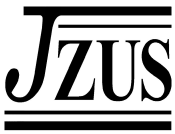
Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn



Video segmentation using Maximum Entropy Model

QIN Li-juan (秦莉娟)†, ZHUANG Yue-ting (庄越挺), PAN Yun-he (潘云鹤), WU Fei (吴 飞)

(*School of Computer Science, Zhejiang University, Hangzhou 310027, China*)

†E-mail: qinlijuan@hotmail.com

Received Jan. 25, 2005; revision accepted June 22, 2005

Abstract: Detecting objects of interest from a video sequence is a fundamental and critical task in automated visual surveillance. Most current approaches only focus on discriminating moving objects by background subtraction whether or not the objects of interest can be moving or stationary. In this paper, we propose layers segmentation to detect both moving and stationary target objects from surveillance video. We extend the Maximum Entropy (ME) statistical model to segment layers with features, which are collected by constructing a codebook with a set of codewords for each pixel. We also indicate how the training models are used for the discrimination of target objects in surveillance video. Our experimental results are presented in terms of the success rate and the segmenting precision.

Key words: Layers segmentation, Maximum Entropy Model, Visual surveillance
doi:10.1631/jzus.2005.AS0047 **Document code:** A **CLC number:** TP391.41

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 (Dick and Brooks, 2003).

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 (Cucchiara *et al*., 2003; Sen-Ching and Cheung, 2004). 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 had 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 BGS. Therefore, we propose the concept of layers segmentation to adaptively differentiate foreground pixels, which should be processed for identification or tracking, from background pixels, which should be ignored.

The BGS techniques can be classified into two broad categories: single-mode modelling and multi-mode modelling. The single-mode modelling (Horprasert *et al*., 1999) is limited to handling multiple backgrounds, like waving trees. There are three representative methods of multi-mode modelling. The mixture of Gaussians (MOG) (Stauffer and Grimson, 1999; Lee *et al*., 2003; Prokili and Tuzel, 2003) has been used to model complex, non-static backgrounds, but backgrounds having fast variations are not easily modelled with just a few Gaussians accurately, and may fail to provide sensitive detection (Elgammal *et al*., 2000). The non-parametric technique (Elgammal *et al*., 2000) cannot be used when long-time periods are needed to sufficiently sample the background

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

(Chalidabhongse *et al*., 2003). The Codebook (CB) algorithm constructs a highly compressed background model to deal with the main problems of BGS, and efficient in memory and speed compared with other background modelling techniques (Chalidabhongse *et al*., 2003). But all these methods focus on detecting moving objects from background no matter whether or not 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; and to (2) have segment layers which can adaptively detect the objects of interest as foreground. We develop the layers segmentation based on the CB background algorithm due to its advantage of meeting the first requirement of our system. The Maximum Entropy (ME) statistical model is introduced to build the layer model with automatic learning ability.

In Section 2 we introduce the mathematical structure of the ME model with refinements to make it practical to implement in layers segmentation. In Section 3 we give a short outline of the system framework and then describe the composing parts in subsections. Section 4 presents experimental results showing that our method is efficient and robust because of the layers segmentation. Finally, Section 5 presents conclusions and touches on future work.

ME MODEL

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

The ME model (Berger *et al*., 1996) constructs an exponential log-linear function that fuses multiple features to approximate the posterior probability of each layer. The estimated model, a posterior probability, is represented as $q_\omega\{b|x\}$, 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 represents the layer codewords for a candidate layer pixel. From x we compute a set of binary features.

$$f_i(x,b)=1_{\{g_i(x)=b\}}\in\{0,1\} \tag{1}$$

where $1_{\{\sim\}}$ is an indication function; g_i is a predictor of layer using the ith binary feature, generated from the codewords set (Hsu *et al*., 2004). f_i equals 1 if the prediction of predictor g_i equals b, and is 0 otherwise.

Given a labelled training set, we construct a linear exponential function for each pixel as

$$q_\omega(b\,|\,x)=\frac{1}{Z_\omega(x)}\exp\left\{\sum_i \omega_i f_i(x,b)\right\} \tag{2}$$

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 Eq.(2) is a valid conditional probability distribution. Basically, ω_i controls the weighting of ith feature in estimation of 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

$$\omega^*=\arg\max_\omega(\tilde{p}\,\|\,q_\omega) \tag{3}$$

where D is the Kullback-Leibler divergence defined as

$$D(\tilde{p}\,\|\,q_\omega)=\sum_x \tilde{p}(x)\sum_{b\in\{0,1\}}\tilde{p}(b\,|\,x)\log\frac{\tilde{p}(b\,|\,x)}{q_\omega(b\,|\,x)} \tag{4}$$

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.

LAYERS SEGMENTATION

Our work is composed of three parts as shown in Fig.1.

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 and features for the ME model are collected. 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.

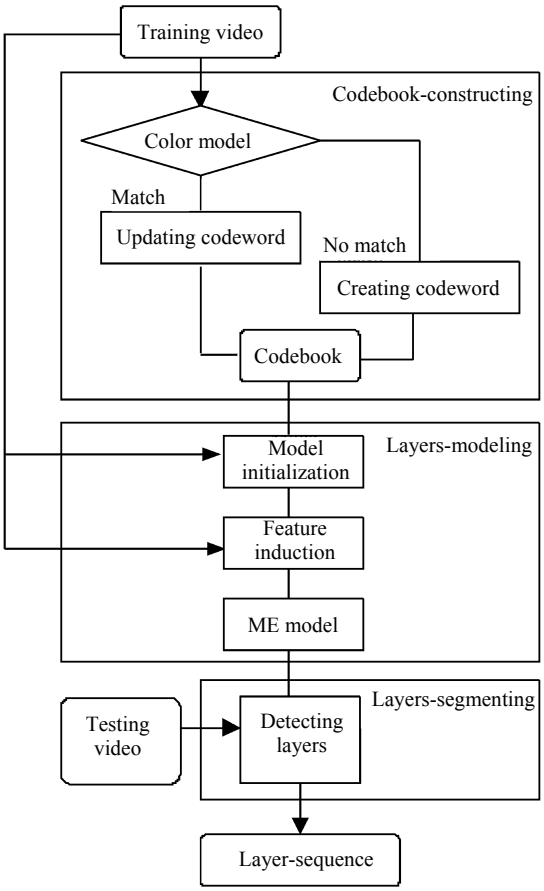


Fig.1 System diagram for layers segmentation

Codebook-constructing

The CB algorithm is encoded on a pixel-by-pixel basis. A pixel is represented by a codebook, consisting of one or multiple codewords. Let x be a training sequence for a single pixel consisting of N RGB-vectors $\boldsymbol{x}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$. Let $\boldsymbol{C}=\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_L\}$ 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,…,L, consists of an RGB vector

$$v_i = (\bar{R}_i, \bar{G}_i, \bar{B}_i) \tag{5}$$

and a 6-tuple

$$aux_i = \langle minI_i, maxI_i, freq_i, \lambda_i, firstT_i, lastT_i \rangle \tag{6}$$

The tuple aux_i contains brightness values and temporal variables described here. *minI*, *maxI* are the min and max brightness, respectively, that the codeword accepted. *freq* 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. *firstT*, *lastT* are the first and last access time, 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 model (Kim *et al*., 2004), which includes a color distortion measure and brightness bounds.

Layers-modelling

The ME model constructs an exponential log-linear function that fuses multiple features in codewords to approximate the posterior probability of each layer.

Given a set of prospective binary features F and an initial ME model q described in Section 2, the model is improved into $q_{\alpha,h}$ by adding a new feature $h \in F$ with a suitable weight α, represented as

$$q_{\alpha,h}(b \mid x) = \frac{\exp\{\alpha h(x,b)\} q(b \mid x)}{Z_\alpha(x)} \tag{7}$$

where $Z_\alpha(x)$ is the normalization factor. A greedy induction process is used to select the feature that has the largest improvement in terms of gains, divergence

reduction, or likelihood increase. The selected feature h^* in each iteration is represented in Eq.(8). h^* is then removed from the candidate pool F. The induction process iterates with the new candidate set $F-\{h^*\}$ till stopping criterion is reached (upper bound of the number of features or lower bound of the gain).

$$h^* = \arg\max_{h\in F}\left\{\sup_{\alpha}\left\{D(\tilde{p}\,\|\,q)-D(\tilde{p}\,\|\,q_{\alpha,h})\right\}\right\}$$
$$= \arg\max_{h\in F}\left\{\sup_{\alpha}\left\{L_{\tilde{p}}(q_{\alpha,h})-L_{\tilde{p}}(q)\right\}\right\} \quad (8)$$

We have described the parameter estimation and feature induction processes from a pool of binary features in Section 2. 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}.

Layers-segmenting

Segmenting the video sequence into layers is straightforward with layer-models. Here, we set a distant threshold DTh for the layers segmentation. Using the layer-models trained with the given surveillance video, we achieve the segmentation probabilities of each layer for a pixel. The distance between the segmentation probabilities is computed as:

$$D_{P1P2} = |P_1 - P_2| \quad (9)$$

P_1 and P_2 represent the segmentation probability of each layer: P_{bg}, P_{mf}, P_{sf}, P_{mb} and P_{sb}. If

$$\min D_{P_1P_2} \geq DTh \quad (10)$$

this pixel is detected as the layer whose probability is the largest among the probabilities. For example, we now have probabilities for a pixel as P_{bg}=0.000; P_{mf}=0.719; P_{sf}=0.024; P_{mb}=0.257; P_{sb}=0.000, and the threshold is set to DTh=0.302, which is learned from the training data. Apparently, all the distances between P_{mf} and the other four are larger than the threshold. So, this pixel is detected as moving foreground. There are also conditions that no layer's probability is satisfied with the limitation of DTh. In our work, such pixels are classified into one of the foreground layers according to the layer probability (the larger one between P_{mf} and P_{sf}) for the high security requirement on false positive alarm.

We explain the segmented layers with the example shown as Fig.2. Layers are respectively displayed in different images.

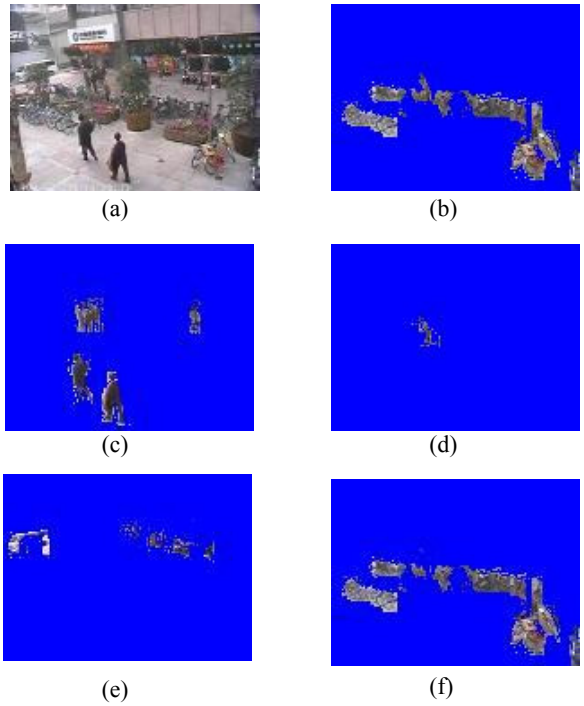


Fig.2 Example of segmented layers

(a) Original image; (b) Foreground; (c) Moving background layer; (d) Moving foreground layer; (e) Stationary background layer; (f) Stationary foreground layer

Fig.2a is the original image in the surveillance video. Fig.2c 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 Fig.2d 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. Fig.2e 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. Fig.2f 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 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 Fig.2b.

The method described above allows us to identify layers pixels in each new frame while updating the description of each pixel's process. These labelled pixels then are segmented into regions by a two-pass, connected components algorithm (Horn, 1986). 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.

EXPERIMENTAL RESULTS

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

We use 25 MPEG-4 video sequences as experimental data, 15 of which are used for CB constructing and ME model training. The remaining 10 sequences are used as test data. All the sequences are generated from a 30 min real-world surveillance video. The monitoring scene is shown as Fig.2, which is a bicycle park in front of a building.

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

We first measure how the performance of the system varies 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 person/m^2;

High complexity (HC): more than four persons, in the scene, more than 0.24 person/m^2;

Ten video sequences of 25 frame/s are used as test data in this experiment. The sequences are classified into LC, MC and HC according to the definition of scene complexity. They have similar object frames which is convenient for the comparison of different levels of complexity. Numerical results are listed in Table 1 and Fig.3 represents success rate mean value in segmenting layers. The performance of our method is compared with CB (Kim *et al*., 2004), Kernel (Elgammal *et al*., 2000) and MOG (Stauffer and Grimson, 1999) (the ordinal rectangles in Fig.3 from left to right are Layer, CB, Kernel and MOG), which simply detect the moving objects as foreground. It is possible to notice that, although the performance of our method decays like the BGS methods with the complexity increasing, good results are obtained also with a medium level of complexity of the scene.

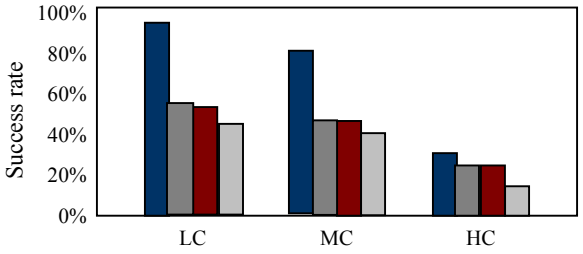


Fig.3 Success rate compared with BGS (The ordinal rectangles from left to right are Layer, CB, Kernel and MOG)

Moreover, performances are measured for the precision of layer segmentation. Evaluations are performed by considering probabilities of success,

Table 1 Numerical results for detection comparison

Video	Sequences No.	Frames No.	Object frames No.	Success detection No.			
				Layer	CB	Kernel	MOG
LC	2	2975	1375	1275	700	575	525
MC	4	6000	1425	1125	575	550	475
HC	4	6000	1300	325	250	250	150

false and miss detection in segmenting layers. They are full success (FS) which means all interest objects and only interest objects have been detected in front layer, partial success (PS) which means all interest objects have been detected but the other objects have also been detected in the front layer, false detection (FD) which means some interest objects have not been detected in the front layer, and miss detection (MD) which means we have lost the interest objects. In this part, the concept of success detection in the first experiment has been subdivided into FS and PS, which helps us to evaluate the success in more detail.

Since the mean success rate for high complexity is relative poor in the first experiment, the precision test is only implemented to the data composed of low complexity and middle complexity. The results are shown in Fig.4, 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.

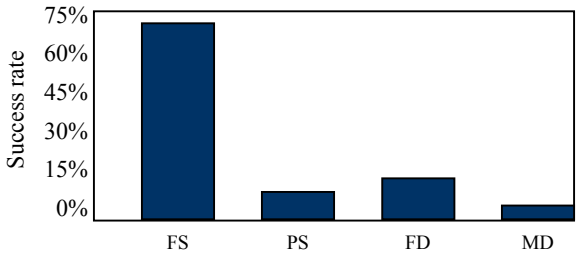


Fig.4 Precision of layers segmentation

CONCLUSION

In this paper, a novel method for layers segmentation is 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 moving and stationary target objects by segmenting layers with ME model. The performances of layers segmentation presented in Section 4 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.

References

Berger, A.L., Della Pietra, S.A., Della Pietra, V.J., 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1):39-71.

Cucchiara, R., Grana, C., Piccardi, M., Prati, A., 2003. Detecting moving objects, ghosts and shadows in video streams. *IEEE Trans. on Patt. Anal. and Machine Intell.*, **25**(1):1337-1342.

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

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

Elgammal, A., Harwood, D., Davis, L.S., 2000. Non-Parametric Model for Background Subtraction. European Conf. Computer Vision, **2**:751-767.

Hsu, W., Chang, S.F., Huang, C.W., Kennedy, L., Lin, C.Y., Lyengar, G., 2004. Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation. IS&T/SPIE Symposium on Electronic Imaging: Science and Technology–SPIE Storage and Retrieval of Image/Video Database.

Horprasert, T., Harwood, D., Davis, L.S., 1999. A Statistical Approach for Real-time Robust Background Subtraction and Shadow Detection. IEEE Frame-Rate Applications Workshop.

Horn, B.K.P., 1986. Robot Vision. The MIT Press, p.66-69, 299-333.

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

Lee, D.S., Hull, J.J., Erol, B., 2003. A Bayesian Framework for Gaussian Mixture Background Modelling. IEEE International Conference on Image Processing.

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

Sen-Ching, S., Cheung, C.K., 2004. Robust Techniques for Background Subtraction in Urban Traffic Video. Proceedings of SPIE.

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