ACCURATE DYNAMIC SCENE MODEL FOR MOVING OBJECT DETECTION

Hong Yang, Yihua Tan, Jinwen Tian, Jian Liu

State Key Laboratory for
Multi-spectral Information Processing Technologies
Huazhong University of Science and Technology, Wuhan 430074, PR China
E-mail: hongle@vip.sina.com

ABSTRACT

Adaptive pixel-wise Gaussian mixture model (GMM) is a popular method to model dynamic scenes viewed by a fixed camera. However, it is not a trivial problem for GMM to capture the accurate mean and variance of a complex pixel. This paper presents a two-layer Gaussian mixture model (TLGMM) of dynamic scenes for moving object detection. The first layer, namely real model, deals with gradually changing pixels specially; the second layer, called on-ready model, focuses on those pixels changing significantly and irregularly. TLGMM can represent dynamic scenes more accurately and effectively. Additionally, a long term and a short term variance are taken into account to alleviate the transparent problems faced by pixel-based methods.

Index Terms— Gaussian mixture model, background subtraction, moving object detection

1. INTRODUCTION

Background subtraction is a common technology used in some vision applications, such as surveillance, where the camera is mechanically fixed. A crucial step of this approach is to model background accurately and robustly.

The statistical method of modeling background involves calculating the likelihood $P(x \mid \Theta_b)$ of each pixel. A simple likelihood may be assumed to consist of a Gaussian distribution [1] [2]. The single Gaussian model (SGM) is suitable for handling gradual illumination changes in a scene. Unfortunately, SGM fails to deal with complex real scenes. In real world, even a pure background pixel may exhibit tremendous changes of intensity or color caused by sudden illumination change, periodic motion such as rippling water, swaying vegetation, flicking flag, raining, and even jittering sensors, etc. These situations imply a bi-modal or multi-modal of a pixel. Additionally, a pixel is usually occupied alternately by both background and foreground randomly. This fact appreciates the need of a two-layer model to apprehend a dynamic scene.

Previous work suggested the efficiency of dealing with these complicated situations mentioned above by Gaussian mixture model (GMM), which is followed in this paper. Stauffer *et al.* [3] modeled pixel intensity by a mixture of Gaussian distributions to account for the multimodality of background. Zivkovic [4] extended the work of Stauffer by adaptively determining the number of models combined in the likelihood. Lee *et al.* [5] proposed an adaptive learning rate schedule for each Gaussian to improve the convergence speed. Porikli *et al.* [6] and Martel-Brisson *et al.* [7] proposed methods for modeling casting shadow as well as background with GMM. Moreover, Elgammal *et al*. [8] proposed non-parametric estimation method for pixel-wise background modeling. They used kernel density estimation (KDE), as a generalization of Gaussian mixture model, to establish membership.

However, the traditional background subtraction based on pixel-wise GMM is subject to two limitations. Firstly, GMM commonly fails to obtain the accurate mean and variance of a complicated pixel, due to the uniform mechanism of update, i.e., the foreground and background pixels are treated by one unified model. Secondly, conventional approach to moving object detection is hard to treat with the situation when foreground and background possess very similar intensity or color, resulting in great false negative (holes in detected regions).

We proposed a new mechanism of two-layer Gaussian mixture model. This scheme endows the model with the ability to learn the parameters of background model effectively and accurately. Additionally, we address the transparent problem (holes) with the usage of temporal information. A long term variance and a short term one of a pixel is taken into account to reduce false negative.

2. TWO-LAYER GAUSSIAN MIXTURE MODEL

In this section, a novel mechanism of two-layer Gaussian mixture model is presented for modeling dynamic scenes. One layer of the model, namely real model, is designed to adapt with gradually changing scenes and to remember the

learned model for a reasonably long period of time. The other layer, called on-ready model, is designated to deal with new pixel values not included in the real model.

These two models consider pixels in a different way, corresponding to background layer and foreground layer of a scene, respectively. The clearly separated framework of TLGMM exhibits wonderful property to acquire the correct values of mean and variance. Stability of real model and swift response of on-ready model give rise to the significantly improved adaptability to dynamic scenes. The architecture of TLGMM and the interaction between real model and on-ready model are described in details in the succeeding sections. Formulas are presented in one dimension for simplicity, while the extension to multiple dimensions is straightforward.

2.1. Real Model

The pixel-based Gaussian mixture model of background is a weighted combination of several Gaussian functions,

$$P(x \mid \Theta_b) = \sum_{i=1}^{N_1} \omega_i p_i(x \mid \mu_i, \sigma_i^2) \tag{1}$$

where $P(x \mid \Theta_b)$ is the probability of a certain pixel of background having intensity of x at time t, the subscript t is omitted for simplicity without confusion. $p_i(x \mid \mu_i, \sigma_i^2)$ is a Gaussian function with mean μ_i and variance σ_i^2, and ω_i is the weight of the i-th Gaussian. The summation of all weights equals 1. N_1 is the number of Gaussians. The model can be denoted as $\Theta_b = \{\omega_i, \mu_i, \sigma_i^2, i = (1, \cdots, N_1)\}$.

The algorithm starts with a coarse decision to learn the exact parameters. When there are not sufficient priors about moving objects, we have to set a global threshold K_{th} of probability to check if a pixel belongs to background or not. For GMM, the decision rule is simplified as that, if a pixel matches anyone of the first B Gaussians, it is classified as background. B is calculated in eq. (5), The match is defined as the value of a pixel falling in the neighborhood of a mean, i.e., $| x - \mu_i | < 2.5\sigma, i = (1, \cdots, N_1)$.

The matching data are used to update the real model with a K-means estimation of EM algorithm. Every new value of the pixel is checked against the real model until a match is found. If there is no match, the pixel value will be treated by on-ready model discussed in the succeeding section. The mean and variance of the matching model are updated as follows,

$$\mu_i \leftarrow (1 - \beta_i)\mu_i + \beta_i x \tag{2}$$

$$\sigma_i^2 \leftarrow \begin{cases} \sigma_{\max}^2, & if \quad \sigma_i^2 > \sigma_{\max}^2 \\ (1-\beta)\sigma_i^2 + \beta((x-\mu_i)^2 + \sigma_{\min}^2), & else \end{cases} \tag{3}$$

where x is the current value of a pixel; $\sigma_{\max}^2$ and $\sigma_{\min}^2$ are the maximum and minimum limitation of variance, respectively; $\beta_i = \beta \cdot P_i(x \mid \mu_i, \sigma_i^2)$, where $P_i(x \mid \mu_i, \sigma_i^2)$ is a Gaussian function, and β is between 0 and 1. β can be set large enough to enable the model to adapt with the rapid and gradual changes of the scene and the learned parameters are more likely the real value. The choice of parameter β_i in eq. (2) makes the model to reach the current value of a pixel quickly and smoothly. The means and variances of other un-matching models remain unchanged for the moment.

Researchers in literatures mostly afford a large initialized value of variance and make it decreasing slowly and almost continuously in common GMM. Stauffer *et al.* [3] introduced the variational attenuation parameter in their update scheme, which implies that the Gaussian model is more difficult to expand than to shrink, and theoretically leads to the variance converging to a very small value even close to zero eventually. This treatment obviously loses the genuine variance and will cause more false positive. An effective way to attain the correct variance is to employ a constant between 0 and 1 as the attenuation parameter. On the other hand, since complex pixel usually changes significantly, the variance of a model is prone to expand wide excessively, resulting in that one dominant Gaussian suppresses all others, which will produce more false negative for moving object detection. To avoid these problems, the variance should be bounded, as in eq. (3), between reasonable mini-max limitations.

The weights are the probability of components in the mixture of Gaussians. We do not expect them changing too quickly to hold the learned model for a reasonably long period of time. A small parameter α is appropriate for this purpose,

$$\omega_i \leftarrow (1-\alpha)\omega_i + \alpha M_i \qquad i = (1, \cdots, N_1) \tag{4}$$

where M_i is 1 for the matching component and 0 for the others. The summation of all weights holds 1 exactly after update and they are not necessary to be renormalized. It is not difficult to know, $K_{th} \approx C\omega_i / \sqrt{2\pi}\sigma_i$, where C is between 0 and 1. Hereby, the components of the model are aligned in descending order according to ω_i / σ_i. The first B components are treated as effective background states,

$$B = \arg\min_k \left\{ \sum_{i=1}^{k} \omega_i > T_1 \right\} \tag{5}$$

A larger T_1 involves more components of real model accounting for the background.

2.2. On-ready Model

Similarly, another Gaussian mixture model accounting for foreground is constructed as follows,

$$P(x \mid \Theta_f) = \sum_{j=1}^{N_2} \omega_j p_j(x \mid \mu_j, \sigma_j^2) \tag{6}$$

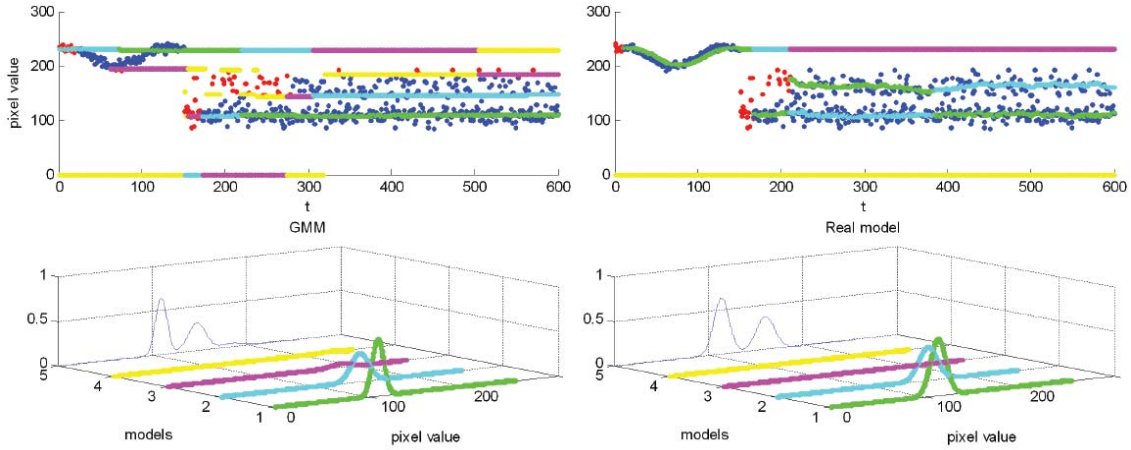

Fig.1. The learning process of GMM (left column) and TLGMM (right column). Red and blue points are synthesized sample data (red indicates pixel misclassified as foreground), and the color curves denote the means of models (up row); the thin blue curves are the combination of 4 Gaussians (down row).

In fact, since the foreground is more complicated, it may be impractical to represent various moving objects with GMM accurately. On-ready model is not designated to address this issue exactly, but aims at capturing the most probability of a pixel state belonging to background while not contained in the existing real model.

The mean and variance are updated the same way as real model. While the weights update as follows,

$$\omega_j \leftarrow (1-\alpha_j)\cdot\omega_j + \alpha_j \cdot M_j, \quad j=(1,\cdots,N_2) \tag{7}$$

$$\alpha_j = \frac{1}{1+e^{-(t_j - T_2)}} \tag{8}$$

$$t_j \leftarrow \quad t_j + P(x \mid \mu_j, \sigma_j^2) \tag{9}$$

M_j is 1 for the matching component and 0 for others. t_j accounts the appearance times regarding to $P(x \mid \mu_j, \sigma_j^2)$. T_2 is mainly in charge of the duration when a pixel is absorbed by background. A large value of T_2 is suitable for scenes containing slowly moving objects; a small T_2 is appropriate for adapting with dynamic scene as fast as possible. When $\omega_j > th_1$, th_1 is a threshold between 0 and 1, an exchange is triggered between real model and on-ready model. The first component of real model is replaced by this newly confirmed background state, and the corresponding component of on-ready model is reinitialized.

3. OBJECT DETECTION WITH TEMPORAL CUES

A significant disadvantage of background subtraction based upon pixel-wise model is that, when the color of moving objects is close to or even falls in the 2.5σ interval of the mean, they will be undoubtedly classified as background. It will cause great false negative especially when the variance is large, unfortunately, which is the case of dynamic scene.

Temporal information may provide a compensatory approach to distinguish similar pixels from background. The variance of a complex pixel usually turns small when it is occupied by moving objects. Thereby, a long term variance model and a short term one for every pixel are established and updated as follows,

$$\sigma_s^2 \leftarrow (1-\beta_s)\sigma_s^2 + \beta_s \sigma_t^2 \tag{10}$$

$$\sigma_l^2 \leftarrow \begin{cases} (1-\beta_l)\sigma_l^2 + \beta_l \sigma_t^2 & if\ x \in BG \\ (1-\beta_l')\sigma_l^2 + \beta_l' \sigma_t^2 & else \end{cases} \tag{11}$$

where σ_l^2 and σ_s^2 are the long term variance and short term one of a pixel, respectively; Their corresponding update parameters β_l and β_s are between 0 and 1, $\beta_l < \beta_s$, $\beta_l' = 0.1\beta_l$; $\sigma_t^2 = (x_t - x_{t-1})^2$ is the variance of consecutive frames. When x is classified as background, the long term variance σ_l^2 updates with a suitable β_l, otherwise it is updated by a smaller parameter to adapt with the situation where a moving object stops for a long interval.

A normalized difference between long term and short term variance are calculated,

$$\sigma_d^2 = \frac{|\sigma_l^2 - \sigma_s^2|}{\sigma_l^2 + \sigma_s^2} \tag{12}$$

For those pixels incorrectly classified as background by GMM, if they satisfy $\sigma_d^2 > th_2$, they will be drawn back to foreground, which will substantially reduce false negative.

4. EXPERIMENTAL RESULTS

Experiments were conducted on both synthesized data and real video sequences to evaluate the performance of the proposed method. Fig.1 shows the learning processes of TLGMM on synthesized data, compared with GMM. The pixel intensity of the first 150 frames changed slightly and

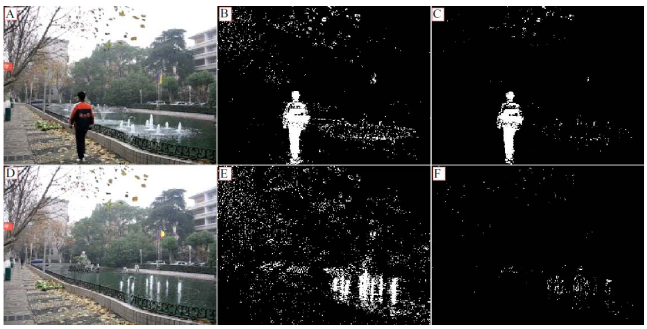

Fig.2. The frames 39 (A) and 1161 (D) of a video sequence; the raw results (not using morphology) obtained by GMM, (B) and (E); TLGMM, (C) and (F).

gradually. TLGMM tracked the correct mean successfully at this stage, while GMM lost the genuine value. From the frame 151, two Gaussian distributions, $N(110,100)$ and $N(160,200)$, simultaneously appeared. After a desired interval (flexibly controlled by T_2), TLGMM jumped to the proper values and kept maintaining them smoothly; while the standard GMM deviated from the true values again, resulting in false positive (indicated by red points).

Two sequences of outdoor scene were tested. The first sequence captures a dynamic scene containing spouting fountains, swaying trees and flicking flags. The fountains ceased in the middle of the sequence and waked at the frame 1158. The raw results are shown in Fig.2. The up row suggests that TLGMM can model dynamic scenes more accurately than GMM. The second row indicates that TLGMM can hold the learned model for a reasonably long interval. This property may be in favor of treating non-strictly periodic motions, such as casting shadows.

The well-known 'WavingTrees' sequence is tested to demonstrate the effectiveness of temporal cue. The results are shown in Fig.3. Large holes in the chest of the man are filled substantially, Fig.3 (D). On the other hand, we noted that the temporal cue is restrictedly suitable for slowly moving objects.

5. CONCLUSIONS

GMM has become a standard method dealing with complicated dynamic scenes, but it is not an effortless issue to attain the accurate mean and variance of a complex pixel. In this paper, we proposed a novel mechanism of two-layer GMM accounting for dynamic scene. The proposed method can achieve accurate background models, significantly improving the performance on moving object detection. Additionally, temporal information is considered to solve the transparent problem of pixel-based methods. Experiments on synthesized data and real video illustrate the good performance of the proposed method.

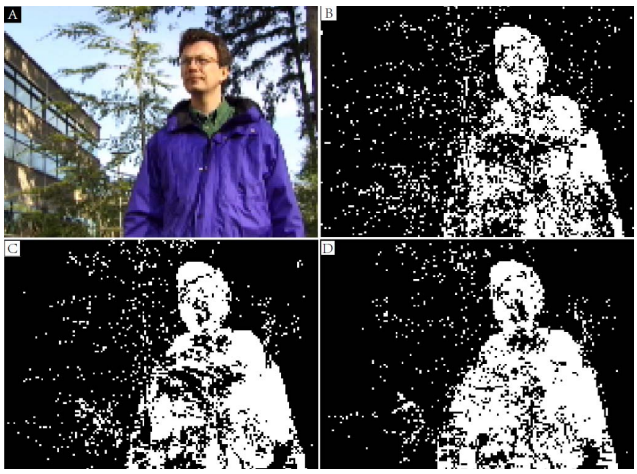

Fig.3. (A) a frame of 'WavingTrees' sequence; the raw results detected by (B) GMM, (C) TLGMM, and (D) TLGMM with temporal information.

6. REFERENCES

[1] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body", *IEEE Trans. PAMI,* vol.19, no.7, July 1997.

[2] N. Friedman, and S. Russell, "Image segmentation in video sequences: A probabilistic approach", In *Proc. of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, Aug. 1-3, 1997.

[3] C. Stauffer, and W. Grimson, "Adaptive background mixture models for real-time tracking," *CVPR,* 1999.

[4] Z. Zivkovic and F. Heijden, "Recursive Unsupervised Learning of Finite Mixture Models", *IEEE Trans. on PAMI,* vol.26, no.5, 2004.

[5] D. S. Lee, "Effective Gaussian Mixture Learning for Video Background Subtraction", *IEEE Trans. PAMI*, vol.27, no.5, 2005.

[6] F. Porikli and J. Thornton, "Shadow Flow: A Recursive Method to Learn Moving Cast Shadows", *ICCV*, 2005.

[7] N. Martel-Brisson and A. Zaccarin, "Moving Cast Shadow Detection from a Gaussian Mixture Shadow Model", *CVPR*, 2005.

[8] A. Elgammal, D. Harwood, L. Davis, "Non-parametric Model for Background Subtraction", *ECCV,* 2000.