# APPEARANCE AND MOTION BASED DEEP LEARNING ARCHITECTURE FOR MOVING OBJECT DETECTION IN MOVING CAMERA

*Byeongho Heo*[⋆]*, Kimin Yun*[†]*, and Jin Young Choi*[⋆]

[⋆]Department of Electrical and Computer Engineering, ASRI
Seoul National University, South Korea
[†]Electronics and Telecommunications Research Institute (ETRI), South Korea
bhheo@snu.ac.kr, kimin.yun@etri.re.kr, jychoi@snu.ac.kr

## ABSTRACT

Background subtraction from the given image is a widely used method for moving object detection. However, this method is vulnerable to dynamic background in a moving camera video. In this paper, we propose a novel moving object detection approach using deep learning to achieve a robust performance even in a dynamic background. The proposed approach considers appearance features as well as motion features. To this end, we design a deep learning architecture composed of two networks: an appearance network and a motion network. The two networks are combined to detect moving object robustly to the background motion by utilizing the appearance of the target object in addition to the motion difference. In the experiment, it is shown that the proposed method achieves 50 fps speed in GPU and outperforms state-of-the-art methods for various moving camera videos.

***Index Terms***— Moving object detection, deep learning, moving camera

## 1. INTRODUCTION

Moving object detection is an important technology for visual surveillance systems, autonomous vehicles, drones, and so on. In moving object detection, there are two kinds of approaches: the object-centric approach [1, 2, 3] and the background-centric approach [4, 5, 6, 7, 8, 9]. Despite the good performance of the object-centric approach, it is not suitable for actual application due to high computational complexity and difficulty of online detection. On the other hand, the background-centric method is suitable for practical systems because of its relatively simple model. In the background-centric approach, the moving object area of the image is defined as the foreground and the other area is defined as the background. The background-centric approach analyzes the background pattern in order to build a background model. The area that is similar to the background model is determined as the background and the other area is determined as the foreground. The Gaussian Mixture Model (GMM) [4, 5, 6, 7], sample consensus [8, 10], and neural network [9, 11, 12] have been used for the background model.

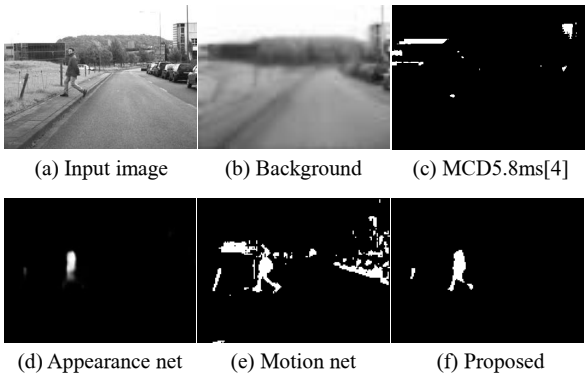


**Fig. 1**. Issue in background modeling in a moving camera. Moving camera image in (a) often yields a noisy background model as shown in (b) due to severe camera motion. The background-centric approach is vulnerable to this situation as shown in (c). However, the proposed method is robust to this situation as shown in (f) by combining the appearance-based result in (d) and motion-based result in (e).

However, the background-centric approach is vulnerable to background contamination due to the camera movements. A background model that does not match the background of the actual image is called a contaminated background model. The contamination occurs when it is difficult to distinguish the foreground from the background, or when the background is moving. In the moving camera environment, the background moves according to the camera movement. Therefore, camera motion should be estimated and the background model should be modified accordingly. Kim *et al.* [5] proposed a motion compensation method to estimate background motion through feature matching. The movement of

This work was partly supported by the ICT R&D program of MSIP/IITP [B0101-15-0552, Development of Predictive Visual Intelligence Technology and B0101-15-0266, Development of High Performance Visual BigData Discovery Platform] and Brain Korea 21 Plus Project.

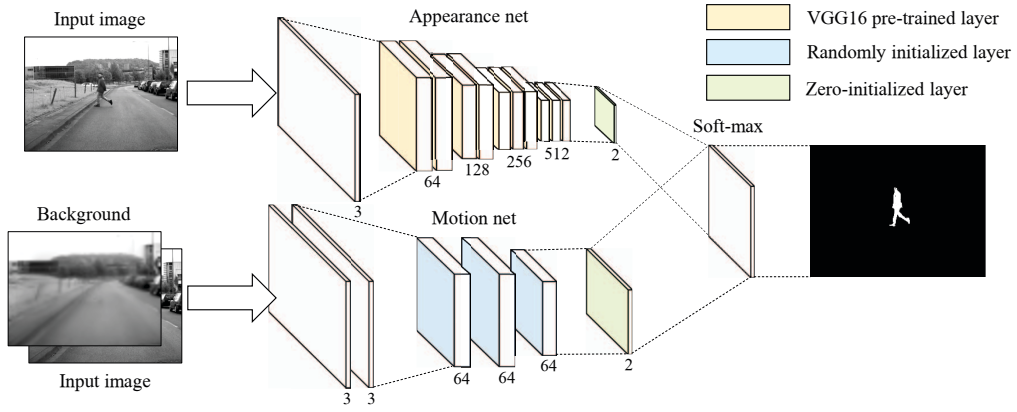


**Fig. 2**. Structure of proposed method. The proposed method consists of the appearance network (A-net) and the motion network (M-net). The A-net detects objects by focusing on appearance in image, and the M-net detects movement by comparing image and background image. The results of the two networks are combined to form a moving object detection result.

the background in time is represented by homography, and the background is warped to follow the camera movement. Yi *et al.* [4] proposed a dual model to prevent contamination of the background model by foreground. Owing to the dual model, the contamination of background by foreground is lessened, which improves the detection performance in the moving background. Yun *et al.* [6] used the spatio-temporal consistency of moving objects to reduce background contamination. These methods compensate for the camera movement in dynamic background modeling. However, camera movement compensation is still incomplete in situations such as a dashcam mounted on a vehicle. Hence the contamination problem remains unresolved in moving camera environments. Fig. 1 shows an example of this situation. Due to inaccurate motion compensation, a noisy background model, as shown in Fig. 1(b), is built. In this case, the existing background-centric algorithm, which is highly dependent on the background model, produces inaccurate detection results as shown in Fig. 1(c). Our goal is to propose a robust real-time moving object detection algorithm to deal with background contamination.

The proposed method is based on deep learning framework and detects moving objects in the moving camera environment. To cope with the background contamination problem, the proposed method adopts appearance features as well as motion features. To this end, we design a deep learning architecture composed of two networks. The first network is the appearance network (A-net) that detects movable objects such as pedestrians and cars. The A-net measures the objectness independent of the background, thus providing robust information to the background contamination. The second network is the motion network (M-net) that distinguishes the foreground from the difference between an image and a background model. The M-net is trained to take the background contamination into account. Therefore, it is possible to detect the foreground region robustly even in the contaminated background model. The proposed architecture combines two networks and performs moving object detection considering both appearance and motion characteristics. As shown in Fig. 1(d), the A-net represents the approximate position of the moving object based on its appearance. The result of M-net in Fig. 1(e) describes the detailed shape of the moving object. However, due to the inaccurate background, a portion of the background is detected as a moving object. The proposed method, which combines the merits of both networks, detects the details of the moving object and shows a small false positive rate as illustrated in Fig. 1(f).

## 2. METHOD

The structure of the proposed method is depicted in Fig. 2. The A-net performs appearance-based detection using only the given image. The M-net compares the input image with the background model to detect the moving object region. The last layers of the two networks are combined to form the final detection result of the proposed method.

### 2.1. Structure of A-net

The A-net is a convolutional neural network that performs appearance based object detection. To focus on the appearance of the object, the A-net uses only one image without background information. Therefore, the A-net detects the appearance of movable objects (people, cars, animals, etc.). The A-net uses objectness information instead of the motion information emphasized in the background-centric approach. Since the A-net performs detection without background information, it is robust against background contamination.

To provide the general object information, we use the VGG-16 network [13] pre-trained in ImageNet dataset [14] as initial weights of the A-net. The structure of A-net is depicted in Fig. 2. For the use of pre-trained information, the network

before pool4 layer of VGG-16 is adopted as the structure of A-net. After the pre-trained network, a zero-initialized convolutional layer is added to perform the detection. The A-net has a fully convolutional structure and performs detection regardless of input image size. Each pooling layer reduces the layer size by a factor of 1/2, and the A-net has three pooling layers. Thus, the detection result of A-net is 1/8 size of the image.

### 2.2. Structure of M-net

In the proposed structure, the M-net is a network that detects motion. In addition to the input image, the M-net receives the background image generated from the background model [4]. The M-net detects motion region based on difference between image and background image. The M-net is similar to the background-centric approach in that it uses the difference between background and image. However, the M-net utilizes the contaminated background itself and so it is possible for the M-net to remember the contaminated situation in background modeling. As a result, the M-net becomes more robust to background contamination than the background subtraction approach.

The motion information is a low-level information compared to appearance information. Therefore, a shallow network is sufficient for motion detection of the M-net. By using a shallow network, the M-net can concentrate only on motion information, excluding high-level appearance information. The structure of M-net is shown in Fig. 2. The M-net is a shallow network composed of three convolutional layers and one pooling layer. The first two layers are randomly initialized, and the last layer is zero-initialized to perform the detection. The number of pooling layers of the M-net is one, and the detection result is half of the image size. It is 4 times higher resolution than the A-net. Thus, the M-net can detect the detailed shape of an object.

### 2.3. Merging of A-net and M-net

As shown in Fig. 2, the proposed method combines the A-net and the M-net to join motion and appearance information. However, this combination creates a problem of learning imbalance between the A-net and the M-net. Since the A-net, with prior knowledge, is more informative than the randomly initialized M-net, the A-net is dominant in training. To solve the imbalance problem of the two networks, we first train the two networks separately. Through separated training, the M-net acquires prior knowledge, and the A-net and the M-net can be learned in a balanced way. After that, the A-net and the M-net are combined, and the combined structure is trained. The method of combining two networks into one is similar to that used by [15]. First, bilinear interpolation is applied to the last layer of the A-net to magnify it by 4. As a result, the size of the last layers of both networks is the same. After that, the last layers of the A-net and the M-net are combined into one layer. The combined layer passes through the pixel-wise soft-max and becomes the final detection score of the proposed method.

## 3. EXPERIMENTS

The proposed method was validated through 15 moving object videos presented in [6, 7, 16]. Among the 15 videos, videos with complex camera movement (*Campus1*, *Campus2*), a video captured by a dashcam mounted on a vehicle (*Daimler*), a video with similar background and foreground (*Fence*), and a video with heavy motion blur (*Cycle*) were selected for the test. Because of lack of training data, the proposed network was trained using 14 videos excluding one video for the test, and it was repeated 5 times for 5 test videos. The proposed method was compared with the state-of-the-art methods of the background-centric approach [4, 6, 16].

### 3.1. Implementation details

The proposed method was implemented through the Tensorflow library. Each frame of video was used as the input image and mean values of Gaussian background model [4] were used as the background image. In the training process, one image was used for one iteration and the learning rate was set to $10^{-6}$. For the separated training, the M-net was trained for 10,000 iterations and the A-net was trained for 5,000 iterations. After that, we combined the two networks and trained the combined network for 5,000 iterations. All training processes were conducted using the pixel-wise soft-max log loss, and the VGG pre-trained layers were also fine-tuned. For the test, we compared two soft-max scores (background and foreground) of each pixels and classified the pixels into higher-scored category. The classified results were compared with the ground truth and the performance was measured. The processing time for detection in each frame was 14ms for computation of the deep network on GTX 1070 GPU and 6ms for background modeling at 3.3GHz CPU. In other words, the proposed method works in real-time at 50 fps.

### 3.2. Results

The proposed algorithm was compared with the state-of-the-art methods [4, 6, 16]. The quantitative comparison of the experiments is shown in graph of Fig. 3 and the qualitative comparison is depicted in Fig. 4. Since *Campus1* and *Campus2* have complex camera motion, background models of [4] and [6] were failed to estimate the camera motion. Therefore, as shown in Fig. 4, some objects were not detected and the recall was dropped. In *Cycle* video, the method of [16] failed to estimate camera motion and repeatedly initialized the background model. As a result, the performance was severely degraded. However, the proposed method shows consistently good performance regardless of background contamination or

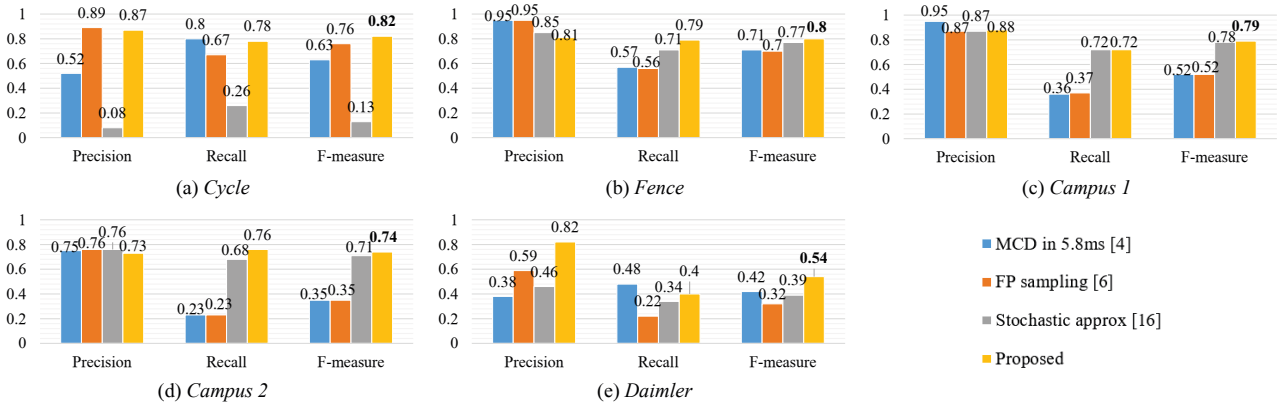


**Fig. 3**. Quantitative results. The graphs illustrate the detection performance of the state-of-the-arts and the proposed method for five moving camera videos. The proposed method shows equal or better performance than the state-of-the-art methods.


**Fig. 4**. Qualitative comparison of detection results. The figure illustrates detection results of the state-of-the-art methods and the proposed method. From the top, it shows the results of *Cycle*, *Fence*, *Campus 1*, *Campus 2*, and *Daimler*.

complex camera motion. In addition, the proposed method shows better performance than other state-of-the-art methods in *Daimler* driving scenes. This result shows that the proposed method effectively copes with the background contamination frequently observed in the dashcam video and outperforms the state-of-the-art methods for the videos captured in the freely moving cameras.

## 4. CONCLUSION

We proposed deep learning based moving object detection framework applicable to freely moving camera videos, such as dashcam videos. The proposed method consists of a network focusing on the appearance as well as a network dedicated to the motion. It is a meaningful contribution that the proposed deep learning approach achieves a robust performance against background contamination, even with the free movement of cameras. In the experiments, the effectiveness of the proposed method has been verified by comparison with the-state-of-the-art methods. Furthermore, the proposed method has a strong advantage in that it is capable of a real-time operation speed of 50 fps, which is suitable for actual application including the challenging situations in autonomous vehicles.

## 5. REFERENCES


[1] W.-D. Jang, C. Lee, and C.-S Kim, "Primary object segmentation in videos via alternate convex optimization of foreground and background distributions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1187 – 1200, 2014.

[3] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] K. M. Yi, K. Yun, S. W. Kim, H. J. Chang, H. Jeong, and J. Y. Choi, "Detection of moving objects with non-stationary cameras in 5.8ms: Bringing motion detection to your mobile device," in *CVPR Workshops*, 2013, pp. 27–34.

[5] S. W. Kim, K. Yun, K. M. Yi, S. J. Kim, and J. Y. Choi, "Detection of moving objects with a moving camera using non-panoramic background model," *Machine Vision and Applications*, vol. 24, pp. 1015–1028, 2013.

[6] K. Yun and J. Y. Choi, "Robust and fast moving object detection in a non-stationary camera via foreground probability based sampling," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 4897–4901.

[7] K. Yun, J. Lim, S. Yun, S. W. Kim, and J. Y. Choi, "Attention-inspired moving object detection in monocular dashcam videos," in *23rd International Conference on Pattern Recognition (ICPR)*, 2016.

[8] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[9] B. H. Chen and S. C. Huang, "An advanced moving object detection algorithm for automatic traffic monitoring in real-world limited bandwidth networks," *IEEE Transactions on Multimedia*, vol. 16, pp. 837–847, 2014.

[10] Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau, and Robert Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2015.

[11] B.-H. Chen and S.-C. Huang, "Probabilistic neural networks based moving vehicles extraction algorithm for intelligent traffic surveillance systems," *Information Sciences*, vol. 299, pp. 283–295, 2015.

[12] Y. Zhang, X. Li, Z. Zhang, F. Wu, and L. Zhao, "Deep learning driven blockwise moving object detection with binary scene modeling," *Neurocomputing*, vol. 168, pp. 454–463, 2015.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] F. J. López-Rubio and E. López-Rubio, "Foreground detection for moving cameras with stochastic approximation," *Pattern Recognition Letters*, vol. 68, pp. 161–168, 2015.