Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 2706–2715
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Efficient hierarchical method for background subtraction

Yu-Ting Chen[a,b], Chu-Song Chen[a,c,*], Chun-Rong Huang[a], Yi-Ping Hung[a,b,c]

[a]*Institute of Information Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan*
[b]*Department of Computer Science and Information Engineering, National Taiwan University, 1 Roosevelt Road, Section 4, Taipei 106, Taiwan*
[c]*Graduate Institute of Networking and Multimedia, National Taiwan University, 1 Roosevelt Road, Section 4, Taipei 106, Taiwan*

Received 5 June 2006; received in revised form 14 November 2006; accepted 21 November 2006

Abstract

Detecting moving objects by using an adaptive background model is a critical component for many vision-based applications. Most background models were maintained in pixel-based forms, while some approaches began to study block-based representations which are more robust to non-stationary backgrounds. In this paper, we propose a method that combines pixel-based and block-based approaches into a single framework. We show that efficient hierarchical backgrounds can be built by considering that these two approaches are complementary to each other. In addition, a novel descriptor is proposed for block-based background modeling in the coarse level of the hierarchy. Quantitative evaluations show that the proposed hierarchical method can provide better results than existing single-level approaches.

Keywords: Hierarchical background modeling; Background subtraction; Contrast histogram; Non-stationary background; Object detection; Video surveillance

1. Introduction

Background modeling is an important module for many vision-based applications, such as visual surveillance and human gesture analysis. To detect moving objects, each incoming frame is compared with the background model learned from previous frames to classify the scene into foreground and background. A difficulty encountered in background modeling is that backgrounds are usually non-stationary in practice. Non-stationary backgrounds would be caused by waving leaves, fluttering flags, ripple water, fluorescent light, monitor flicker, and so on. Even when the background is static, camera jittering and signal noise may still cause non-stationary problems. Furthermore, shadows [1–3] and sudden lighting changes [4] are also important issues. In Ref. [5], Toyama et al. summarized 10 important problems in background subtraction. Except for these problems, real-time performance is also an important problem.

Most background modeling methods are pixel-based. Gaussian distribution has become a popular choice for modeling. Since the background is often non-stationary, a single Gaussian model used in Refs. [2,6] is not sufficient for its representation. In Ref. [7], Stauffer and Grimson proposed *Mixture of Gaussians* (MoG) by using k Gaussians to model each pixel. In MoG, an online K-means approximation was used instead of using the exact Expectation–Maximization (EM) algorithm. The MoG approach is modified or extended in several researches. For example, instead of using RGB color as features, Harville et al. [8] used YUV color plus depth measured by a stereo camera as features instead of using RGB color. Harville [9] introduced a framework to guide pixel-level evolution in Ref. [8] with high-level information. In Ref. [10], Lee proposed an effective learning algorithm for MoG.

Instead of Gaussian mixtures, Ridder et al. [11] used Kalman filter for adaptive background estimation. In Ref. [12], Zhong and Sclaroff developed a foreground–background segmentation algorithm via a robust Kalman filter to segment the foreground objects from time-varying and textured backgrounds. Stenger et al. [13] presented a framework for Hidden-Markov-Model topology and parameter estimation in an online and dynamic fashion. In Ref. [5], Toyama et al. proposed *Wallflower* to use

* Corresponding author. Institute of Information Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan.
Tel.: +886 2 27883799x1310 ; fax: +886 2 27824814.
E-mail address: song@iis.sinica.edu.tw (C.-S. Chen).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.11.023

pixel-level, region-level, and frame-level components to automatically identify people, objects, or events of interest in different kinds of environments. Elgammal et al. [1] proposed a non-parametric background subtraction method utilizing Parzen-window density estimation for representing the background. In Ref. [14], Kim et al. presented a real-time algorithm called Codebook that is efficient in either memory or speed.

Pixel-based methods model each pixel independently. Approaches of this type have the advantage of extracting detailed shapes of moving objects, but may suffer from the drawback that their segmentation results are sensitive to non-stationary scenes or backgrounds. Though some approaches such as Refs. [1,4] considered relative relationships between neighboring pixels in either space or time domain, a local vibration of the scene may still cause problems of false detection. In fact, high false-alarm rates have become a serious problem for many practical visual-surveillance systems.

Recently, some researches used block-based approaches instead of pixel-based approaches for background modeling and subtraction. In block-based approach, an image is often divided into overlapped or non-overlapped blocks, and specific block features are used for background modeling. Since a block can monitor more global changes in the scene than a single pixel, block-based approaches are insensitive to local movements and are more capable of dealing with non-stationary backgrounds. In addition, by using efficient features for each block of images, block-based approach is possible to be implemented very fast. Nevertheless, a primary limitation of block-based approach is that only a coarse-resolution foreground can be extracted, and so it is not suitable for applications requiring detailed shape information.

In recent years, researches for block-based background modeling are proposed. In Ref. [15], Matsuyama et al. proposed normalized vector distance (NVD) to measure the correlations between blocks. In Ref. [16], Mason et al. calculated edge and color histograms in each block as features to describe the block, and histogram similarity is computed to detect the foreground region. In Ref. [17], Monnet et al. used incremental PCA and an online auto-regressive model to predict a dynamic scene. In Ref. [18], Heikkilä et al. used local binary pattern (LBP) [19] histogram to capture background statistics of each block.

Both pixel- and block-based approaches have their pros and cons. An interesting issue is that they are complementary to each other. The simplest way to combine them might be running these two approaches independently, and then taking the intersections or unions of the detected foreground regions as the results. However, this combination is not the most efficient way. In this paper, a hierarchical method is proposed to combine a block-based approach, referred to as a *coarse-level* approach, and a pixel-based approach, referred to as a *fine-level* approach, in an asymmetric feed-forward framework. A novel discriminative descriptor called *contrast histogram* that is extended from Ref. [20] is used as a feature to describe each block, and Gaussian mixtures are used for maintaining a coarse-level model. For the fine-level model, existing pixel-based methods can be adopted with merely a slight modification, and a feed-forward framework is introduced to effectively dispatch the detected coarse-level information to the fine-level stage.

The paper is organized as follows: In Section 2, coarse-level background modeling using contrast histogram is introduced. The strategy for combining coarse- and fine-level models is presented in Section 3. Experimental results are shown in Section 4. Conclusion and future work are given in Section 5.

2. Coarse-level modeling

In coarse level, an efficient descriptor is expected to be built for an image block. In object recognition, invariant descriptors such as scale-invariant feature transformation (SIFT) [21] have shown their convincing performance for representing a region centered at a feature point. In SIFT, significant keypoints are identified and each keypoint is assigned with a descriptor composed of the orientation histograms computed from the gradient magnitudes and orientations sampled around the keypoint. Though SIFT has shown its powerfulness in several applications, it is not suitable for describing a background region in our experience. The primary difficulty is that the SIFT descriptor is suitable for representing clutter scenes. However, when some backgrounds are featureless, the gradient information used in SIFT causes unstable representations.

In Ref. [19], LBP has shown its powerful means for texture recognition. To represent a feature point, the circular neighboring pixels are labeled by thresholding the difference between neighboring pixels and the center (feature) pixel. Then the labeled results are considered as a binary number (LBP code). In Ref. [18], Heikkilä et al. used histogram of LBP codes in image blocks to capture background statistics. Nevertheless, LBP histogram is not suitable to describe non-stationary backgrounds according to our experimental results (see Section 4).

In this research, we construct the contrast histogram descriptor extended from Ref. [20] to describe each block directly based on the pixel colors in a block. The proposed contrast histogram is insensitive to center-point drifts and pixel reshuffles.

2.1. Contrast histogram of gray-level images

Given an image I, we first smooth this image by applying Gaussian kernels and obtain a Gaussian smoothed image L:

$$L(\mathbf{p}, \sigma) = G(\mathbf{p}, \sigma) * I(\mathbf{p}), \tag{1}$$

where $\mathbf{p}$ is a pixel at location (x, y), $*$ is convolution, and $G(\mathbf{p}, \sigma)$ is a Gaussian function with variance σ^2.

After dividing an image into blocks, our next step is to build a descriptor for each block $\mathbf{B}_c$. One obvious approach would be to sample the local image intensities in $\mathbf{B}_c$ as a template and perform template matching by using normalized correlation as in Ref. [15]. However, this method is sensitive to noise [5].

Our approach does not involve the gradient computation, and is stable to compute. The descriptor is constructed based on the contrast value defined below:

$$C(\mathbf{p}, \mathbf{p}_c) = L(\mathbf{p}) - L(\mathbf{p}_c), \tag{2}$$