A Recurrent Neural Network Approach to Rear Vehicle Detection Which Considered State Dependency

Kayichirou Inagaki
Department of Engineering and Computer Science, Chubu University
Kasugai, Aichi, JAPAN

Shozo Sato
Research Institute of System Sciences, Nihon Fukushi University
Handa, Aichi, JAPAN

Taizo Umezaki
Graduate School of Engineering, Nagoya Institute of Technology
Nagoya, Aichi, JAPAN

Abstract

Experimental vision-based detection often fails in cases when the acquired image quality is reduced by changing optical environments. In addition, the shape of vehicles in images that are taken from vision sensors change due to approaches by vehicle. Vehicle detection methods are required to perform successfully under these conditions. However, the conventional methods do not consider especially in rapidly varying by brightness conditions. We suggest a new detection method that compensates for those conditions in monocular vision-based vehicle detection. The suggested method employs a Recurrent Neural Network (RNN), which has been applied for spatiotemporal processing. The RNN is able to respond to consecutive scenes involving the target vehicle and can track the movements of the target by the effect of the past network states. The suggested method has a particularly beneficial effect in environments with sudden, extreme variations such as bright sunlight and shield. Finally, we demonstrate effectiveness by state-dependent of the RNN-based method by comparing its detection results with those of a Multi Layered Perceptron (MLP).

Keywords: Recurrent Neural Network, Vehicle Detection, State Dependency, Back Propagation Through Time

1. INTRODUCTION

In the context of Intelligent Transportation Systems (ITS), vision-based detection is watched with keen interest as an effective method in cost and adaptability of various environments. Techniques of image understanding and artificial neural networks have been used extensively in vision-based systems. In vision-based vehicle detection, researchers often encounter variations in visibility conditions, weather, and perspective, which complicate the task of vehicle detection. A design of the vision-based vehicle-detection systems that successfully addressed these crucial challenges would be of great practical value.

Several approaches to resolving these issues have been proposed in previous research [1−4]. M. Betck et. al. [2] applied the symmetry algorithm that detects the rear lights to detection under condition of reduced visibility and night. R. Cucchiara et. al suggested a hybrid method consisting of day and night detection [3]. Z. Sun et. al. employed a *Low Light Camera System* and support vector machine for detection [4]. Many of the conventional approaches are based on detection of each image frame such as a single frame and a very short sequence. The reliability of these approaches, however, is uncertain under visibility conditions that are reduced in consequence of bright sunlight and halation such that the target vehicle is lost sight.

We propose a new way of addressing those conditions through a monocular vision-based vehicle detection system. Our method uses a Recurrent Neural Network (RNN), which has been used elsewhere to analysis signals with the time variation, for vehicle detection. Presented consecutive scenes including target vehicles in learning, the RNN is able to respond to the target and can track its movements using network outputs depend on the past network states by the feedback loops between neurons. Namely, the proposed method is regarded as a sequence detection-based method. In addition, this network is able to cope with features such as changing shapes and activity of the vehicle in time. Therefore, our method resolves the miss-segmentation in rapidly varying brightness conditions and perspectival changes in shapes.

The rest of paper is organized as follows: The RNN-based vehicle detection method is detailed in Section II. A description of the data used in learning and detection experiments is given in Section III. In Section IV, we provide an experimental demonstration of the detecting efficiency of suggested method.

2. RNN-BASED DETECTION

2.1 RNN-based Detection

The RNN is suited for signal processing that involves time variation. Especially, *Jordan network* [5] and *Elman network* [6] which are proposed as the network model, can provide spatiotemporal processing and are often exploited in that field.

Shown in Figure 1 is the architecture of the RNN that we selected in this paper. The network is composed of input, hidden and output units in a layered structure with inter-layer connections, similar to a multi-layered perceptron (MLP). The difference is that the RNN has feedback connections from the output units to the hidden units. Thus, there are two types of neurons in the network: feed-forward and feedback units. The states of the feedback unit are computed by the following differential equation:

$$\tau_i \frac{dx_i(t)}{dt} = -x_i(t) + \sum_j w_{ij} y_j(t) + I_i(t)$$
$$y_i(t) = F_i(x_i(t)) \quad (1)$$

where $x_i, y_j, \tau_i, w_{ij}, I_i, F_i$ denote the i th neuron' state, the output state, the time constant, the i th neuron's connective weight with the j th neurons, the external inputs, and the unit's squashing function, respectively. Equation (1) is referred to as "the Additive Net". The feed-forward unit is defined by Eq. (1) when $\tau_i = 0$. The target signals $O_i(t)$ are presented to the output units, and the set of the output units is indexed with V. We selected squares error between outputs $y_i(t)$ and the target signals $O_i(t)$ integrated from time T_1 to T_2 as a cost function computed by the following equation:

$$E = \int_{T_1}^{T_2} dt \sum_{i \in V} \frac{1}{2} (y_i(t) - O_i(t))^2 \quad (2)$$

We adopted the Back Propagation Through Time (BPTT) learning procedure based on variation method [6] for modifications of the connective weights. In learning, the learning coefficient was set at 0.001.

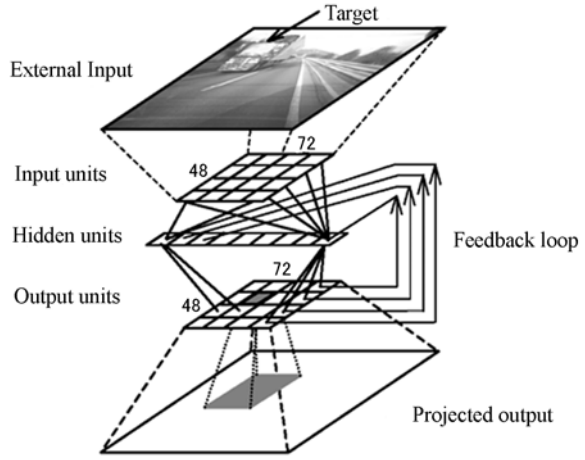

Fig.1 The architecture of RNN

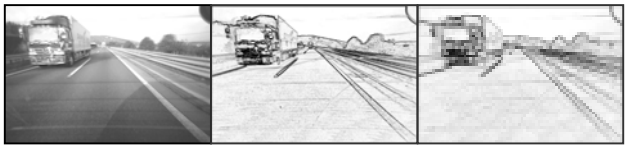

Fig.2 An illustration of feature extraction data, which are presented on input units.

Shown in Figure 2 are feature extracted data before learning, which are presented to input units in learning. In preprocessing, the horizontal and vertical edges of each frame image in the learning data were extracted by Sobel filtering and the image is compressed into one tenth of the original image size in order to detect the target vehicle easily.

Therefore, the network structure of input and output units applied to the learning and detection consisted of 3456 (72×48) units. Before learning, the connective weights were set at a random value [-1.0 – 1.0].

2.2 Learning and Target Pattern

We chose 12 learning scenes that had color variation and trajectory variation of vehicle movement as the learning data. Shown in Figure 3 is a sample learning scene. In Fig.3, the black region in target data (c) corresponds to the vehicle region, the value of which was set as "1". The region corresponding to the background was set as "0". A random interval of frames was sampled from the scene of learning data and presented to the input units; this process was repeated until the end of each scene in the learning process. Whenever the learning scene changes, the network states were reset to its initial state.

2.3 Standards of Detection

The learning process is truncated at an appropriate timing yielded by the following equation:

$$E_t = \frac{1}{N} \sum_{t=0}^{N} \frac{1}{2} (R_{out}(t) \oplus R_{cri}(t)) \quad (3)$$

where $R_{out}(t), R_{cri}(t), N$ denote the region where the outputs $y_i \geq 0.2$, the criterion region of the target and the number of data frames used, respectively. When E_t reaches the threshold during learning, it is hypothesized as the truncation point.

In evaluation, the success criterion of vehicle detection is adopted from the result of the following equation:

$$R_{eval} = \frac{R_{out} \cap R_{cri}}{R_{cri}} \quad (4)$$

The detection of a frame in which the matching rate R_{eval} satisfied the threshold level ($th = 0.5$) was judged a success.

3. DATABESE

The vehicle database consisted of scenes taken from a CCD video camera on Japanese highways at daytime. The acquired images were 256 gray level, and 720×480 pixels in size. We chose 12 and 5 independent scenes as the learning and evaluation data set, respectively. The learning data included vehicle color variation, trajectory variation, lane changing, and approaches by vehicle from behind. About 11% (19/172) of the frame images in the entire evaluation data were ones in which the following vehicle was indistinct due to sunlight.

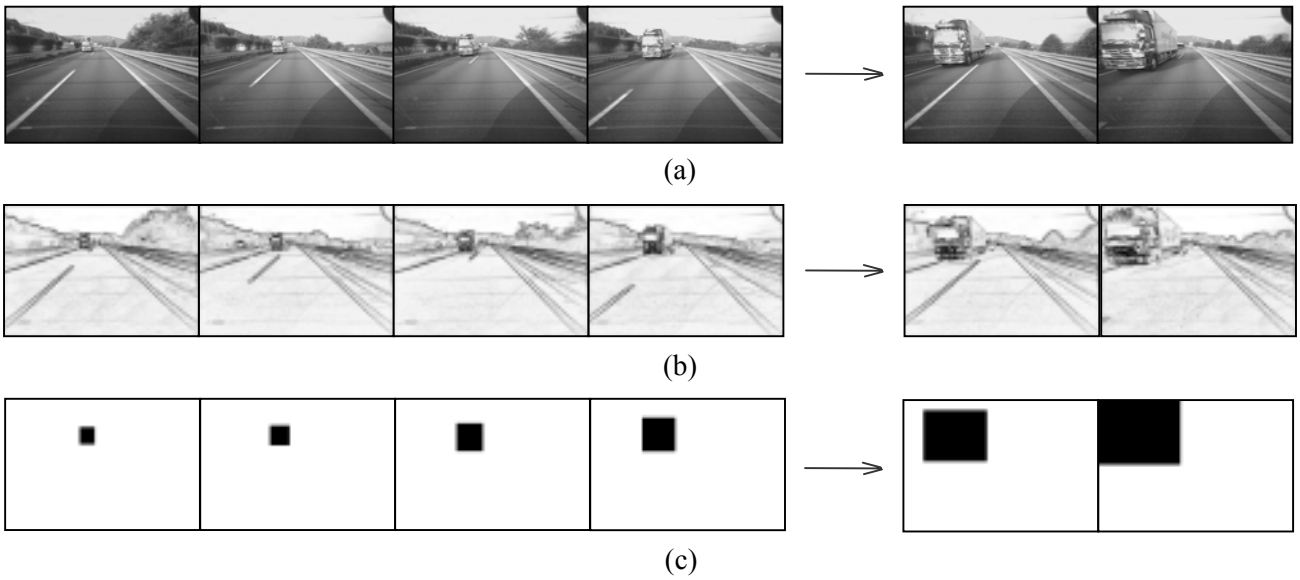

Fig.3 Learning a scene, which has time variations,: (a) shows original image, which is taken from a monocular vision sensor. (b) is an input scene, which is extracted feature by Sobel filtering and image compression. (c) is a target data, which are presented to the output units. The black region correspond to the vehicle region, and the other region correspond to the background

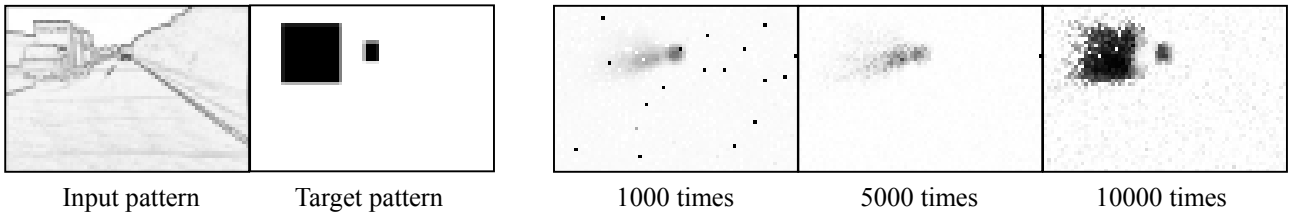

Fig.4 Output states of each learning time (1000, 5000, 10000). In the 1000–times result, some isolated pixels were observed, and the network reacted to the high density region of the learning signals. In 5000–times, the isolated pixels have disappeared. By 10000–times learning, the appropriate response was observed in the low–density region.

In these situations, conventional methods based on single-frame detection, rather than sequential detection, are unable to execute robustly.

4. RESULTS AND DISCUSSION

4.1 Learning Process

In learning, the RNN trains for the desired response of the learning data for vehicle detection. This section describes its learning process. Figure 4 and Figure 5 show the output states of each learning time, and the distribution of the target signal of the learning data that we chose, respectively. As shown in Fig.4, the output states after 1000-times has the some isolated pixels; the network reacted appropriately to the high-density region of the learning data, but not to the low density region. In 5000-times learning, the isolated pixels had disappeared; however, the network did not react appropriately to the low density region.

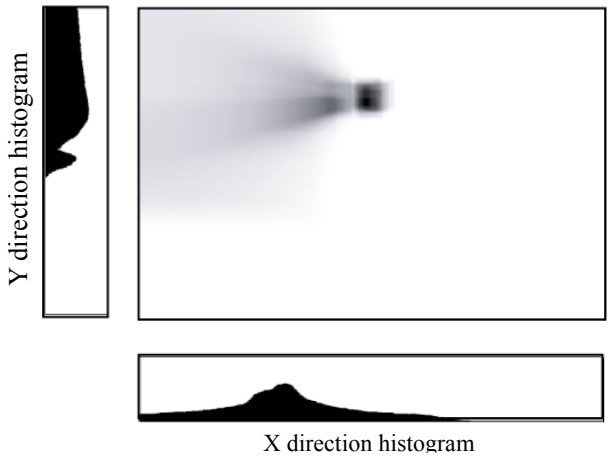

Fig.5 An illustration of the distribution of target signals and XY direction histogram of target distribution. The region where it is shown black are high density.

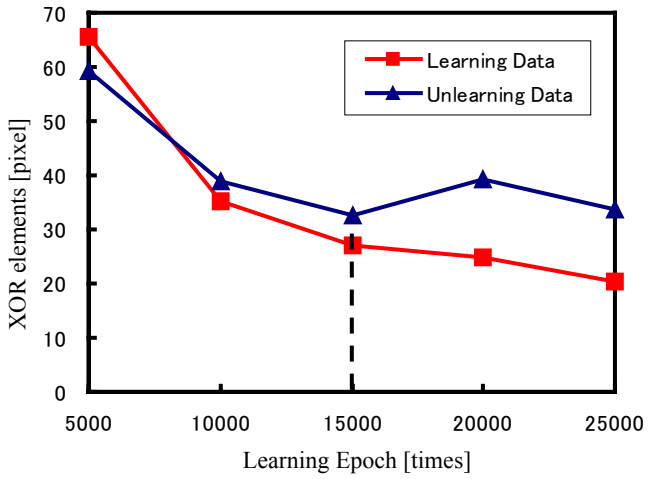

Fig.6 Illustration of result of XOR between RNN outputs and the target signals. A rectangular symbol denotes the evaluation findings on learning data. A triangular symbol denotes the evaluation findings on unlearned data.

After 10000-times learning, the appropriate response was observed in the low-density region. The network outputs of 10000-times learned RNN resembles the target pattern. These results denote that the RNN approximated the learning pattern variation.

In order to decide the learning truncation point, the error between the RNN outputs and target data were evaluated by Eq.(3). Shown in Figure 6 is the evaluation result of XOR. In the learning data, XOR appears to decrease linearly with an increasing learning epoch. The evaluation result on unlearned data behaves similarly to learning data; however, the tendency in which the data decreases is not observed after 15000 times. Therefore, we define that the truncation point in learning is 15000 times, and the15000-times learned RNN could be used effectively in vehicle detection.

4.2 Detection Result

After learning, an RNN is chosen that indicates the best performance from RNN's learned under the conditions of the number of hidden units set at 48, 64, 96, and 128. Shown in Figure 7 is an example of the detected results under the rapid varying brightness conditions due to the sunlight effect. In Fig.7, the frame (b) displays a sudden brightness variation, but the network reacted in the target region. The succeeded detection is considered to be due to the RNN's state dependency. Furthermore, vehicle detection time per one frame was 134[ms] in ALPHA-CPU 600MHz

We defined the disparity in brightness between two frames, difference *D*, by the following formula intended to analyze the network response under rapidly varying brightness conditions.

$$D = \frac{|G_t - G_{t-1}|}{M} \tag{5}$$

where, *G* and *M* denote the image frame (current: *t*, before: *t*-1), and all pixels in the image, respectively. Shown in Figure 8 is a result of vehicle detection with varying visibility conditions using the learned network. In Fig.8, the matching rate between the network outputs and the target R_{eval} is computed by Eq.(4). The gray areas in Fig.8 illustrate that the brightness variation computed by Eq.(5) is large. Although the following vehicle in the frame image becomes hard to distinguish from the background in those areas, R_{eval} doesn't decrease; that is, the detection is successful. Therefore, detection under the reduced visibility conditions is remarkably successful: by considering the previous network states, the RNN is able to regard the vehicle activity as a sequential signal.

4.3 MLP Based Detection vs. RNN Based Detection

In this section, we compare the MLP-based method as an example of single frame detection and the suggested RNN-based method as an example of consecutive frame detection with a state-dependency of the past network states. In other words, this investigation measured the effectiveness of state dependency. The MLP was as follows: input layer, hidden layer, and output layer have 1125 (45×25) units, 100 units, and 1125 (45×25) units, respectively. The learning coefficient in weight modification was set at 0.3. These parameters were obtained in pre-experimentation. The back propagation method [8] was applied for MLP learning. The 18 learning data including color variation and the vehicle shapes variation were selected as the learning data of MLP. The input data were the extracted feature presented to the input layer. In learning, the input and target data were shifted by minute random digits at every learning time. This process was intended to inhibit the fluctuation of the camera angle of the view due to the road conditions. Shown in Figure 9 is an example of the detection results with the MLP-based and RNN-based methods. The MLP-based method reacts to the vehicles traveling in the opposite direction; that is, this detection fails. This result is attributed to the fact that the MLP-based method recognizes only a single frame of the consecutive scenes. With the RNN-based method, only the appropriate response for the vehicles moving forward is observed. The experimental comparison between the MLP-based and the RNN-based methods indicate that the sequential detection method with a state dependency is effective in reducing miss-perception in vehicle detection.

We investigated the detection ability of the MLP and RNN –based detection methods in relation to unlearned data. The evaluation data consisted of 5 unlearned scenes of 624 image frames total. The matching rate between the criterion data and the network response was calculated by Eq.(4). Shown in Figure 10 are the detection rates of the MLP-based and RNN-based methods. A detection rate of 93.4% was obtained for the unlearned data in 64 hidden units, while the MLP method yielded a maximal 80% detection rate in pre experimentation. The RNN-based method's superiority is attributed to the fact that RNN output reflects the past states of the network. With the RNN-based method, increasing of the disparity and decreasing of the network output were observed with 48 hidden units and lower. On the other hand, when 96 hidden units and upper, the network tended to miss-segmentation in background and regions where vehicles do not exist. Furthermore, the network detection failed in cases when the vehicle strongly deviated from the learned moving trajectory. This problem warrants future work.

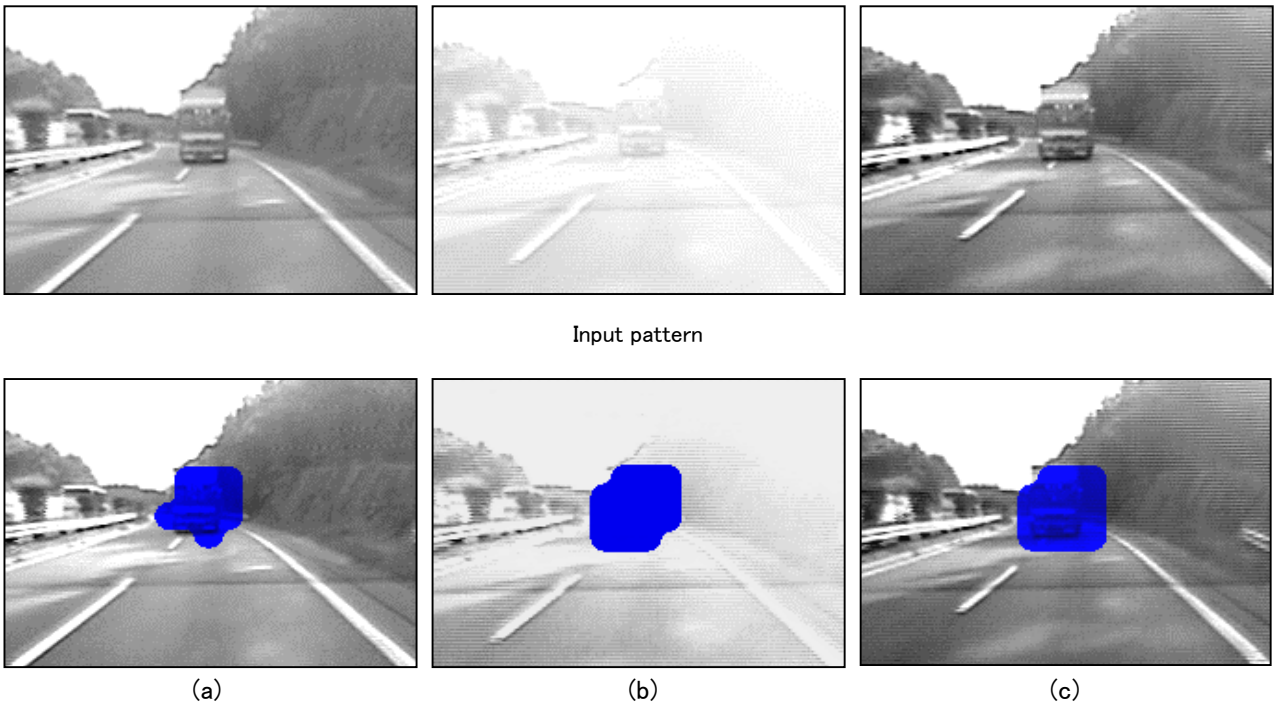

Fig.7 Illustration of detection results under the sudden variation of brightness condition due to sunlight. Frames (a) and (c), which are the result before varying brightness condition and after varying one are detected result. Frame (b) shows a dramatic variation due to sunlight.

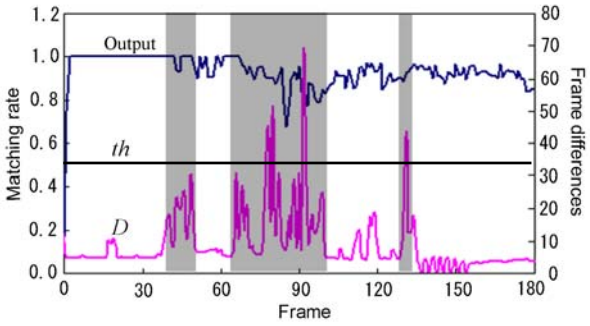

Fig.8 Example of vehicle detection results with varying brightness conditions. The gray areas show where the brightness variation is large. The line *th* designates the threshold level for success detection.

Fig.9 Illustration of the comparison between the MLP-based and RNN-based detection methods. (a)Detection result with the MLP (conventional) method. (b)Detection result with the RNN (suggested) method.

5. CONCLUSION

In this paper, we suggested a new vehicle detection method employing a recurrent neural network, and investigated its detection ability. The RNN is able to react to consecutive scenes including target vehicles and can track the movements of targets due to network outputs depending on the past network states found in the feedback-loops between neurons. Therefore, the RNN is an appropriate detection method for feature variations within a time series. Experimental results indicated that the RNN-based detection was remarkably successful for frames in which there was sudden variation of brightness due to reflected sunlight. This result is attributed to the RNN's ability to exploit features hidden in the time series and the state dependency.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewer for valuable comments. This works is supported by OMRON corporation. The vehicle data were provided by Takashi Nose.

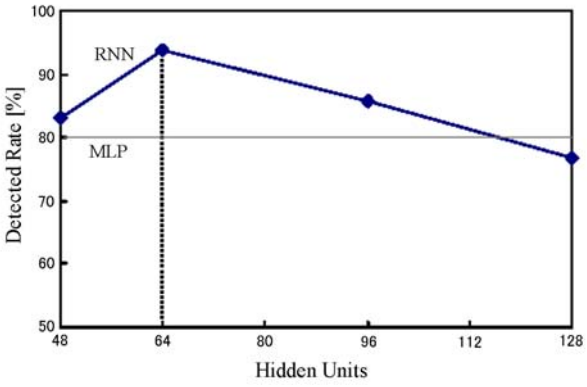

Fig.10 The detection rate of MLP-based and RNN based methods. With the MLP-based method, the maximal detection rate in pre-experimentation was 80.0%. The maximum detection rate of the RNN-based method was 93.4%.

REFERENCES

[1] P. H. Batavia, D. A. Pomerleau, C. E. Thorpe, "Detecting Overtaking Vehicles With Implicit Optical Flow," CMU-RI-TR-97-28, Mar. 1998.

[2] M. Betck, and H. Nguyen, "Highway Scene Analysis from a Moving Vehicle under Reduced Visibility Conditions," *IEEE International Conference on Intelligent Vehicles*, pp. 131-136, Oct. 1998.

[3] R. Cucchiara, M. Piccardi, "Vehicle Detection under Day and Night Illumination," *in Proc. of IIA'99-Third Int. ICSC Symp. on Intelligent Industrial Automation, Special Session on Vision Based Intelligent Systems for Surveillance and Traffic Control*, pp. 789-794, 1999.

[4] Z. Sun, R. Miller, G. Bebis, D. DiMeo, "A Real-Time Precrash Vehicle Detection System," *IEEE Workshop on Application of Computer Vision, Orland*, FL, 2002.

[5] J. L. Elman, "Finding Structure in Time," *Cognitive Science*, vol.14, pp.179-212, 1990.

[6] M. I. Jordan, "Attractor Dynamics and Parallelism in a Connectionist Sequential Machine," *in Proc. the Eighth Annual Conference of the Cognitive Science Society*, Amherst, MA, 1986, pp. 531-546.

[7] M. Sato, "A Learning Algorithm to Teach Spatiotemporal Patterns to Recurrent Neural Networks," *Biological Cybernetics*, vol. 62, pp. 259-263, 1990.

[8] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, pp. 533-536, Oct. 1986.