Moving Shadow Detection from Background Image and Deep Learning

Jong Taek Lee(✉), Kil-Taek Lim, and Yunsu Chung

Electronics and Telecommunications Research Institute (ETRI),
Daegu, South Korea
jtlee@utexas.edu, {ktl,yoonsu}@etri.re.kr

Abstract. We present a novel approach for moving shadow detection, which is applicable to various environments. Although there have been extensive studies of shadow detection since 1980s, the problem is still considered as a challenging and important issue in the most visual surveillance systems. Herein, we propose a shadow region learning method using a deep structure for moving shadow detection. Unlike previous approaches which are usually based on hand-crafted features using chromacity or physical properties of shadow regions, our approach is able to automatically learn features of shadow region from input source and its background image. The proposed approach is relatively simpler to implement than previous approaches as we don't need to consider intensity and color properties of video sequences. However, its performance is comparable to that of state-of-the-art approaches. Our algorithm is applied to five different datasets of moving shadow detection for comprehensive experiments.

Keywords: Moving shadow detection · Convolutional deep neural network · Visual surveillance

1 Introduction

Shadows are always present both in-door and out-door environments. The effect of shadows varies depending on light sources, but it has a high impact on the most surveillance systems. Shadows often make scene analysis more difficult rather than help scene understanding. Although shadow detection is a difficult problem, detected shadows can help analyzing the direction of light sources and the geometry of scenes. Also, shadow removal aids the localization, segmentation, and tracking of interesting objects in surveillance systems. A substantial number of studies for shadow detection ever exist, but the shadow detection is an important and challenging problem in the computer vision field.

Therefore, we propose a new general approach for shadow detection. By using convolutional deep neural network (ConvNets), we can learn features for shadow regions from both input images and background images. Without adjusting parameters for environments, we are able to learn our network with sufficient amount of input data.

F. Huang and A. Sugimoto (Eds.): PSIVT 2015 Workshops, LNCS 9555, pp. 299–306, 2016.
DOI: 10.1007/978-3-319-30285-0_24

The rest of the paper is organized as follows: Sect. 2 shortly introduces previous works. Section 3 presents a detailed description of our approach. Section 4 exhibits quantitative and qualitative experimental results. Section 5 concludes the paper.

2 Related Work

2.1 Moving Shadow Detection

Prati et al. [1] and Sanin et al. [2] classified and presented comparative evaluation of moving shadow detection methods developed until 2000 and 2010, respectively. Sanin et al. [2] placed shadow detection methods in a feature-based taxonomy comprised of four categories such as chromacity, physical properties [3], geometry [4] and textures [5]. As our method learns features from regions of the input frame and its background frame, it combines a spectral feature-based method and a texture-based method. Yet, none of methods introduced in [2] used convolutional networks for feature learning in a moving shadow detection framework.

2.2 Deep Learning

Many deep representation learning architectures have been proposed in the last decade and the successes of these deep learning methods in speech recognition [6] and image classification challenge [7] have motivated its use in computer vision applications. Khan et al. [8] proposed a ConvNets architecture to learn the most relevant features for shadow detection instead of using hand-crafted features. The learned features are further fed to conditional random field (CRF) model for shadow edge detection. Shen et al. [9] also proposed a structured ConvNets for shadow detection, but its output spaces are high dimensional to learn shadow edge structure. Both approaches are able to learn features of shadow boundaries to segment shadow regions from the other regions.

On the other hand, our approach learns shadow regions instead of shadow edges. Our approach can be less accurate than these two approaches for detecting shadow edges and their surrounding pixels, but our approach can efficiently localize shadow blobs without significant errors. Also, to the best of our knowledge, no studies have tried to use both input image and background image to learn features for shadow region classification.

3 Approach

3.1 System Overview

This subsection provides an overview of our shadow detection method. Given an input image, we first generate an over-segmented image. For superpixels containing positive pixels from foreground mask, their locations are measured to extract patches of an input image, a foreground mask, and a background image. Input

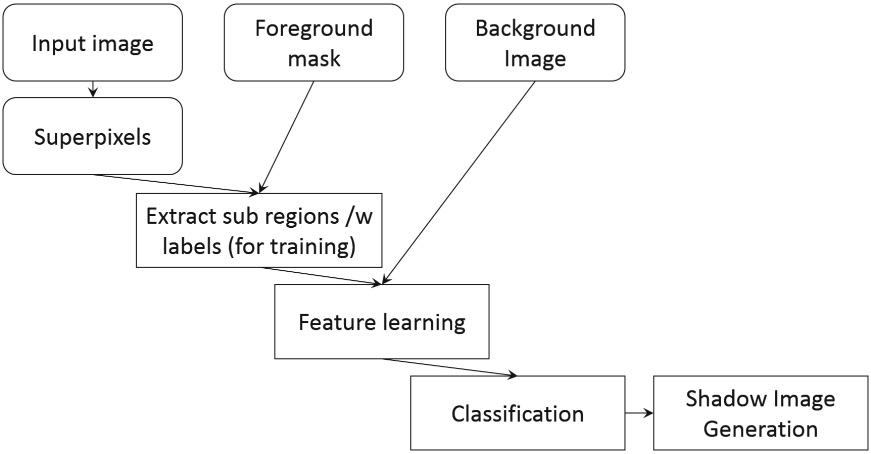


Fig. 1. The framework of our shadow detection.

image patches and background image patches are merged so that the number of channels of input source for our ConvNets is six. (three from input image and three from background image) The number of output from the ConvNets is three: foreground class, shadow class, and invalid class. From the superpixels of input image and their shadow classification results, we can generate a foreground/shadow map. This system overview is summarized in Fig. 1.

3.2 Sub-region Extraction

The size of shadow blobs and foreground blobs varies depending on the size of interesting objects and light conditions. In the case that we use a grid of uniformly spaced cells as a sub-region, the quantization error will increase if the grid is sparse, and the computational requirements will increase if the grid is dense. Therefore, we use a over-segmented image by using a superpixel algorithm. To form superpixels, we use the *Superpixels Extracted via Energy-Driven Sampling* (SEEDS) algorithm [10] because the algorithm runs in real-time and performs a comparable achievement to the state-of-the-art. The extracted superpixels are further resized to a fixed size window for feature learning, as shown in Fig. 2.

3.3 Deep Shadow Region Learning with ConvNets

Our ConvNets consists of seven layers as shown in Fig. 3. The first four layers are two convolutional layers and two subsampling layers, and the other three layers are fully-connected layers. The last fully-connected layer can generate a probability for 3 classes such as foreground, shadow, and the others. The Rectified Linear Unites (ReLUs) nonlinearity is applied to the output of two convolutional layers and two fully-connected layers to reduce the training time. The first and second convolutional layers consist of twenty and sixty 5 × 5

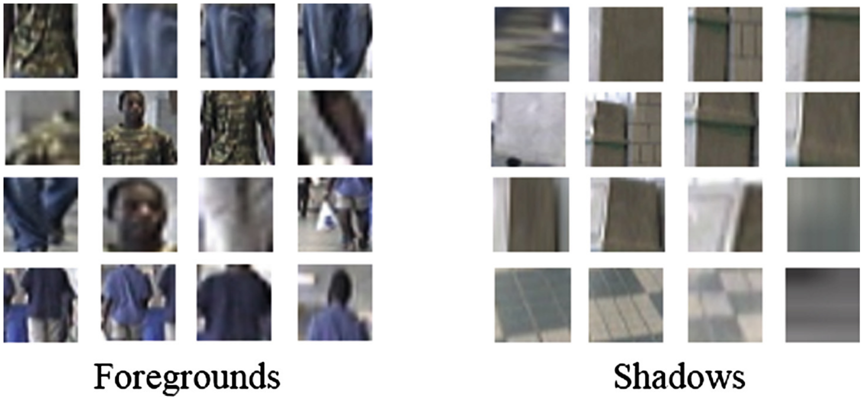


Fig. 2. A collection of resized sub-regions with label - foregrounds (not shadows) on the left side and shadows on the right side.

kernels, respectively, with unit pixel stride. The pooling size is 2×2 with unit-pixel stride. Also, the fully-connected layers have 500 units each.

The dimension of input source of our ConvNets is 32 (width) × 32 (height) × 6 (channels) by integrating an input image and background image. We trained our ConvNets without background image (i.e. 32 × 32 × 3 input source), but the network could not learn features for shadow detection. The network may be able to learn boundaries near shadows, but classifying shadow regions only depending on image patch seems impractical.

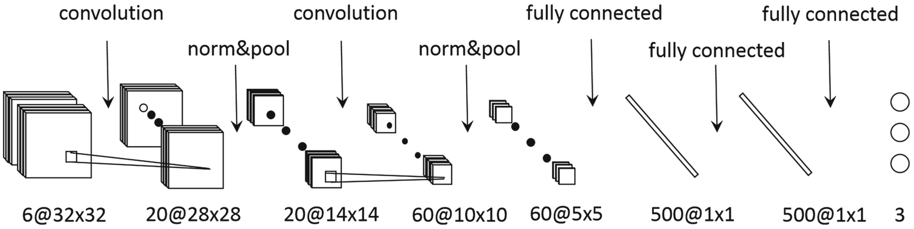


Fig. 3. An illustration of our ConvNet architecture.

4 Experiments

4.1 Datasets

To evaluate the proposed approach, we used a challenging video sequences introduced in [1,11,12]: The Campus, Hallway, Lab, Room, and CAVIAR sequences. The details of the video sequences are present in [2]. These video sequences are challenging because of variations in the scene, illumination, and type/size of objects and shadows. The Campus sequence is an outdoor sequence where some of the shadows are extremely long and the Hallway sequence is an indoor sequence which has a textured background. The Lab sequence has a dynamic scene of two people crossing each other and the Room sequence is taken by a wide-angle camera with distortion. show two laboratory rooms in two different perspectives and lighting conditions. The CAVIAR sequences has a reflective background and human clothing appearance varies significantly. Our shadow detection method does not need to change any parameters for different sequences, which implies our method is applicable to wide range of shadow detection problem. The first row of Fig. 4 shows a sample images for shadow detection. We implemented an shadow detection system using the Caffe [13] implementation of the ConvNets.

4.2 Experimental Framework

Network Configuration. The input layer corresponds to input image patch and background image patch, and the output layer corresponds to class probability measure whether the image patch is of shadows or not. The size of input image patch has an impact on the efficiency and accuracy of the shadow detection system. We tested three shadow detection image sizes (32 × 32, 20 × 20, 48 × 48) to find the optimal solution. Using a 20 × 20 sized detection window, each iteration takes less time than the others and the network converges faster. However, the accuracy rate of the network is too low for shadow detection. For a 48 × 48 window, the performance of the network is slightly better than using a 32 × 32 window after 4 times of computational load is used.

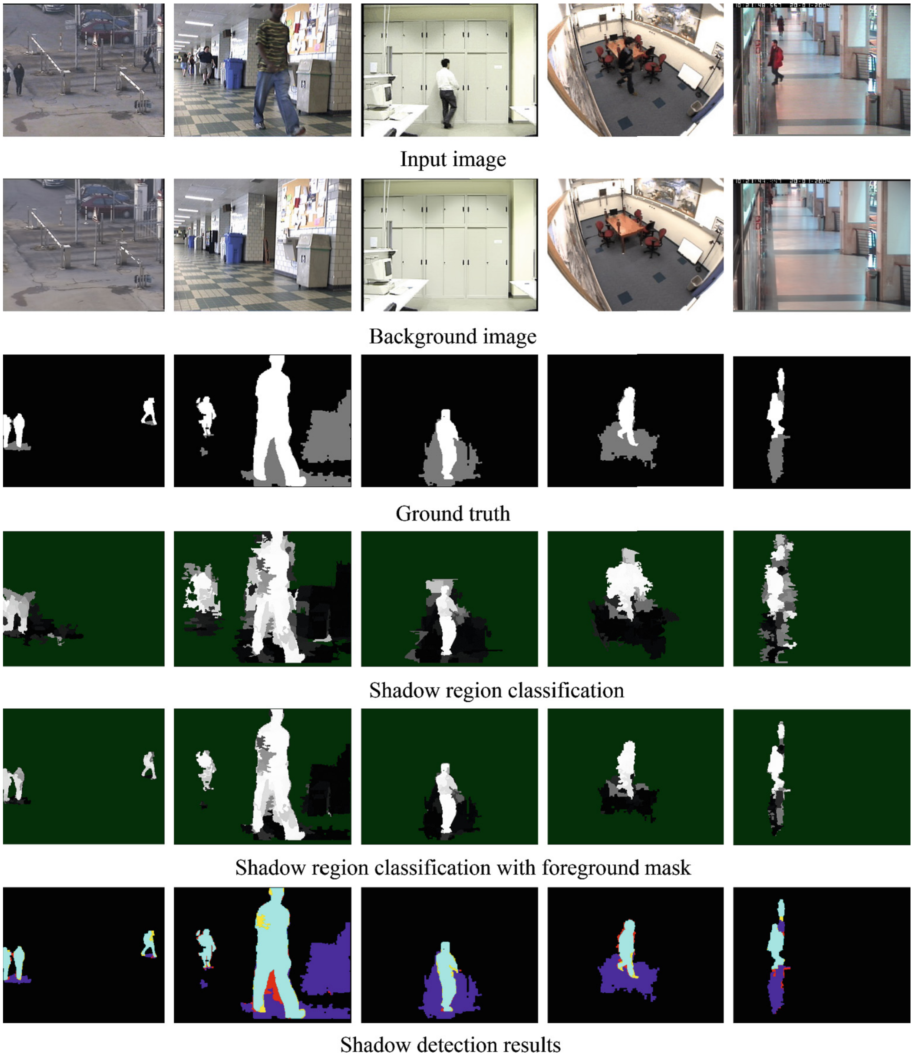
Optimization and Class Balancing. We trained our network using the standard stochastic gradient descent with momentum. Initial learning rate is 0.0001, momentum is 0.9, and weight decay is 0.0005. We used drop-out layers in the first two fully-connected layers to learn more robust features. The performance of our network was not improved by class balancing, while our labels are slightly unbalanced.

Table 1. Comparison of percentage (%) of shadow detection rate (η) and shadow discrimination rate (ξ).

Sequence	Campus		Hallway		Lab		Room		CAVIAR	
Method	η	ξ	η	ξ	η	ξ	η	ξ	η	ξ
DNM [1]	82.9	86.7	–	–	76.3	89.9	78.6	90.3	–	–
SILTP [5]	62.6	43.1	82.3	91.1	–	–	–	–	87.5	94.8
RMSDAP [14]	33.8	91.0	88.3	96.4	–	–	–	–	89.7	96.4
Proposed	74.3	84.9	89.6	92.2	86.2	84.3	80.9	93.7	84.0	93.7

4.3 Quantitative Comparisons

To compare the accuracy of our proposed method with the baseline moving shadow detection method, we tested our network on five different video sequences and present the two terms for analyzing shadow detection performance: shadow detection rate (η) and shadow discrimination rate (ξ) [1]. The two metrics are defined as follows:



Input image

Background image

Ground truth

Shadow region classification

Shadow region classification with foreground mask

Shadow detection results

Fig. 4. Qualitative results of our methods on five video sequences: Campus, Hallway, Lab, Room, and CAVIAR from left to right (Color figure online).

$$\eta = \frac{TP_S}{TP_S + FN_S}; \xi = \frac{TP_F}{TP_F + FN_F}, \tag{1}$$

where TP and FN denote true positive and false negative, respectively, and the subscript S stands for shadow and F stands for foreground. Shadow detection rate can be calculated by the ratio of the number of the correctly classified shadow pixels to that of the number of the shadow pixels. Shadow discrimination rate can be calculated by the ratio of the number of the correctly classified foreground pixels (not shadow) to that of the number of foreground pixels. The performance of our method and baseline methods are compared in Table 1. Our method performs as good as other methods.

4.4 Qualitative Results

Figure 4 shows successful examples of the proposed method detecting shadows in various environments. The images of the first three rows are given in the datasets. The fourth row shows the probability map, where brighter pixels are more likely foregrounds (not shadows) and darker pixels are more likely shadows. The pixels of dark green color indicate regions which are neither foregrounds nor shadows. The fifth row shows the probability map with foreground (including shadows) mask. The sixth row presents how many shadow/foreground pixels are correctly classified. The pixels of blue, teal, red, and yellow indicate true positive of shadows, true positive of foregrounds, false negative of shadows, and false negative of foregrounds, respectively.

5 Conclusion and Future Work

We proposed a novel shadow region learning method using a deep structure for moving shadow detection in various environments. Due to the simplicity and generality of our method, it can easily incorporate other existing methods. However, our system requires sufficient amount of labeled training data. In future work, we consider to use autoencoder neural networks [15] to increase learning rates for scenarios where small labeled training data and large unlabeled training data. To enhance our shadow detection, we could incorporate a CRF model [16], which is able to enforce label consistency and local smoothness. Also, we consider a fully convolutional network for shadow detection as it is proven effective for semantic segmentation.

References

1. Prati, A., Mikic, I., Trivedi, M.M., Cucchiara, R.: Detecting moving shadows: algorithms and evaluation. IEEE Trans. Pattern Anal. Mach. Intell. **25**(7), 918–923 (2003)
2. Sanin, A., Sanderson, C., Lovell, B.C.: Shadow detection: a survey and comparative evaluation of recent methods. Pattern Recogn. **45**(4), 1684–1695 (2012)

3. Huang, J.-B., Chen, C.-S.: Moving cast shadow detection using physics-based features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, pp. 2310–2317 (2009)
4. Lalonde, J.-F., Efros, A.A., Narasimhan, S.G.: Detecting ground shadows in outdoor consumer photographs. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 322–335. Springer, Heidelberg (2010)
5. Qin, R., Liao, S., Lei, Z., Li, S.Z.: Moving cast shadow removal based on local descriptors. In: 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, pp. 1377–1380 (2010)
6. Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, pp. 1097–1105 (2012)
8. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R.: Automatic feature learning for robust shadow detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, pp. 1939–1946 (2014)
9. Shen, L., Wee Chua, T., Leman, K.: Shadow optimization from structured deep edge detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, pp. 2067–2074 (2015)
10. Van den Bergh, M., Boix, X., Roig, G., de Capitani, B., Van Gool, L.: SEEDS: superpixels extracted via energy-driven sampling. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 13–26. Springer, Heidelberg (2012)
11. Martel-Brisson, N., Zaccarin, A.: Kernel-based learning of cast shadows from a physical model of light sources and surfaces for low-level segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, pp. 1–8 (2008)
12. CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint (2014). arxiv:1408.5093
14. Dai, J., Han, D.: Region-based moving shadow detection using affinity propagation. Int. J. Signal Process. Image Process. Pattern Recogn. **8**(3), 65–74 (2015)
15. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
16. Zhu, J., Samuel, K.G.G., Masood, S.Z., Tappen, M.F.: Learning to recognize shadows in monochromatic natural images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, pp. 223–230 (2010)