Motion Background Modeling based on Context-encoder

Zhenshen Qu, Shimeng Yu, Mengyu Fu
Dept. of Control Science and Engineering,
Harbin Institute of Technology,
Harbin, China
yushimeng@hotmail.com

***Abstract*—A background modeling method for motion-based background of a video made by moving camera is proposed in this paper. We utilize the recently proposed context-encoder to model the motion-based background from a dynamic foreground. This method aims to restore the overall scene of a video by removing the moving foreground objects and learning the feature of its context. An advantage of this method is that the performance of background modeling will not be affected when the camera is moving fast.**

Keywords—motion-based background modeling; context-encoder; convolutional neural networks

I. INTRODUCTION

Background modeling is an important component of many computer vision systems and widely used before tasks such as foreground detection [1], object segmentation [2], tracking [3] and video surveillance [4].

Numerous research and studies have been done and a huge amount of methods have been developed in this area over recent years. These methods can be classified into following categories [5]: Basic Background Modeling, Statistical Background Modeling, Fuzzy Background Modeling, Background Clustering, Neural Network Background Modeling, Wavelet Background Modeling and Background Estimation. More classifications can be found in [6].

Conventional background modeling methods require a fixed camera position to keep a stationary background, and a great deal of work has been done with a stationary camera about moving objects [7]. However, in some certain conditions, the camera's position changes and a modeling of non-stationary background is needed. The Mixture of Gaussians (MOG) background model shows its high efficiency in multi-modal distribution background modeling and has been widely used. The MOG can adjust to the condition when some little changes happen to the background (for example, the waving leaves and gradual light change). But the MOG background modeling cannot work well when the scene changes a lot. [8] presented an approach of background modeling which is able to immune to the variations of the background, but it does not work when the movement of camera is fast and the background changes a lot. [9] introduced a Spatial Distribution of Gaussians (SDG) model which can detect foreground objects with non-stationary background. Some similar and earlier studies can be found in [10]. [11] proposed a real-time optical flow algorithm to detect moving objects in a dynamic scene. [12] is a further study of [11]. Another background modeling method dealing with dynamic scenes, which computes and utilizes optical flow in a higher dimensional space towards the modeling of dynamic characteristics has been proposed in [13]. The motion-based background modeling method presented in [14] used optical flow to detect moving objects. But motion field computation based on optical flow can be time consuming.

In this paper, a new idea is proposed to estimate the motion-based background while the camera is moving. We use convolutional neural networks (CNNs) to achieve this goal. We apply an unsupervised visual feature learning algorithm presented in [15] to the process of our motion-based background estimation. [15] introduced the context-encoder which is used to predict the missing part of an image according to the surroundings of the missing region in order to make a prediction that approximate to the original scene as much as possible. We utilize the context-encoder in the process of restoring a complete background of a video made by a moving camera which has dynamic obstacles. An obvious advantage of this method on background extraction is that the performance will not be affected even when the camera is moving fast.

The rest of this paper is organized as follows: Some related work is introduced in Section 2. Details of the proposed motion-based background modeling method are described in Section 3. We also discuss the problems we met in the process of experiments and propose solutions in that section. Then, in Section 4, the experimental results are presented. Finally, the conclusion is given in the Section 5.

II. RELATED WORK

CNNs have worked well in many semantic image understanding tasks including unsupervised understanding and natural images generating [15]. Autoencoders [16, 18] which can learn features of an image are typical deep unsupervised learning method in this field. Denoising autoencoders [17] can "make the learned representations robust to partial corruption of the input pattern". The context-encoder [15] could be thought of as a variant of

*Research supported by Chinese National Natural Science Foundation(61375046, Scene flow computation based on dynamic primitive feature)

denoising autoencoders. Given an image with a missing region, the context-encoders "can both understand the content of an image, as well as produce a plausible hypothesis for the missing parts".

The context-encoder in [15] is a convolutional neural network trained to predict the pixels of a missing region of a scene based on the surroundings of the missing region. The overall architecture of the context-encoder is similar to autoencoders [16, 18].

According to the theory of context-encoder, it aims to restore the missing part of an image, while we utilize the principle of context-encoder to predict our motion-based background. In contrast to [15], in which the missing regions chosen to be removed should not include a complete object otherwise it will be impossible to predict the content, the obstacles must be involved in the "dropped out" regions in our prediction process.

Another method used to inpaint an image is PatchMatch algorithm [25]. The Content-Aware Fill function in Photoshop is using this algorithm. It is an interactive image editing tool using a randomized algorithm and can find approximate nearest neighbor matches between image patches (as described in [25]).When inpainting relatively big and semantic regions, context-encoder outperforms Photoshop with our text samples. We made a comparison between using Content-Aware Fill function and context-encoder. The results are discussed in the Section 5.

III. Motion Background Modeling based on Context-encoder

This motion-based background modeling method applies context-encoder [15] in the background generation process. Context-encoder is a convolutional neural network which can be used to predict the missing pixel values of a natural image with a missing region and to fill the hole by generating pixels based on surrounding context. This characteristic of context-encoder provides us a thought of background modeling method. Following this line of thinking, we apply the context-encoder to the modeling of motion-based background. After removing the foreground dynamic obstacles, we use context-encoder to predict the missing pixels of the "dropped out" region according to its context and to generate a background model of each frame. The flow chart of the motion-based background modeling method is shown in Fig. 1.

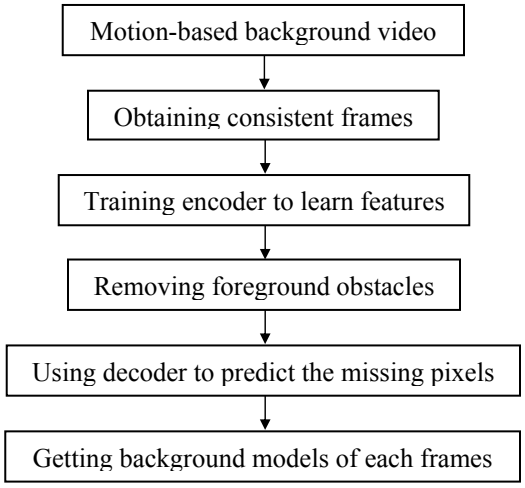

Fig. 1. Flow Chart of Motion-based Background Modeling

A. Background Modeling

Substantial studies of the background modeling methods based on stationary camera have been done in recent years and the approach that has been most widely used is the Mixture of Gaussians (MOG). In the algorithm of MOG, pixels are characterized by the intensity in the RGB color space. The possibility of the current pixel value of K Gaussians is shown in the function below [21].

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \tag{1}$$

As described in [5], $\{X_1, \ldots\ldots, X_t\}$ are pixel values, K is the number of distributions, $\omega_{i,t}$ is a weight associated to the i^{th} Gaussian at time t with mean $\mu_{i,t}$ and standard deviation $\Sigma_{i,t}, \eta$ is a Gaussian probability density function:

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{t/2}} e^{-\frac{1}{2}(X_t-\mu)\Sigma^{-1}(X_t-\mu)} \tag{2}$$

The compositions of RGB color are independent from each other and the variances are same for some computational reasons, as described in [21]. So, the covariance matrix is:

$$\Sigma_{i,t} = \sigma_{i,t}^2 I \tag{3}$$

According to [21], each pixel is characterized by a mixture of K Gaussians. The different parameters of the mixture of Gaussians should be initialized at the time the model of the background is defined. $K, \omega_{i,t}, \mu_{i,t}$ and $\Sigma_{i,t}$ are parameters of the MOG model. If the foreground objects are dense, the performance of MOG method is not very good. In such a case, we can use the output background models as training samples of the context-encoder and get a better result.

Some methods use dense optical flow algorithm to model the motion-based background. Consider $I_1(x, y)$ and $I_2(x, y)$ are two consecutive frames and assume that corresponding pixels have equal grey values, the determination of the optical flow from I_1 to I_2 comes down to find formulations (as described in [23]) below:

$$\bar{h}(x, y) = (u(x, y), v(x, y)) \tag{4}$$

$$I_2(x, y) = I_1(x - u(x, y), y - v(x, y)), \qquad \forall (x, y) \in \mathbb{R}^2 \tag{5}$$

According to Nagel and Enkelmann [24], it is a minimization problem:

$$E_{NE}(\bar{h}) = \int_{\mathbb{R}^2} (I_1(x-u(x,y), y-v(x,y)) - I_2(x,y))^2 dx + C\int_{\mathbb{R}^2} trace((\nabla\bar{h})^T D(\nabla I_1)(\nabla\bar{h}))dx \quad (6)$$

Where C is a positive constant and $D(\nabla I_1)$ is a regularized projection matrix in the direction perpendicular of ∇I_1 :

$$D(\nabla I_1) = \frac{1}{|\nabla I_1|^2 + 2\lambda^2}\left\{\begin{pmatrix} \frac{\partial I_1}{\partial y} \\ -\frac{\partial I_1}{\partial x}\end{pmatrix}\begin{pmatrix} \frac{\partial I_1}{\partial y} \\ -\frac{\partial I_1}{\partial x}\end{pmatrix}^T + \lambda^2 Id\right\} \quad (7)$$

In this formulation, Id denotes the identity matrix. The advantage of this method is that it inhibits blurring of the flow across boundaries of I_1 at locations where $|\nabla I_1| \gg \lambda$.

The motion-based background modeling method in this paper utilizes the context-encoder to restore the background. The context-encoder and the details of the method will be introduced in the next two parts.

B. Context-encoder

The encoder of context-encoder is derived from the AlexNet [19] and is used to produce a latent feature representation of input image samples with "dropped out" regions. The decoder which has a series of five up-convolutional [22] layers with learned filters utilizes the feature representation produced by encoder to inpaint the missing regions of the input samples. The encoder and the decoder are connected through a channel-wise fully-connected layer. This layer allows information to be propagated within activations of each feature map and can observably decrease the number of parameters (ignoring the bias term) comparing to a fully-connected layer of AlexNet (as is described in [15]).

The context-encoder uses a joint loss function (the reconstruction loss and the adversarial loss [20]) in the process of image restoration. The reconstruction (usually L2 or L1 loss) is used to predict a blurry outline of the missing region. And the adversarial loss can help the context-encoder make the substance of prediction more realistic and more distinct (as is described in [15]).

The reconstruction loss function L_{rec} , adversarial loss function L_{adv} , and overall loss function L presented in [15] are shown respectively as below:

$$L_{rec}(x) = \left\| \hat{M} * (x - F((1-\hat{M}) * x)) \right\|_2 \quad (8)$$

$$L_{adv} = \max_D \mathrm{E}_{x\in\chi}[\log(D(x)) + \log(1 - D(F((1-\hat{M}) * x)))] \quad (9)$$

$$L = \lambda_{rec} L_{rec} + \lambda_{adv} L_{adv} \quad (10)$$

(a) Original Image (b) Input Context (c) Output Image

Fig. 2. Qualitative Results of Context-encoder

The $*$ in above functions is the element-wise product operation. $\hat{M}$ is a binary mask. In the missing region the values of $\hat{M}$ are 1s and for the input pixels are 0s. More details of the context-encoder can be found in [15]. The performance test results of context-encoder are shown in Fig. 2.

C. Background Modeling based on Context-encoder

In our method of motion-based background modeling, we utilize the context-encoder [15] to predict the missing pixels according to the context around the "dropped out" region in the process of background extraction and restoration of each frame and get the whole background of a video taken by a moving camera.

We choose about one-tenth frames as training samples. The chosen frames should not be completely consecutive from the first one to the last one. And the total contents of the samples selected should reflect the features of the whole original motion-based background as the groundtruth. This is because each feature of the motion-background should be learned by the context-encoder. It means that there is no part of the background that we have not learnt.

After the training process, the obstacles of each frame of the video are removed with square regions. Since the moving foreground objects, which are different from the background we need, should be limited completely to the inpainting regions, we must adjust the square regions, in the middle of the input images, according to the size of the foreground objects. If not so, the moving obstacles will not be removed completely and we could not get the model of the whole background.

According to [15], the figure of the input images for training can only be squares. It is a limitation using context-encoder currently for image inpainting under most conditions. This is so because the adversarial discriminator

(a) Original image (b) Context-encoder (c) Content-Aware Fill

Fig. 3. Comparison Results between Context-encoder and Content-Aware Fill method

net has to be designed to fit some particular sizes, according to Pathak, the author of [15]. However, the actual pixel aspect ratio of a video is not always strictly 1:1 and it is not a good idea to change the original aspect ratio since it may cause a distortion to some degree. If we use rectangle images as training samples directly, the context-encoder will choose a size of 128x128 square region for each image automatically and randomly. Under this condition, we could not make sure that our foreground objects are within the square regions selected. In order to avoid this situation, we should make image clipping to ensure the foreground objects are inside the square region and as in the middle of the image as possible.

IV. Experimental Results

We ran experiments on several video samples. Each sample has a motion-based background and contains thousands of video frames. The parameters suggested in [15] are proven performing well in our experiments. The background modeling results of single frames are shown in Fig. 4. We select some typical examples from our video frame samples. You could find that the backgrounds of the examples are practically continuous.

We also use the Content-Aware Fill method to restore the background of several image samples selected from the video frames and compare the results with the ones using context-encoder as shown in Fig. 3.

V. Conclusions

In this work, we have presented an idea to model the background of a motion-based background video using the context-encoder. And the proposal was proven feasible by experiments.

However, some issues still remain to be explored. For example, it would be more convenient if we modify the context-encoder to adjust the input image size designedly with the video samples instead of being limited to 1:1 aspect ratio. Another issue remains to be improved is that when we inpaint two or more removed parts coincidently, the performance is not as good as inpainting one "dropped-out" region only. The remaining problems could be explored with further studies. The efficiency of feature learning and the quality of training results could also be improved.

Acknowledgment

We would like to thank Deepak Pathak for providing details about the context-encoder in his studies. And we thank anonymous reviewers for suggesting improvements to the writing.

References

1. H. Hassanpour, M. Sedighi, and A. R. Manashty. "Video Frame's Background Modeling: Reviewing the Techniques." Journal of Signal & Information Processing 02.2(2011):72-78.
2. D. Culibrk, O. Marques, D. Socek, H. Kalva and B. Furht. Neural network approach to background modeling for video object segmentation. IEEE Transactions on Neural Networks, 18(6), 1614-27. "Neural Network Approach to Background Modeling for Video Object Segmentation." IEEE Transactions on Neural Networks 18.6(2007):1614-27.
3. C. Stauffer and W.E.L. Grimson. "Adaptive Background Mixture Models for Real-Time Tracking." cvpr IEEE Computer Society, 2007:2246.
4. K. Toyama, J. Krumm, B. Brumitt and B. Meyers. "Wallflower: Principles and Practice of Background Maintenance." IEEE International Conference on Computer Vision IEEE, 1999:255-261 vol.1.
5. T. Bouwmans. "Recent Advanced Statistical Background Modeling for Foreground Detection - A Systematic Survey." Recent Patents on Computer Science 4.3(2011):147-176.
6. H. Wang and D. Suter. "A Novel Robust Statistical Method for Background Initialization and Visual Surveillance." Computer Vision – ACCV 2006. Springer Berlin Heidelberg, 2006:328-337.
7. A. Elgammal, D. Harwood and L. Davis. "Nonparametric model for background subtraction." European Conference on Computer Vision Springer-Verlag, 2000:751-767.
8. S. S. Huang, L. C. Fu and P. Y. Hsiao. "Region-level motion-based background modeling and subtraction using MRFs." IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 16.5(2007):1446-56.
9. Y. Ren, C. S. Chua and Y. K. Ho. "Motion detection with nonstationary background." International Conference on Image Analysis and Processing, 2001. Proceedings 2001:332-343.
10. D. Murray and A. Basu. "Motion tracking with an active camera."IEEE Transactions on Pattern Analysis & Machine Intelligence 16.5(1994):449-459.
11. A. Talukder and L. Matthies. "Real-time detection of moving objects from moving vehicles using dense stereo and optical flow."Ieee/rsj International Conference on Intelligent Robots and Systems IEEE, 2004:3718-3725 vol.4.
12. A. Talukder, S. Goldberg, L. Matthies and A. Ansar. "Real-time detection of moving objects in a dynamic scene from moving robotic vehicles." Ieee/rsj International Conference on Intelligent Robots and Systems 2003:1308-1313 vol.2.
13. A. Mittal and N. Paragios. "Motion-Based Background Subtraction Using Adaptive Kernel Density Estimation." IEEE Computer Society Conference on Computer Vision & Pattern Recognition 2004:II-302-II-309 Vol.2.
14. M. Y. Shih, Y. J. Chang, B. C. Fu, and C. C. Huang. "Motion-based Background Modeling for Moving Object Detection on Moving

Platforms.” International Conference on Computer Communications & Networks IEEE, 2007:1178-1182.

15. D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell and A. A. Efros. "Context Encoders: Feature Learning by Inpainting." (2016).
16. Y. Bengio. "Learning Deep Architectures for AI." Foundations & Trends® in Machine Learning 2.1(2009):1-127.
17. P. Vincent, H. Larochelle, Y. Bengio and P. A. Manzagol. "Extracting and composing robust features with denoising autoencoders." *International Conference, Helsinki, Finland, June* ACM, 2008:1096-1103.
18. G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks. " Science 313.5786(2006):504-507.
19. A. Krizhevsky, I. Sutskever and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25.2(2012):2012.
20. I. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. "Generative Adversarial Nets." *Advances in Neural Information Processing Systems* (2014):2672-2680.
21. C. Stauffer and W. E. L. Grimson. "Adaptive Background Mixture Models for Real-Time Tracking." cvpr IEEE Computer Society, 2007:2246.
22. A. Dosovitskiy, J. T. Springenberg and T. Brox. "Learning to generate chairs with convolutional neural networks." Computer Vision & Pattern Recognition IEEE, 2014:1538-1546.
23. L. Alvarez, J. Weickert and J. Sánchez. "Reliable Estimation of Dense Optical Flow Fields with Large Displacements." International Journal of Computer Vision 39.1(2000):41-56.
24. H. H. Nagel and W. Enkelmann. "An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences." IEEE Transactions on Pattern Analysis & Machine Intelligence 8.5(1986):565-93.
25. C. Barnes, E. Shechtman, A. Finkelstein and D. B. Goldman. "PatchMatch: a randomized correspondence algorithm for structural image editing." Acm Transactions on Graphics 28.3, article 24(2009):341-352.

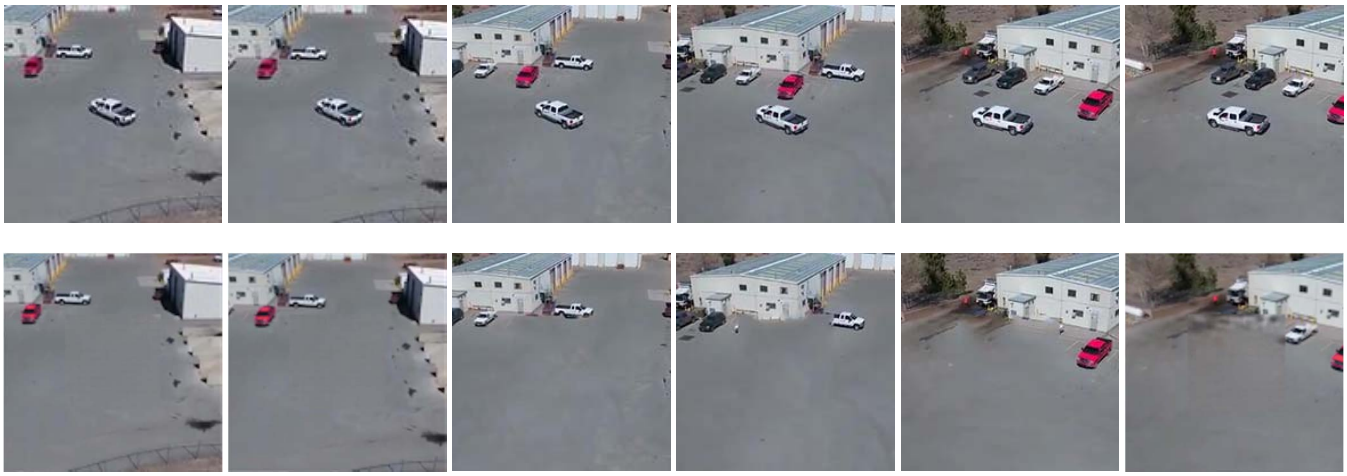

Fig. 4. Motion-based Background Modeling Results: samples in the first row are original images, the white car in the middle of the images is the dynamic foreground object and the images in the second row are background models.