Background subtraction using a convolutional neural network

Fedor Morozov[1], Anton Konushin[1,2]
[1]Lomonosov Moscow State University, Moscow, Russia
[2]NRU Higher School of Economics, Moscow, Russia
f-morozov@yandex.ru, ktosh@graphics.cs.msu.ru

Abstract

Background subtraction is a first step in many video content analysis applications including object tracking and event recognition.

Algorithms that take into account the neighborhood of each analyzed pixel tend to outperform pixel-based approaches. Most of these spatial-based algorithms use hand-crafted local features combining color and texture information to compare the patches.

We propose to use a convolutional neural network (CNN) to predict similarity between patches in the current image and in the background model. Pairs of patches extracted from existing background subtraction dataset with pixel-accurate labeling are used to train the neural network.

The use of a siamese network architecture with a dot product layer allows to compute the convolutions for each patch only once resulting in a local feature descriptor optimized for background subtraction. Therefore, the proposed similarity metric can be incorporated in existing adaptive background models without significant computational overhead.

We demonstrate the robustness of the proposed approach in terms of precision and recall on a subset of ChangeDetection.net dataset and show that our method outperforms existing color- and texture-based descriptors used for background subtraction.

Keywords: *Background subtraction, convolutional neural networks, local feature descriptors, similarity learning.*

1. INTRODUCTION

Background subtraction is an important task in the area of computer vision as it often serves as the first step in many video surveillance and video content analysis algorithms including object tracking, people counting, event recognition and traffic analysis.

The problem can be formulated as follows: for each pixel of the video frame background subtraction algorithm should determine whether the pixel is part of a background or a foreground object. In video surveillance scenarios the algorithm has to deal with shadows, lighting changes, dynamic background and other challenges while maintaining real-time processing speed.

Most background subtraction algorithms work by building and updating a background model of a scene. Different metrics can be used to compare frames to the model, including color distances and local descriptors.

Convolutional neural networks have recently proven their superiority in many computer vision tasks like object classification and facial recognition. Authors of [13] used CNN to compare image patches for dense stereo matching, demonstrating robustness of this technique and its compatibility with existing stereo matching algorithms.

In this work we propose to train a convolutional neural network (CNN) to predict similarity between patches in the current image and in the background model. We use the outputs of the network in a simple background subtraction algorithm and compare it to existing methods in terms of precision and recall.

2. RELATED WORK

The performance of a background subtraction algorithm depends on its background modeling method and the way it compares individual pixels or patches to the model.

Most background models are either variations of mixture of Gaussians [7] or sample consensus model [1]. Alternative non-parametric models use codebook based approach proposed in [6]. Some state-of-the art algorithms employ more sophisticated background models like split Gaussian models [11] or sharable models [3] and rely on color comparison of individual pixels. Other algorithms use information about the neighborhood of each analyzed pixel in the form of a local descriptor.

Authors of [5] proposed to use local binary patterns (LBP) for background subtraction robust to illumination changes. [15] describes a spatiotemporal version of the LBP descriptor (STLBP) that extracts both spatial texture and temporal motion information of a pixel.

Local binary similarity patterns descriptor [2] is a spatiotemporal modification of LBP that thresholds pixel differences instead of directly comparing values. In [8] it was used to demonstrate that simply incorporating a local feature component in a regular pixel-based method can lead to a significant performance increase. Two state-of-the-art algorithms use combinations of LBSP with sample consensus [9] and codebook [10] models.

Algorithms described above use hand-crafted local features to compare image patches. Patch comparison is a common task in many computer vision applications including local feature matching, dense stereo and optical flow. Some approaches use machine learning, specifically convolutional neural networks to predict similarity between patches.

In [13] and [14] a convolutional neural network is trained for stereo matching. Authors utilize a siamese network architecture to compute patch descriptors and design two versions of the network: an accurate architecture that uses fully-connected layers to compare the descriptors and a fast architecture that uses an inner product layer. Authors of [12] make a review of different network architectures applied to local feature matching and wide baseline stereo.

In this work we use a network architecture similar to the fast architecture from [14] and apply it to background subtraction.

3. PROPOSED METHOD

3.1 Comparing patches

To compare image patches we use a convolutional neural network shown in **Figure 1**. The network has siamese architecture, i.e. it consists of two joined sub-networks with shared weights.

Each subnetwork consists of a few convolutional layers with 3 × 3 kernels. All convolutional layers except the last one are followed by rectified linear unit (ReLU) activation layers. After the convolutional layers we obtain a single vector of numbers that is later normalized. Input patch size depends on the number of convolutional layers. In our experiments we used sub-networks with 3 convolutional layers processing 7 × 7 patches.

The siamese architecture allows to compute the output of a sub-network only once for each patch, so it's output can be considered to be a descriptor of the patch. To predict similarity score between two patches we use the dot product of their descriptors.

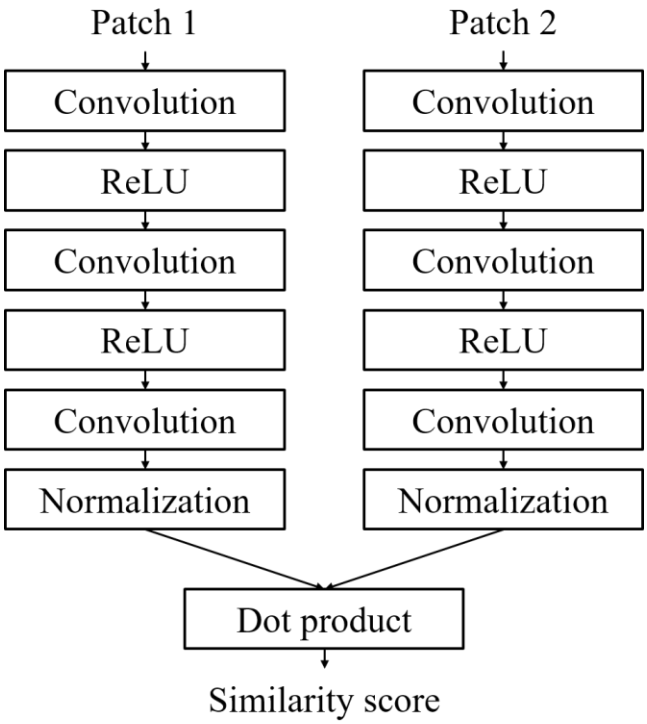

Figure 1: Proposed network architecture.

3.2 Training the network

To train the network a pixel-accurate groundtruth dataset is required. Since the number of foreground pixels is relatively small we utilize all of those from the training sequences obtaining about 20 million examples from the *baseline* category of *ChangeDetection.net* dataset [4].

We extract a patch centered at the foreground pixel and two background patches at the same location on different frames. The pair of background patches forms a positive example for similarity training, while the pair of background and foreground forms a negative example. Following [14] the extracted pairs of examples are used to minimize hinge loss with margin value of 0.2 using mini-batch gradient descent with batch size of 128.

3.3 Background modeling

In this work we focus on the CNN similarity measure and its comparison to existing descriptors used for background subtraction. We use a simple background modeling scheme described in [2]. The background model is constructed over the first 200 frames by choosing a descriptor repeated at least 20 times at each pixel location, there is no background update mechanism. Median filter is applied as a post-processing step.

Using a simple background model highlights performance of the proposed similarity measure and makes it possible to directly compare different descriptors.

Since sub-network output of each patch does not depend on other patches and can be computed once, the proposed similarity measure can be used with more complex non-parametric background models like ViBE [1] and word consensus [10] without great overhead.

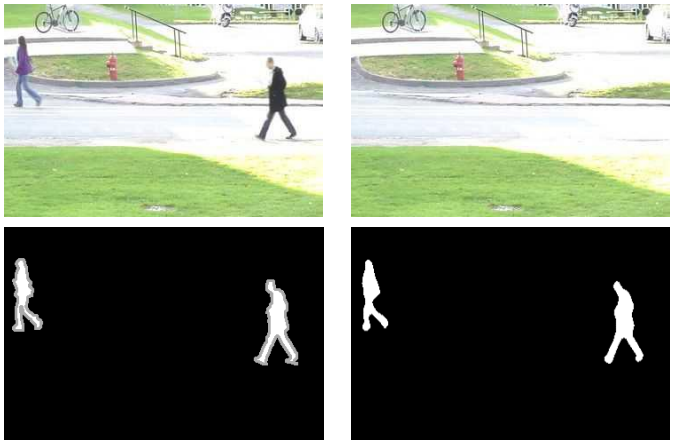

Figure 2: Background subtraction results. Top left – input frame, top right – background image, bottom left – groundtruth mask (gray denotes unknown areas), bottom right – algorithm result.

4. EVALUATION

For convolutional neural network training and overall algorithm evaluation we use videos from *ChangeDetection.net* dataset [4]. The dataset provides a set of camera-captured videos with manually labeled pixel-accurate annotations. We only use videos from *baseline* since performance in most other categories heavily depends on background modeling scheme. Pixel classification quality is evaluated in terms of precision, recall and f-measure.

Evaluation results are shown on Figure 3. CNN denotes the proposed method; other results were obtained from respective papers or from *ChangeDetection.net*. Proposed descriptor tends to outperform the existing ones: spatiotemporal local binary patterns with Gaussian mixture model [15] and local binary patterns with similar background modeling scheme [2].

While the proposed method performs better than pixel-based ViBE algorithm [1] it has lower precision and recall than advanced algorithms that use LBSP descriptor: a modification of ViBE described in [8] and one of the top *ChangeDetection.net* algorithms SubSENSE [9].

Method	Precision	Recall	F-measure
SuBSENSE	0.95	0.95	0.95
LBSP (ViBE)	0.96	0.9	0.92
CNN	0.94	0.85	0.89
ViBE	0.93	0.82	0.87
LBSP	0.93	0.8	0.86
STLBP (GMM)	0.92	0.8	0.84

Figure 3: Evaluation on *baseline* category from *ChangeDetection.net* dataset.

5. CONCLUSION

In this paper, we proposed a local descriptor for background subtraction based on a convolutional neural network and presented a simple change detection algorithm that utilizes the proposed descriptor. We demonstrated the robustness of our method on *ChangeDetection.net* dataset in comparison to other background subtraction algorithms. In the future we plan to incorporate the proposed descriptor into a more complicated background modeling scheme and to design a spatiotemporal descriptor based on the proposed one.

6. ACKNOWLEDGMENTS

This work was supported by RFBR grant no. 14-01-00849 and by the Skolkovo Institute of Science and Technology, the contract 081-R, Annex A2.

7. REFERENCES

[1] Barnich, Olivier, and Marc Van Droogenbroeck. "ViBe: a powerful random technique to estimate the background in video sequences." *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009.

[2] Bilodeau, Guillaume-Alexandre, Jean-Philippe Jodoin, and Nicolas Saunier. "Change detection in feature space using local binary similarity patterns." *Computer and Robot Vision (CRV), 2013 International Conference on*. IEEE, 2013.

[3] Chen, Yingying, Jinqiao Wang, and Hanqing Lu. "Learning sharable models for robust background subtraction." *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2015.

[4] Goyette, Nil, et al. "Changedetection. net: A new change detection benchmark dataset." *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012.

[5] Heikkilä, Marko, Matti Pietikäinen, and Janne Heikkilä. "A texture-based method for detecting moving objects." *BMVC*. 2004.

[6] Kim, Kyungnam, et al. "Real-time foreground–background segmentation using codebook model." *Real-time imaging* 11.3 (2005): 172-185.

[7] Stauffer, Chris, and W. Eric L. Grimson. "Adaptive background mixture models for real-time tracking." *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*. Vol. 2. IEEE, 1999.

[8] St-Charles, Pierre-Luc, and Guillaume-Alexandre Bilodeau. "Improving background subtraction using local binary similarity patterns." *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014.

[9] St-Charles, Pierre-Luc, Guillaume-Alexandre Bilodeau, and Robert Bergevin. "Flexible background subtraction with self-balanced local sensitivity." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014.

[10] St-Charles, Pierre-Luc, Guillaume-Alexandre Bilodeau, and Robert Bergevin. "A self-adjusting approach to change detection based on background word consensus." *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015.

[11] Wang, Rui, et al. "Static and moving object detection using flux tensor with split gaussian models." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014

[12] Zagoruyko, Sergey, and Nikos Komodakis. "Learning to compare image patches via convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[13] Zbontar, Jure, and Yann LeCun. "Computing the stereo matching cost with a convolutional neural network." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[14] Zbontar, Jure, and Yann LeCun. "Stereo matching by training a convolutional neural network to compare image patches." *Journal of Machine Learning Research* 17 (2016): 1-32.

[15] Zhang, Shengping, Hongxun Yao, and Shaohui Liu. "Dynamic background modeling and subtraction using spatio-temporal local binary patterns." *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008.

About the authors

Fedor Morozov is a Master student at Lomonosov Moscow State University, Department of Computational Mathematics and Cybernetics. His contact email is f-morozov@yandex.ru.

Anton Konushin is an associate professor at Lomonosov Moscow State University and at National Research University Higher School of Economics. His contact email is ktosh@graphics.cs.msu.ru.