

Online vehicle detection using gated recurrent units

Arezoo Sedghi¹[0000-0002-1752-6303] and Esmat Rashedi²[0000-0002-2539-5817]

And Maryam Amoozegar³[0000-0001-7161-8623] and Fatemeh Afsari⁴[0000-0003-4165-5233]

¹ Faculty of Electrical and Computer Engineering, Graduate University of Advanced Technology, Kerman, Iran
a.sedghi@student.kgut.ac.ir

² Faculty of Electrical and Computer Engineering, Graduate University of Advanced Technology, Kerman, Iran
e.rashedi@kgut.ac.ir

³ Department of Computer and Information Technology, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran
amoozegar@kgut.ac.ir

⁴ Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran
afsari@uk.ac.ir

Abstract. As road transportation increases and inner and outer city roads are extended, traffic control systems and intelligent monitoring of traffic flow are crucial. Vehicle detection is one of the primary tasks in computer vision applications, which is widely used in surveillance-related tasks, especially in intelligent traffic monitoring systems. To address this issue, a number of approaches have been presented out, the most of which are based on deep learning frameworks. We propose a method that solves the vehicle detection problem by online detection without needing labeled data so it can be used for traffic flow supervision. The proposed approach utilizes the deep gated recurrent neural networks that can extract moving objects (vehicles) from the original scenes by applying background model maintenance. The method is evaluated on BMC database of traffic videos. As the evaluation process is carried out from the first frame to the end of the video, the experimental results are presented by the well-known metrics and visually demonstrated, which is remarkable considering the unsupervised training and real-time performance.

Keywords: Vehicle detection, Background model maintenance, Deep learning, Gated recurrent unit, Online training, Traffic surveillance systems.

1 Introduction

Vehicle detection is essential in computer vision with various applications, such as self-driving cars [1], traffic flow monitoring [2], parking lot management [3], video

surveillance [4], and robotics [5]. Object detection, especially vehicle detection, is a challenging problem due to the complex and diverse appearance of vehicles and the variability in the environmental conditions in which they operate such as different illumination conditions, dynamic backgrounds, etc. From traditional machine learning algorithms [6] to deep learning-based approaches [7], numerous methods have been developed over the years to detect vehicles and moving objects. In recent years, deep neural networks have demonstrated superior performance in detecting vehicles in images and videos. In the following, some recent researches are reviewed on object detection especially vehicle detection using deep neural networks.

In particular, convolutional neural networks (CNNs) have become the state-of-the-art method for object detection due to their ability to automatically learn high-level features from raw image data. One of the most influential CNN-based object detection frameworks is the region-based CNN (R-CNN) family [8]. Since its introduction in 2014, this family of models has undergone several improvements, including Fast R-CNN, Faster R-CNN [9], and Mask R-CNN [10], which have significantly improved the accuracy and speed of object detection. In [11], the authors present an end-to-end generative adversarial network (GAN) named RMS-GAN for moving object segmentation. RMS-GAN consists of two generators, each one uses a different recurrent technique for foreground probability map estimation. A cascaded approach is incorporated to enhance the spatial coherence of the estimated foreground probability maps with the help of generator-1. Generator-2 uses a decoder-encoder feature fusion and previous frame outputs to estimate the current frame foreground objects. In [12], a recurrent encoder-decoder network with a convolutional long short-term memory (LSTM) for dense pixel-wise prediction of video frames was proposed. The network compresses the spatiotemporal features at the encoder and restores them to the original size at the decoder. The LSTM successfully learns the spatiotemporal relation with relatively fewer parameters, enabling the network to handle the long-term dependencies of temporal events. In [13], the authors studied various methods for object detection in videos, specifically those using Recurrent Neural Networks. They compared feature-based methods, box-level methods, and flow networks. The study found that incorporating temporal context into object detection is beneficial, and provides conclusions and guidelines for video object detection networks.

What motivates us to choose GRUs? Recurrent Neural Networks suffer from short-term memory. If a sequence is long enough, it may have difficulty conveying information from one time step to another. The vanishing gradient problem is a common issue in recurrent neural networks during backpropagation. In this process, the neural network weights are updated through gradients that backpropagate through time. However, in some cases, these gradients can shrink significantly, resulting in vanishing gradients. When the gradient values become too small, they do not contribute much to the learning process. By using Gated Recurrent Units (GRUs), Cho et al [14] solve the vanishing gradient problem that occurs in conventional recurrent neural networks by improving a standard recurrent neural network. In a sequence, it is possible for GRUs to identify which data should be saved or discarded, and to make predictions, it can transmit pertinent information.

In this paper, a model containing online gated recurrent units is proposed for detecting vehicles in traffic surveillance videos. The following, methodology section

describes the proposed method in three main subsections. In Section 4, network implementation is described in detail. Furthermore, quantitative and quantitative results are reported and the proposed model is compared with state-of-the-art methods. Finally, the main conclusions and future scopes are outlined.

2 Methodology

In this paper, a method is proposed based on deep neural networks and unsupervised learning techniques to solve the moving object detection (MOD) problem. It has been aimed to have online performance and be able to face real-world situations such as real-time traffic sequences from surveillance cameras, accordingly it is designed unsophisticated computationally. As it can be seen in Figure 1, the data is given to the model as a grayscale and converted to a one-dimensional signal before entering the network. Each frame is presented to the model one at a time during training, and it does not see any repeated frames. Deep neural networks of the type of gated recurrent units are employed, which are capable of predicting and estimating the constant patterns and low dimensional components. A fully connected layer comes last. We implement the proposed online and GRU-based approach, which we refer to as OGRU, in three principal phases.

2.1 Background Model Maintaining

Our proposed MOD approach works based on background model maintenance. Over time, the model learns the background pattern and maintains the estimated model of the background scene that is obtained by feeding the original scene to the GRU cells. It is possible to train GRUs to maintain information from the past, without losing its impact on network training over time, as well as to remove extraneous information from the estimation, using their update and reset gates. As the background scene holds a constant pattern over time and has the least changes, the possibility of background estimation exists for GRUs. Output of the last GRUs layer is passed to a fully connected layer which produces output that is the background.

2.2 Moving Object Detection

In the following step, moving objects considered as foreground are extracted by thresholding the difference between entered scenes and the corresponding estimated background to determine the movements of objects that can be detected precisely.

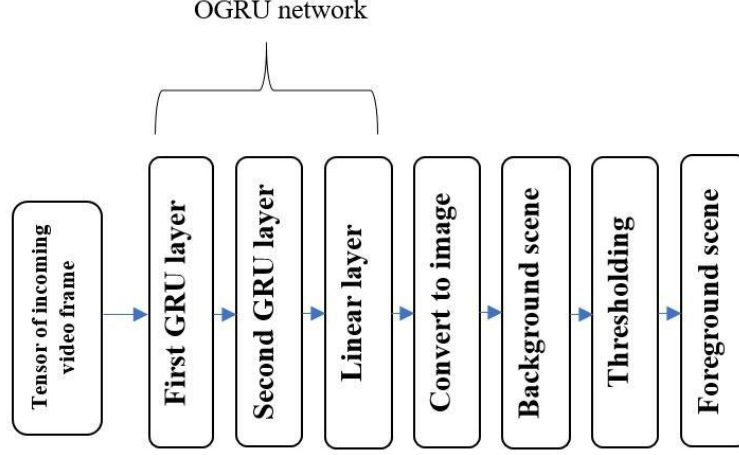


Fig. 1. The architecture of the proposed model.

2.3 Online Training

The training process of the proposed method is performed using the backpropagation algorithm. In this way, solving the optimization problem by setting the loss function as the RMSE function can train the model by updating the weights of its neurons. The purpose of each training step of this procedure is to reduce the difference between the current estimated background scene and the previous one. Here is the pseudo-code that shows how the proposed model is trained online.

Algorithm 1: The online training procedure of OGRU

Input: Incoming frame $X_t \in X = \{X_1, X_2, \dots, X_t, \dots, X_N\}$

Output: Background of X_t

Initialization:

1. Background matrix is set as zero.
2. Randomly Initialize the weights of the OGRU randomly.

While True:

1. Background model maintaining:

$$B_t = OGRU(X_t)$$

2. Moving object detection (MOD):

$$F_t = X_t - B_t$$

3. Training:

Minimize the term $RMSELoss(X_t, B_t)$

End of video

3 Evaluation Results and Discussion

3.1 Implementation Details

The proposed deep learning-based model is implemented using Python and the Pytorch framework. Details of feeding the data to the model, the value of parameters, and details for implementations are given in Table 1.

Table 1. Details of the model architecture and hyper-parameter tuning.

Number of GRU layers	2
Hidden size of GRUs	40
Number of linear layers	1
Size of linear layer	307200
Size of video frame	640×480
Size of input data	307200×1
Loss function	RMSE
Optimization algorithm	R-prop[15]
Learning rate	0.02

3.2 Quantitative Analysis

In this study, the BMC dataset [16] is utilized, which consists of 20 synthetic traffic videos for training and testing the proposed model. Each synthetic video in the dataset has a frame size of 480x640 pixels and contains approximately 1500 frames. The synthetic traffic videos in the BMC dataset are designed to simulate real-world scenarios such as urban streets, squares, and highways.

In the evaluation phase of our study, we present a quantitative analysis of the performance of our deep learning-based moving object detection model. To enable the evaluation, we convert the grayscale foreground frames produced by the model to binary masks using a thresholding technique, followed by applying morphological operators to enhance the quality of the masks. This process enables us to calculate the confusion matrix and obtain quantitative evaluation metrics.

In this study, we evaluate the performance of our deep learning-based moving object detection model using a variety of commonly used evaluation metrics. These metrics include recall, precision, F-measure, intersection over union (IoU), Matthews

correlation coefficient (MCC), and Detection Rate (DR). Recall, precision, and the most important metric, F-measure are computed based on [16]. F-measure is the harmonic mean of recall and precision, providing a balanced measure of both metrics, and is used in the evaluation of most moving object detection algorithms. Intersection over Union (IoU) measures the overlap between the predicted moving object and the ground truth. Matthews correlation coefficient (MCC) is a correlation coefficient between predicted and ground truth labels.

Finally, the Detection Rate measures the proportion of frames in which a moving object is correctly detected. By using these evaluation metrics, we aim to provide a comprehensive assessment of the performance of our model and to compare it to other existing approaches in the field of moving object detection. In Table 2, the results are reported. The evaluation process is done from the first frame to the end of the video.

Table 2. Evaluation results on the synthetic videos of BMC dataset as percentages.

Video ID	Recall	Precision	F-Measure	IoU	MCC	DR
111.mp4	61.55	88.79	75.62	19.22	45.28	23.10
112.mp4	60.69	92.77	75.30	20.40	44.98	22.66
121.mp4	57.38	87.65	72.03	12.38	36.97	14.77
122.mp4	55.95	88.31	68.28	11.54	28.91	13.03
211.mp4	61.30	88.19	75.38	18.64	44.58	22.60
212.mp4	60.50	92.11	74.98	19.95	44.29	22.21
221.mp4	61.02	90.17	75.96	18.61	46.55	22.34
222.mp4	57.55	95.16	71.60	14.87	35.52	16.23
311.mp4	61.98	88.14	76.10	19.68	14.18	23.97
312.mp4	61.56	90.29	75.44	21.94	46.48	25.05
321.mp4	61.58	90.51	76.52	19.46	47.63	23.17
322.mp4	58.73	94.38	72.33	17.17	37.65	18.95
411.mp4	60.54	85.38	73.34	17.13	40.56	21.10
412.mp4	60.23	82.50	70.06	17.94	35.46	23.63
421.mp4	59.73	85.13	71.79	16.19	38.16	19.51
422.mp4	56.76	89.96	69.27	12.88	31.25	16.70
511.mp4	60.58	63.40	62.77	11.94	24.23	21.26
512.mp4	60.31	79.69	68.64	17.61	33.73	21.94
521.mp4	61.03	86.24	73.00	19.11	41.85	22.06
522.mp4	57.55	90.59	70.08	14.61	33.74	16.28
Average	59.82	87.46	72.42	17.19	37.59	20.51

In order to have a comprehensive evaluation, we compare the OGRU method with some of the state-of-the-art methods evaluated on 10 videos of the BMC dataset. The results are shown in Table 3. As the results indicate, our method outperforms some methods and the average F-measure is equal to or slightly different from others. Considering that our method is not time-consuming and has low computational complexity compared to classic methods and it is trained online and evaluated

simultaneously from the first frame of the video to the end, during the unsupervised training. The proposed method is able to process high-resolution and long video sequences.















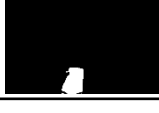
Table 3. Reported F-measures for state-of-the-art algorithms on synthetic videos of BMC in [17].

Algorithm	F-measure	Algorithm	F-measure
RLS [18]	0.712	Variational BRPCA [19]	0.70
PCA [20]	0.701	MOG-RPCA [21]	0.757
EALM [22]	0.649	IALM [22]	0.711
Go-Dec [23]	0.732	SemiSoft GoDec [23]	0.726
BLWS [24]	0.610	ROSL [25]	0.754
FAM [26]	0.713	GRASTA [27]	0.618
NSA [28]	0.730	pROST [29]	0.718
PSPG [30]	0.698	GOSUS [31]	0.714
TFOCS [32]	0.735	PRMF [33]	0.727
ALM [34]	0.649	Adaptive MOG [35]	0.698
Bayesian RPCA [36]	0.643	OGRU (Proposed model)	0.7159

3.3 Qualitative Analysis

To visually assess the performance of our deep learning-based moving object detection model, we provide qualitative analysis through images. We demonstrate the effectiveness of our model in detecting vehicles in different scenarios, including scenarios such as streets, squares, foggy or polluted weather, dynamic background, and different sizes of vehicles. We present sample frames from the dataset, where OGRU model has successfully detected and tracked vehicles. The visualizations include the predicted foregrounds overlaid on the original frames and their corresponding ground truth demonstrating the ability of our model to capture the shape and motion of moving objects. Table 4 shows the visual results of the proposed model.

Table 4. Qualitative analysis of the OGRU model.

Video ID	Original Scene	Ground truth	OGRU
111.mp4			
121.mp4			
122.mp4			
322.mp4			
412.mp4			

4 Conclusion and future scope

This paper presents a new method using deep recurrent neural networks for detecting vehicles in real-time moving object detection applications. The proposed approach is unsupervised and estimates the background model of each incoming video frame by utilizing a GRU-based network. The model learns the background scene pattern by receiving video frames sequentially and maintains the estimated background image using GRU cells. The model is trained until the end of the task, resulting in improved effectiveness. The BMC2012 benchmark dataset's traffic surveillance videos are used to train and evaluate the proposed model, which demonstrated strong performance. Since the method is unsupervised and operates frame by frame, it is suitable for real-time applications. In the future, it is planned to enhance the OGRU model to make it more robust to sudden changes, outliers, and noise.

References

- [1] W. Farag and Z. Saleh, "An advanced vehicle detection and tracking scheme for self-driving cars," in *2nd Smart Cities Symposium (SCS 2019)*, 2019, pp. 1–6.
- [2] P. P. Tasgaonkar, R. D. Garg, and P. K. Garg, "Vehicle detection and traffic estimation with sensors technologies for intelligent transportation systems," *Sens. Imaging*, vol. 21, pp. 1–28, 2020.
- [3] L. Lou, Q. Li, Z. Zhang, R. Yang, and W. He, "An IoT-driven vehicle detection method based on multisource data fusion technology for smart parking management system," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 11020–11029, 2020.
- [4] A. Appathurai, R. Sundarasekar, C. Raja, E. J. Alex, C. A. Palagan, and A. Nithya, "An efficient optimal neural network-based moving vehicle detection in traffic video surveillance system," *Circuits, Syst. Signal Process.*, vol. 39, pp. 734–756, 2020.
- [5] A. Petrovskaya and S. Thrun, "Model based vehicle detection and tracking for autonomous urban driving," *Auton. Robots*, vol. 26, no. 2–3, pp. 123–139, 2009.
- [6] M. Shakeri and H. Zhang, "COROLA: A sequential solution to moving object detection using low-rank approximation," *Comput. Vis. Image Underst.*, vol. 146, pp. 27–39, 2016.
- [7] X. Ou *et al.*, "Moving object detection method via ResNet-18 with encoder–decoder structure in complex scenes," *IEEE Access*, vol. 7, pp. 108152–108160, 2019.
- [8] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

- [9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
- [10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [11] P. W. Patil, A. Dudhane, and S. Murala, "End-to-End Recurrent Generative Adversarial Network for Traffic and Surveillance Applications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14550–14562, Dec. 2020, doi: 10.1109/TVT.2020.3043575.
- [12] S. Choo, W. Seo, D. Jeong, and N. I. Cho, "Multi-scale recurrent encoder-decoder network for dense temporal classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 103–108.
- [13] A. B. Qasim and A. Pettirsch, "Recurrent neural networks for video object detection," *arXiv Prepr. arXiv2010.15740*, 2020.
- [14] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv Prepr. arXiv1409.1259*, 2014.
- [15] M. Riedmiller and H. Braun, "Rprop-a fast adaptive learning algorithm," in *Proc. of ISCIS VII*, Universitat, 1992.
- [16] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievre, "A benchmark dataset for outdoor foreground/background extraction," in *Asian Conference on Computer Vision*, 2012, pp. 291–300.
- [17] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *Comput. Sci. Rev.*, vol. 23, pp. 1–71, 2017.
- [18] F. De la Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, 2001, vol. 1, pp. 362–369.
- [19] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, 2012.
- [20] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, 2000.
- [21] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang, "Robust principal component analysis with complex noise," in *International conference on machine learning*, 2014, pp. 55–63.
- [22] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv Prepr. arXiv1009.5055*, 2010.
- [23] T. Zhou and D. Tao, "Godec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 2011.
- [24] Z. Lin and S. Wei, "A block Lanczos with warm start technique for accelerating nuclear norm minimization algorithms," *arXiv Prepr.*

- arXiv1012.0365*, 2010.
- [25] X. Shu, F. Porikli, and N. Ahuja, “Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3874–3881.
 - [26] P. Rodriguez and B. Wohlberg, “Fast principal component pursuit via alternating minimization,” in *2013 IEEE International Conference on Image Processing*, 2013, pp. 69–73.
 - [27] J. He, L. Balzano, and J. Lui, “Online robust subspace tracking from partial information,” *arXiv Prepr. arXiv1109.3827*, 2011.
 - [28] N. S. Aybat, D. Goldfarb, and G. Iyengar, “Fast first-order methods for stable principal component pursuit,” *arXiv Prepr. arXiv1105.2126*, 2011.
 - [29] C. Hage and M. Kleinsteuber, “Robust PCA and subspace tracking from incomplete observations using ℓ_0 -surrogates,” *Comput. Stat.*, vol. 29, no. 3–4, pp. 467–487, 2014.
 - [30] N. S. Aybat, D. Goldfarb, and S. Ma, “Efficient algorithms for robust and stable principal component pursuit problems,” *Comput. Optim. Appl.*, vol. 58, pp. 1–29, 2014.
 - [31] J. Xu, V. K. Ithapu, L. Mukherjee, J. M. Rehg, and V. Singh, “GOSUS: Grassmannian online subspace updates with structured-sparsity,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3376–3383.
 - [32] S. Becker, E. Candes, and M. Grant, “TFOCS: flexible first-order methods for rank minimization,” in *Low-rank Matrix Optimization Symposium, SIAM Conference on Optimization*, 2011.
 - [33] N. Wang, T. Yao, J. Wang, and D.-Y. Yeung, “A probabilistic approach to robust matrix factorization,” in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VII 12*, 2012, pp. 126–139.
 - [34] G. Tang and A. Nehorai, “Robust principal component analysis based on low-rank and block-sparse matrix decomposition,” in *2011 45th Annual Conference on Information Sciences and Systems*, 2011, pp. 1–5.
 - [35] A. Shimada, D. Arita, and R. Taniguchi, “Dynamic control of adaptive mixture-of-Gaussians background model,” in *2006 IEEE international conference on video and signal based surveillance*, 2006, p. 5.
 - [36] X. Ding, L. He, and L. Carin, “Bayesian robust principal component analysis,” *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, 2011.