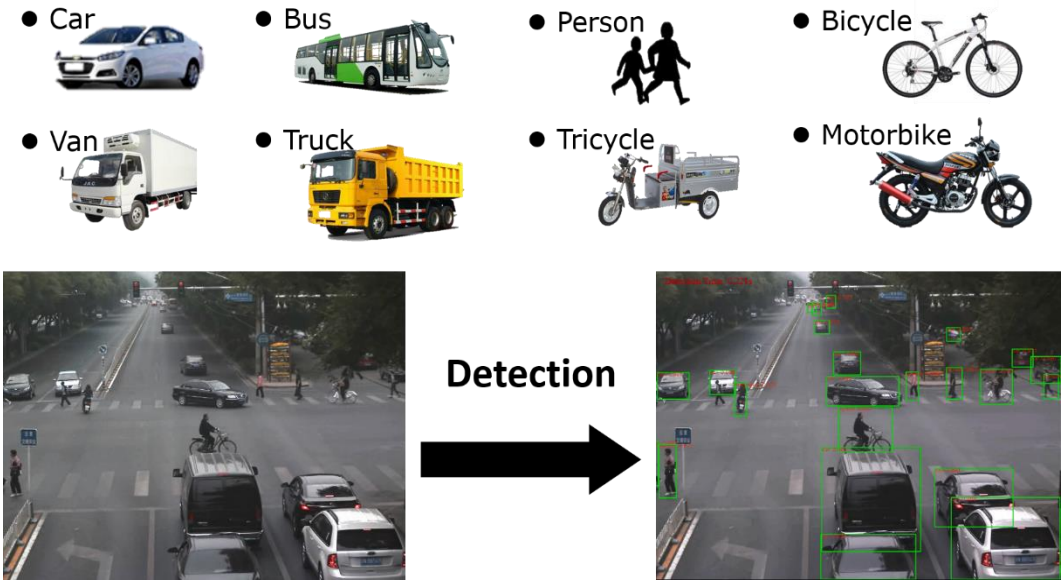

Figure 4. The performance evaluation of multimodal detection.

Table 1. Multimodal Detection Results

Object Type	Number of Training Samples	Recall	Precision
Car	21278	93.49%	97.09%
Bus	1765	88.76%	98.34%
Van	472	91.86%	94.95%
Truck	731	89.36%	94.38%
Pedestrian	6229	90.17%	96.56%
Bicycle	821	87.44%	93.53%
Motorbike	789	91.91%	94.72%
Tricycle	628	87.61%	94.84%

From the results, we can see multimodal in the visible region are detected simultaneously. The recall and precision of multimodal are 85%~100%, which indicate a satisfactory performance of the detection method. Since the detection accuracy and the number of samples are positively correlated, the precision of car and pedestrian is above 90%. Therefore, the accuracy of detection can be further improved with the increase of the number of samples.

Performance Evaluation of Multimodal Tracking

Tracker should indefity the correct locations of objects as precisely as possible. The performance of our system is evaluated by *TR* (Tracking Rate), which is defined as (8):

$$TR = \frac{N_{tracked}}{N_{total}} \quad (8)$$

where, $N_{tracked}$ is the number of objects which are tracked successfully, and N_{total} is the number of tested objects.

1000 moving objects including all types are used for detection experiment. 500 of them are in freeway, while the other 500 objects are in intersections. If the rectangle can always cover 80% of an object while this object is moving from the specified starting position to the end position, $N_{tracked}$ increases by one. Table 2 shows the tracking results of each type objects.

Table 2. Multimodal Tracking Results

Object Type	Tracking Rate in Freeway	Tracking Rate in Intersection
Car	94.05%	67.40%
Bus	91.22%	63.26%
Van	90.47%	45.45%
Truck	92.70%	86.04%
Pedestrian	-	40.30%
Bicycle	-	69.23%
Motorbike	-	67.44%
Tricycle	-	66.67%

Generally, pedestrian and non-motorized vehicle should not appear in freeway. Due to objects moving directions are almost parallel, the probability of temporary occlusion is tiny. However, the surveillance videos in intersection are more clutter. Temporary occlusions of objects interrupt the tracking. As a result, tracking rate in intersection is lower than in freeway. In addition, due to the length of the pedestrian in vertical direction is almost three times the length in horizontal direction, there are many failures of KCF tracker to capture pedestrian moving along horizontal direction in frames, which leads to the relatively low tracking rate for pedestrian.

CONCLUDING REMARKS

In this paper, an automatic multimodal detection and tracking system for road monitoring is proposed. The CNN-based Faster R-CNN algorithm is employed for multimodal detection. Meanwhile, the KCF method is applied for tracking. The experimental results show that the proposed system can automatically detect multimodal from different scenarios and track those objects robustly in the majority of scenes. Future research will be focusing on detecting objects at night and tracking the temporary occluded objects.

ACKNOWLEDGEMENTS

This work is partially supported by the Beijing Municipal Science and Technology Project under Grant #D161100004116001. The authors would also like to thank the insightful and constructive comments from anonymous reviewers.

REFERENCES

Alan, M. M. (2001). "Background subtraction techniques." *Proc of Image & Vision Computing*, 2(2), 1135–1140.

Angel, A., Hickman, M., Mirchandani, P., and Chandnani, D. (2003). "Methods of analyzing traffic imagery collected from aerial platforms." *IEEE Transactions on Intelligent Transportation Systems*, 4(2), 99–107.

Bradski, G. R. (1998). "Real Time Face and Object Tracking as a Component of a Perceptual User Interface." *Applications of Computer Vision, 1998. WACV '98. Proceedings. Fourth IEEE Workshop on* (pp. 214).

Canny, J. (1986). "A computational approach to edge detection." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 8(6), 679–98.

Clippard, D. (1996). "Overhead microwave detectors cut out the loop." *Traffic Technology International 96 Annual Review Issue*.

Dong, Y., Wang, C., Xue, L., and Ming, Y. (2013). "Pedestrian detection and tracking based on tld framework." *Journal of Huazhong University of Science & Technology*.

Girshick, R. (2015). "Fast R-CNN." *IEEE International Conference on Computer Vision*, IEEE, 1440–1448.

Hare, S., Saffari, A., and Torr, P. H. S. (2011). "Struck: Structured output tracking with kernels." *International Conference on Computer Vision*, 23, 263–270.

Henriques, J. F., Rui, C., Martins, P., and Batista, J. (2015). "High-speed tracking with kernelized correlation filters." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(3), 583–596.

Kalal, Z., Mikolajczyk, K., and Matas, J. (2012). "Tracking-learning-detection." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 34(7), 1409–1422.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems*, 25(2), 2012.

Lécun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, 86(11), 2278–2324.

Liu, Z., Chen, Y., and Li, Z. (2009). "Camshift-Based Real-Time Multiple Vehicle Tracking for Visual Traffic Surveillance." *CSIE 2009, 2009 WRI World Congress on Computer Science and Information Engineering, March 31 - April 2, 2009, Los Angeles, California, USA*, 5, 477–482.

Lucas, B. D., and Kanade, T. (1981). "An iterative image registration technique with an application to stereo vision." *International Joint Conference on Artificial Intelligence*, 73, 285–289.

Masher, D. P. (1975). *Digital loop detector system*.

Meng, C., Yanmei, Y. U., and Cai, C. (2015). "Design of vehicle and parking cell detection system based on computer vision." *Journal of China Maritime Police Academy*.

Mimbela, L. E. Y., and Klein, L. A. (2000). "Summary of vehicle detection and surveillance technologies used in intelligent transportation systems." *Vehicle Detectors*.

Nagel, H. H., & Haag, M. (1998). "Bias-corrected optical flow estimation for road vehicle tracking." *International Conference on Computer Vision*, IEEE, 8, 1006.

Osuna, E., Freund, R., and Girosit, F. (1997). "Training support vector machines: an application to face detection." *Computer Vision and Pattern Recognition, 1997. Proceedings. 1997 IEEE Computer Society Conference*, IEEE, 130–136.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). “Faster r-cnn: towards real-time object detection with region proposal networks.” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1–1.

Sotelo, M. A., Parra, I., Fernandez, D., and Naranjo, E. (2006). “Pedestrian detection using SVM and Multi-Feature combination.” *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, 103–108.

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features." *Proc Cvpr*, 1, 511.

Wang, C. (2011). “Moving vehicle detection combined contourlet transform with frame difference in highways surveillance video.” *Advances in Electrical Engineering and Electrical Machines*, Springer Berlin Heidelberg.

Zeiler, M. D., and Fergus, R. (2013). “Visualizing and understanding convolutional networks.” *European Conference on Computer Vision*, 8689, 818–833.

Zhan, C., Duan, X., Xu, S., Zheng, S., and Min, L. (2007). “An improved moving object detection algorithm based on frame difference and edge detection.” *International Conference on Image and Graphics*, 519–523.

Zou, B., Liu, Y., and Zhou, X. (2013). “A robust vehicle tracking for real-time traffic surveillance.” *20th ITS World Congress.*