SEGMENTING LAYERS IN AUTOMATED VISUAL SURVEILLANCE

Lijuan Qin Yueting Zhuang Yunhe Pan Fei Wu

College of Computer Science
Zhejiang University
Hangzhou,310027,P.R.China
qinlijuan@hotmail.com, {yzhuang, panyh, wufei}@zju.edu.cn

ABSTRACT

Detecting objects of interest from a video sequence is a fundamental and critical task in automated visual surveillance. Those objects can either be moving or stationary. However, most of current approaches only focus on discriminating moving objects by background subtraction. In this work, we propose layers segmentation to detect both of moving and stationary target objects from surveillance video. We first construct a codebook with set of codewords for each pixel and then extend the Matrix Entropy statistical model to segment layers with codewords features. Our experimental results are presented in terms of success layer segmentation rate.

1. INTRODUCTION

Video surveillance systems seek to automatically identify people, objects, or events of interest in different kinds of environments. Typically, these systems consist of stationary cameras directed at offices, parking lots, and so on, together with computer systems that process the images and notify human operators or other processing elements of salient events [1].

A common element of such surveillance systems is a module that performs background subtraction (BGS), which identifies objects from the portion of a video frame that significantly differs from a background model. The subtraction leaves only moving objects as foreground. But sometimes, moving objects (like passersby in a hurry on the street) are not the objects we are interested in, while stationary objects (like the lost luggage), which have been subtracted as background, are the target objects we are looking for. It is difficult to discriminate such kinds of moving background and stationary foreground objects by background subtraction. Therefore, we propose an idea of layers segmentation to adaptively differentiate foreground pixels, which should be processed for identification or tracking, from background pixels, which should be ignored.

Related Work The background subtraction techniques can be classified into two broad categories: single-mode modeling and multi-mode modeling. The single-mode modeling [2] is limited to handle multiple backgrounds, like waving trees. Come to three representative methods of multi-mode modeling, the mixture of Gaussians (MOG)[3][4][5] has been used to model complex, non-static backgrounds, but backgrounds having fast variations are not easily modeled with just a few Gaussians accurately, and it may fail to provide sensitive detection [6]; the non-parametric technique [6] cannot be used when long-time periods are needed to sufficiently sample the background [7]; the Codebook (CB) algorithm constructs a highly compressed background model handling the main problems of background subtraction, and is efficient in memory and speed compared with other background modeling techniques [7]. But all techniques simply detect the moving objects from background no matter whether they are the objects of interest.

In this paper, we present a novel method to segment layers in surveillance video. It is designed to 1) have the capability of encoding multiple changing backgrounds and coping with local and global illumination changes; 2) segment layers to adaptively detect the objects of interest as foreground. We develop the layers segmentation based on the CB background subtraction since its advantages of satisfying the first requirement of our work. The Matrix Entropy (ME) statistical model is introduced to build the layer model with automatically learning. Our work is composed of three main parts, which are described in Section 2. Experimental results show, in Section 3, that our method is efficient and robust for the layers segmentation. Finally, conclusions and future work are presented in Section 4.

2. LAYERS SEGMENTATION

Our method is composed of three parts as shown in Figure1.

0-7803-9332-5/05/$20.00 ©2005 IEEE

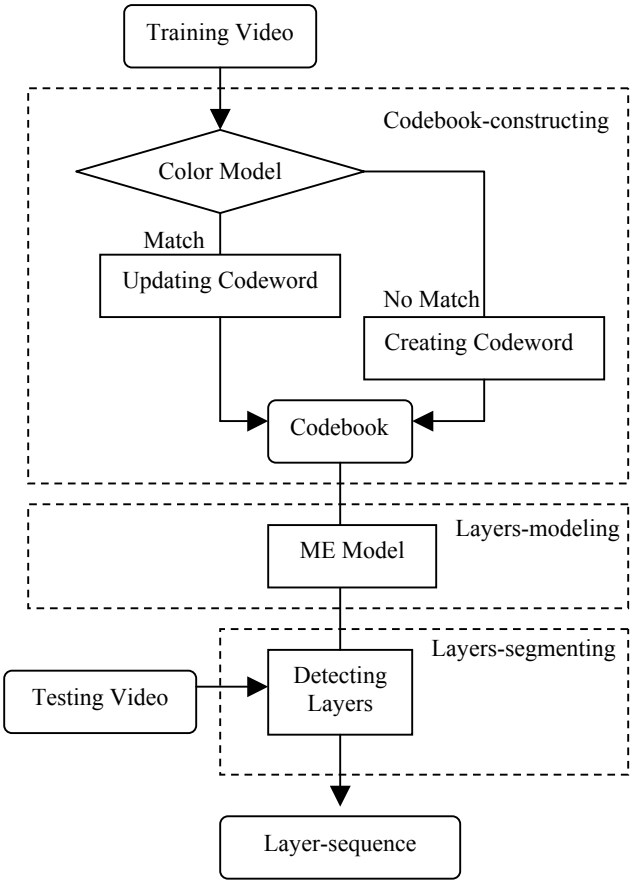


Figure 1: System diagram for layers segmentation

First, we construct a codebook with one or more codewords for each pixel. Samples at each pixel are clustered into a set of codewords based on a color distortion metric together with a brightness ratio. The video image is then encoded on a pixel-by-pixel basis. In a second step, we introduce ME model to build models adaptive to each layer. The ME model constructs an exponential log-linear function that fuses multiple features from codewords to approximate the posterior probability of a layer. In a last step, layers are segmented according to the layer-models and the foreground is detected automatically.

2.1. Codebook-constructing

The CB algorithm is encoded on a pixel-by-pixel basis. A pixel is represented by a codebook, which consists of one or multiple codewords. Let χ be a training sequence for a single pixel consisting of N RGB-vectors $\chi = \{x_1, x_2, \cdots, x_N\}$. Let $C = \{c_1, c_2, \cdots, c_L\}$ represent the codebook for the pixel consisting of L codewords. Each pixel has a different codebook size based on its sample variation. Each codeword c_i , $i = 1 \cdots L$, consists of an RGB vector $v_i = (\overline{R_i}, \overline{G_i}, \overline{B_i})$ and a 6-tuple $aux_i = \langle \min I_i, \max I_i, f_i, \lambda_i, p_i, q_i \rangle$. The tuple aux_i contains brightness values and temporal variables described here. $\min I$, $\max I$ are the min and max brightness, respectively, that the codeword accepted. f is the frequency with which the codeword has occurred. λ is the maximum negative run-length defined as the longest interval during the training period that the codeword has not recurred. p , q respect the first and last access times, respectively, that the codeword has occurred. In the training period, each value, x_t , sampled at time t is compared to the current codebook to determine which codeword c_m (if any) it matches (m is the matching codeword's index). We use the matched codeword as the sample's encoding approximation. To determine which codeword will be the best match, we employ a color distortion measure and brightness bounds. Details can be found in [8].

2.2. Layers-modeling

We propose to model the video layers by using statistical framework. The assumption is that there exist consistent statistical characteristics within pixels and with adequate learning, a general model with generic pool of computable features can be systematically optimized to construct effective segmentation tools for surveillance video.

The ME model [9] constructs an exponential log-linear function that fuses multiple features in codewords to approximate the posterior probability of each layer. The estimated model, a posterior probability, is represented as $q_\omega(b \mid x)$, where $b \in \{0,1\}$ is a random variable corresponding to the presence or absence of a layer in the context x and ω is the estimated parameter set. Here x is the layer codewords for a candidate layer pixel. From x we compute a set of binary features, $f_i(x,b) = 1_{\{g_i(x)=b\}} \in \{0,1\}$. $1_{\{\cdot\}}$ is an indication function; g_i is a predictor of layer using the i th binary feature, generated from the codewords set. f_i equals 1 if the prediction of predictor g_i equals b , and is 0 otherwise. Given a labeled training set, we construct a linear exponential function as

$$q_\omega(b \mid x) = \frac{1}{Z_\omega(x)} \exp\{\sum_i \omega_i f_i(x,b)\} \qquad (1)$$

where $\sum_i \omega_i f_i(x,b)$ is a linear combination of binary features with real-valued parameters ω_i. $Z_\omega(x)$ is a normalization factor to ensure equation is a valid conditional probability distribution. Basically, ω_i controls the weighting of *i* th feature in estimation the posterior probability. The parameters $\{\omega_i\}$ are estimated by minimizing the Kullback-Leibler divergence measure computed from the training set that has empirical distribution $\tilde{p}$. The optimally estimated parameters are $\omega^* = \arg\max_\omega D(\tilde{p} \| q_\omega)$, where *D* is the Kullback-Leibler divergence defined as

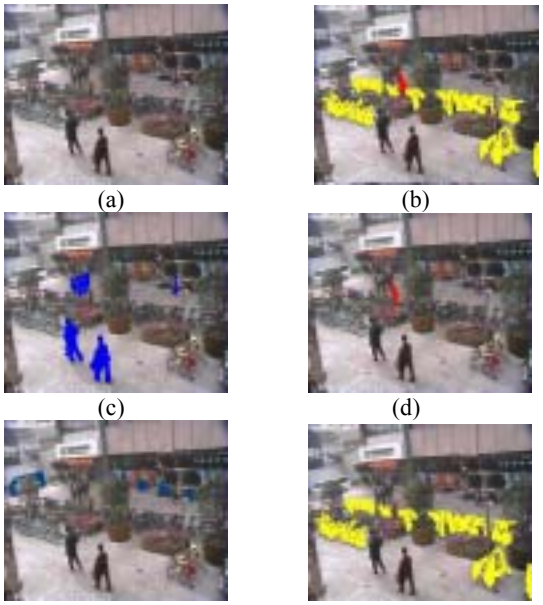
$$D(\tilde{p} \| q_\omega) = \sum_x \tilde{p}(x) \sum_{b\in\{0,1\}} \tilde{p}(b \mid x) \log \frac{\tilde{p}(b \mid x)}{q_\omega(b \mid x)} \qquad (2)$$

When the exponential model underestimates the expectation value of features f_i, its weight ω_i is increased. Conversely, ω_i is decreased when overestimation occurs.

Let *M* denote the layer model, besides the traditional background model M_{bg}, we trained the other four layer models: moving foreground model M_{mf}, stationary foreground model M_{sf}, moving background model M_{mb}, and stationary background model M_{sb}.

2.3. Layers-segmenting

Segmenting the video sequence into layers is straightforward with layers model. We here explain layers segmentation with the example shown as Fig2. Layers are segmented in different colors. Fig2-a is the original image in surveillance video. Fig2-c shows the moving objects, which are segmented with M_{mb}. They are passersby on the street. Objects of this layer are moving but not the target objects we need to detect in this system. So they are segmented to the moving background layer. The moving object in Fig2-d is segmented with M_{mf} to moving foreground layer. It is a suspicious bicycle theft, which is discriminated from the video as moving foreground layer. Fig2-e shows the motionless objects, which are segmented as stationary background layer. They are idle person in front of the building. But they cannot be simply segmented to traditional background layer, since they have the possibility to change into moving objects in foreground. Fig2-f shows the stationary foreground layer, which is segmented by M_{sf}. The layer consists bicycles parked in front of a building. The bicycles, which are



(a) (b) (c) (d) (e) (f)

Fig 2 (a-f): Segmented layers in different colors

target objects in the sequence are usually segmented as background in traditional BGS. Finally, we successfully detect all the objects of interest from the surveillance video in the moving foreground layer and stationary foreground layer, shown as Fig2-b.

The method described above allows us to identify layers pixels in each new frame while updating the description of each pixel's process. These labeled pixels then are segmented into regions by a two-pass, connected components algorithm [10]. Because this procedure is effective in determining all kinds of candidate objects, target objects can be easily characterized by their position, size, moments, and other information. These characteristics are useful for later processing and classification, and they can aid in the tracking process for not only moving objects but also stationary objects.

3. EXPERIMENTAL RESULTS

In this section, we describe the performances of layers segmentation by assuming that the segmented foreground layers really correspond to interest events.

The performances of layers segmentation are measured by counting how many times the system is able to segment the layers sequence containing objects related with particular events.

We first measure how the performances of the system vary with the complexity of the scene, where the complexity is identified with the number of persons moving in the guarded environment. For tests, we define three levels of complexity:

- low complexity (LC): two persons, at maximum, in the scene, corresponding to a maximum density of 0.12 person/m^2;
- medium complexity (MC): four persons, at maximum, in the scene, corresponding to a maximum density of 0.24/ m^2;
- high complexity (HC): more than four persons, in the scene, more than 0.24person/m^2;

Numerical results are shown in Fig3 that represents success rate mean value in segmenting layers considering different levels of scene complexity. The performance is compared with BGS (right rectangle), which simply detect the moving objects as foreground. It is possible to notice that, although the performances decay with the complexity increasing, good results are obtained also with a medium level of complexity of the scene.

Moreover, performances are measured for the precision of layer segmentation. Evaluations are performed by considering probabilities of success, false and miss detection in segmenting layers. They are full success (FS), partial success (PS), false detection (FD), and miss detection (MD). Two different kinds of success are considered. The results are shown in Fig4, in which it is possible to notice that the success rate is quite high (81%), and the probability of full success (69%) is considerably higher than the number of partial success (12%). It proves the efficiency of the proposed method.

4. CONCLUSIONS

In this paper, a novel method for layers segmentation has been presented with applications to automated visual surveillance. Our work aims at differentiating foreground pixels of both moving and stationary objects of interest from background pixels. Our method has the advantage over previous techniques in the sense that it dose not only handle scenes containing multiple backgrounds and illumination variations, but also effectively detects the target objects by segmenting layers with ME model. The performances of layers segmentation presented in Section 3 prove the efficiency of our work.

In the future, we would like to work on identifying event of interest based on the segmented layers. And it has still to be explored, which algorithm provides a good compromise between accuracy and computational complexity.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 60272031), Technology Plan Program of Zhejiang Province (2003C21010), Zhejiang Provincial Natural Science Foundation of China (M603202).

REFERENCES

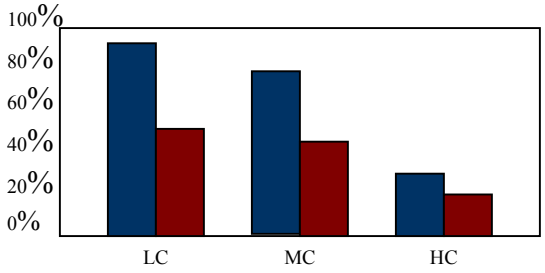


Fig 3: Success rate compared with BGS

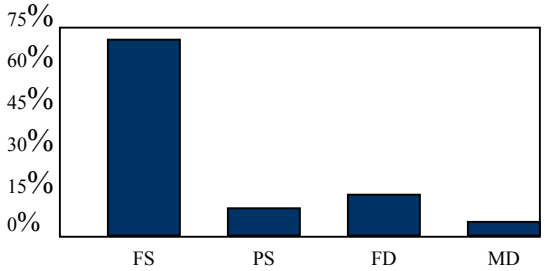


Fig4: Precision of layers segmentation

[1] A.R. Dick, and M.J. Brooks, "Issues in Automated Visual Surveillance," *International Conference on Digital Image Computing: Techniques and Applications*, 2003.

[2] T. Horprasert, D. Harwood, and L.S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," *IEEE Frame-Rate Applications Workshop*, 1999.

[3] C. Stauffer, and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-time Tracking," *IEEE Int. Conf. Computer Vision and Pattern Recognition*, Vol. 2, pp. 246-252, 1999.

[4] D.S. Lee, J.J. Hull, and B. Erol, "A Bayesian Framework for Gaussian Mixture Background Modeling," *IEEE International Conference on Image Processing*, 2003.

[5] F. Porikli and O. Tuzel, "Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis," *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2003.

[6] A. Elgammal, D. Harwood, and L.S. Davis, "Non-Parametric Model for Background Subtraction," *European Conf. Computer Vision*, Vol. 2, pp. 751-767, 2000.

[7] T.H. Chalidabhongse, K. Kim, D. Harwood, L. Davis, "A Perturbation Method for Evaluation Background Subtraction Algorithms," *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[8] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Background Modeling and Subtraction by Codebook Construction," *IEEE International Conference on Image Processing*, 2004.

[9] A.L. Berger, S.A. Della Pietra, and V.J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, Vol. 22, No.1, 1996.

[10] B.K. P. Horn, *Robot Vision*, pp.66-69, 299-333. The MIT Press, 1986.