Pattern Recognition 47 (2014) 2952–2961

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Detect foreground objects via adaptive fusing model in a hybrid feature space

Zhangjian Ji, Weiqiang Wang*

School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 14 June 2013
Received in revised form
19 January 2014
Accepted 18 March 2014
Available online 29 March 2014

Keywords:
Foreground detection
ST-SILTP
Adaptive fusion model
Lateral inhibition filter

ABSTRACT

Accurate extraction of foreground objects is crucial for tracking and recognition, and it is still a challenging problem in complex scenes with illumination variations and dynamic backgrounds. In this paper, we propose a foreground object detection approach with three aspects of contributions. First, considering the temporal persistence of texture sequences, we extend Liao's method [1] to the spatiotemporal domain and propose a modified local image descriptor called ST-SILTP. Second, we present an adaptive fusion approach of color and texture to compensate for their respective defects. Unlike those existing fusion methods, we do not need to adjust the parameters manually to adapt actual situations. Third, since a pixel of foreground or not depends on not only itself but also its neighborhood, we utilize the lateral inhibition filter model to incorporate the neighborhood information into calculating the pixel's confidence score. The comprehensive experiments on the dataset containing complex scenes demonstrate that the proposed approach is superior to the existing state-of-the-art algorithms.

1. Introduction

In video surveillance system, background subtraction is widely used to detect moving objects in various applications, and the accurate extraction of foreground objects is crucial for many high-level vision tasks, such as object tracking, action recognition and behavior understanding. In the framework of background subtraction, the performance of extracting moving objects highly depends on the reliability of background model. The challenges in background modeling mainly come from dynamic environments. The environmental changes result from gradual or sudden illumination variation or subtle periodic movements of objects in a scene, such as waving trees, spouting fountains, rippling water, camera jitter. To address these problems, some researchers have proposed diverse approaches. The efforts mainly focus on two aspects, *i.e.*, complex feature representations and sophisticated learning models.

In the early research stage, the single Gaussian model [2] was used to model the statistical distribution of intensities at each pixel location. Since it cannot accurately model the temporal variance at some background locations in dynamic scenes, the mixture of Gaussian model (MoG) is proposed as a popular solution [3]. There are a lot of variants or modified versions of MoG. Friedman and Russell [4] adopted a mixture of three Gaussians in the traffic surveillance applications. In [5], the authors extended this idea by using the flexible number of Gaussians to model each pixel and proposed a fast on-line approximate approach to update the model parameters, aiming at making the complexity of the algorithm meet the real-time requirement. Although these methods are claimed to be effective for dynamic scenes in practical surveillance systems, a group of appropriate parameters have to be carefully tuned, including the learning rate and standard deviation. For example, a high learning rate can make the slow moving foreground objects be mistaken as background, which results in a high false negative rate. To enhance the robustness of the MoG, diverse representation or modeling techniques are incorporated with it, such as Bayesian framework [6], dense depth data [7], color and gradient information [8], mean shift analysis [9] and region-based information [10,11] and pixel-wise information [12,13].

Besides the parametric probabilistic models like the MoG, the non-parametric probabilistic approaches are also very popular in pixel-wise background modeling [14]. The kernel density estimation techniques (KDE) are capable to establish the probability density function of arbitrary shape to approximate the distribution of pixel values, but more storage spaces are required. Further, by considering the correlation of neighboring pixels, a novel KDE method was developed by joint domain-range density estimation [12], and its performance has been proven superior to the earlier

* Corresponding author.
E-mail addresses: jizhangjian08@mails.ucas.ac.cn (Z. Ji), wqwang@ucas.ac.cn (W. Wang).

http://dx.doi.org/10.1016/j.patcog.2014.03.016
0031-3203/

neighbor-independent models. In this work, a foreground model is also introduced based on the temporal consistence criterion. Another kind of background modeling approach is based on the codebook, where the value of a given pixel is quantified into a group of codewords by the clustering computation [15,16]. During the detection phase, if the current pixel value is similar enough with one of the codewords, it is judged as a background pixel; otherwise, it belongs to a foreground one. To reach the real-time requirement, Guo et al. [17] proposed a hierarchical mechanism to model background based on the codebook. Further, an improved approach [18], whose codebook incorporated the spatial-temporal context of each pixel, was proposed. Recently, Moshe et al. [19] proposed a novel background modeling technique which directly modeled the statistics of 3D space-time video patches based on the codebook.

As more effective models appear, more robust feature representations are also developed to better adapt to illumination variation, especially sudden changes. These features include normalized color, spatial gradient, texture, optical flow and so on. The local binary pattern (LBP) features, used by Heikkila and Pietikainen [20], are tolerable against illumination changes but cannot handle the moving shadows very well. Recently, Liao et al. [1] have extended the LBP feature to a scale invariant local ternary pattern (SILTP) operator, which is more effective in dealing with illumination changes and shadows in scenes. In addition, Hu and Su [21] proposed a photometric invariant model in RGB space to handle the intensity change of each pixel. Kim et al. [16] proposed a similar approach, but it was directly embedded in the codebook algorithm, not as a post-processing step. They also proposed a multi-layer background framework, but it needed more storage space and computation cost. Javed et al. [8] emphasized on integrating multiple cues (color and gradient) in computing the background model. Yao and Odobez [22] proposed a layer-based method to detect foreground objects by exploiting both color and texture features in complex scenes. In [23], the authors combined the bottom-up and top-down information in constructing a background model and their approaches well avoided the adjustment of parameters.

There still exists many other methods to model background, such as the Kalman filtering [24], Weiner filter [25], PCA [26], ICA [27], and Barnich's ViBe [28,29]. In this paper, we present a novel adaptive fusion framework of color and texture features based on online learning, and our contributions lie in three aspects. First, since background has a significant coherence between frames, we extend the SILTP descriptor to the temporal domain and propose a novel spatiotemporal scale invariant ternary pattern (ST-SILTP) descriptor. Second, we propose an adaptive feature fusion method, which avoids manually adjusting the fusion parameters for diverse actual situations and improves its autonomy in practical applications. Third, since a pixel of foreground or not depends on not only itself but also its neighborhood, we utilize the lateral inhibition filter model to incorporate the neighborhood information into calculating the pixel's confidence score.

The remaining parts of the paper are organized as follows. In Section 2, we introduce an adaptive fusion framework of the color and texture information based on online learning to construct the background model and further the segmentation method of foreground pixels. In Section 3, the comprehensive experimental results are reported, including qualitative and quantitative analysis. Finally, a brief discussion is given in Section 4 to summarize this paper.

2. The proposed approach

In this section, we first introduce spatiotemporal scale invariant local ternary pattern (ST-SILTP) that is to model texture in Section 2.1. Then, we utilize both color and texture information to model backgrounds in Section 2.2, to make the model more robust for lighting changes and dynamic scenes. Simultaneously, in Section 2.3, we give the online learning framework of the whole algorithm. Next, we propose the similarity measurement model of color and texture respectively in Sections 2.4 and 2.5. Finally, we give an adaptive fusion framework of color and texture information in constructing the background model and how to segment the foreground pixels in Section 2.6.

2.1. Spatiotemporal scale invariant local ternary pattern

Local Binary Pattern (LBP) has been proved to be a powerful local texture descriptor. In order to make it more robust, Tan and Triggs [30] propose a Local Ternary Pattern (LTP) operator by introducing a small offset into LBP, however, it cannot adapt to sudden light changes. Liao et al. [1] extend LTP to Scale Invariant Local Ternary Pattern (SILTP) by multiplying a scale transform factor, which can keep its invariance against scale transform. But these texture features are only spatially invariant. When pixel values change in a part of local area, especially in outdoor scenes, spatial invariant features are no longer good enough. To solve this problem, Shimada et al. [31] propose a new spatiotemporal local binary pattern (st-LBP) with spatial invariance and temporal invariance. Considering their respective advantages, we also further use temporal invariance information to extend SILTP to spatiotemporal SILTP, namely ST-SILTP by combining the motion and appearance together. Due to its invariance, robustness to gray-scale variations and multi-resolution analysis, the proposed approach is very promising for practical application.

The proposed ST-SILTP descriptor is a string of 0 s and 1 s, as shown in Fig. 1, which encodes the intensity change of the pixels in the spatiotemporal neighborhood of a given pixel $\mathbf{p}$ belonging to the image grid. Formally, it is defined as

$$\Omega(I(\mathbf{p});\tau,N_{\mathbf{p}}) = \oplus_{\rho\in N_{\mathbf{p}}} s(I(\mathbf{p}),I(\rho);\tau) \tag{1}$$

where $I(\mathbf{p})$ denotes the gray value of the current pixel, $I(\rho),\rho=1,2,\ldots$ denote the gray values of neighboring pixels in the spatiotemporal neighborhood $N_{\mathbf{p}}$ of pixel $\mathbf{p}$. $\oplus$ denotes the computation of concatenating two binary strings, and $s(I(\mathbf{p}),I(\rho);\tau)$ is a piecewise function defined as

$$s(I(\mathbf{p}),I(\rho);\tau) = \begin{cases} 10 & \text{if } I(\rho) > (1+\tau)I(\mathbf{p}) \\ 01 & \text{if } I(\rho) < (1-\tau)I(\mathbf{p}) \\ 00 & \text{otherwise} \end{cases} \tag{2}$$

where τ is a scale factor to make the ST-SILTP descriptor more robust under different illumination conditions. Fig. 1 gives an example to show the encoding procedure of the ST-SILTP descriptor. Concretely, for a given pixel with intensity value 188 in current frame at time t, its 4-neighbors and the five corresponding neighbors in its previous frame at time $t-1$ are involved in the descriptor extraction computation. The nine intensity values are first quantized according to Eq. (2) with $\tau=0.05$ into three cases, *i.e.*, '00', '01', '10', and then the nine 2-bits are concatenated together to form a 18-bit string '010000000000000001', *i.e.*, the ST-SILTP descriptor of the given pixel.

Since the ST-SILTP descriptor is the extension of SILTP descriptor, it remains the advantages of SILTP descriptor. Besides, by introducing the temporal information, the ST-SILTP descriptor is more robust for dynamic background than the SILTP descriptor. Fig. 2 shows an example to illustrate the discriminative power of the two descriptors. Two video frames from sequence WaterSurface are shown in Fig. 2(a), where two 15×15 pixel blocks are labeled by red and blue squares at the same positions of the two frames. Apparently, for the two frames, the red blocks both contain