Moving Foreground Detection Based On Spatio-temporal Saliency

Yang Xia[1], Ruimin Hu[1], Zhongyuan Wang[1] and Tao Lu[1, 2]

[1] National Multimedia Software Engineering Research Center, Computer School of Wuhan University, Wuhan University, Wuhan, 430072, China

[2] Hubei Province key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, 430074, China

Abstract

Detection of moving foreground in video is very important for many applications, such as visual surveillance, object-based video coding, etc. When objects move with different speeds and under illumination changes, the robustness of moving object detection methods proposed so far is still not satisfactory. In this paper, we use the semantic information to adjust the pixel-wise learning rate adaptively for more robust detection performance, which are obtained by spatial saliency map based on Gaussian mixture model (GMM) in luma space and temporal saliency map obtained by background subtraction. In addition, we design a two-pass background estimation framework, in which the initial estimation is used for temporal saliency estimation, and the other is to detect foreground and update model parameters. The experimental results show that our method can achieve better moving object extraction performance than the existing background subtraction method based on GMM.

Keywords: *Moving Object Detection, Background Subtraction, Visual Saliency, Gaussian Mixture Model*

1. Introduction

Detection of moving objects in video sequences is essential in many applications, such as visual surveillance [1], object-based video coding [2], etc. The most popular approach for moving object detection is the background subtraction, which maintains an up-to-date model of the background and detects moving objects as those that deviate from the background model. A classical parametric background model with adaptive Gaussian mixture models was introduced by C.Stauffer and W.Grimson [3]. Each pixel is modeled as a mixture of Gaussians (MOG) which can be updated on-line. Object detection is performed by matching luminance and color of every pixel with the most likely background Gaussians distribution. But this kind of parametric approach may fail when the density function is complicated. To solve the problem, Elgammal [4] and Mittal [5] used kernel density estimate to model more complex background. Recently, Zivkovic [6] improved the GMM based approach [3] by adaptively selecting an appropriate number of Gaussian components for each pixel. However, since the update rate of background model parameters are usually slower than the rate of illumination changes, above methods are not robust to illumination changes. At meanwhile, the background models detect foreground pixel only by the temporal information of the pixel, so the background models are inaccurate when the objects move slowly, with only the edges of outstanding objects labeled salient.

Visual saliency is another way to detect objects from video, which can use more spatial information to increase detection robustness. As a representative work of visual saliency, Itti set up computational visual saliency models for still image [7] and video [8] by simulating the pre-attentive selection mechanism. The regions which have high saliency values may be considered as salient objects. However the experiment results [8] show that the method may mistake a lot of background regions for foreground.

It is observed that background subtraction based methods are sensitive to illumination changes, while visual saliency is more robust to illumination changes but mistakes many background areas for foreground in practice. If the background subtraction based methods incorporate some spatial information obtained by visual saliency, the foreground detect performance will be improved. In this paper, a novel object detection algorithm based on spatio-temporal saliency is proposed, which takes advantage of background subtraction and visual saliency. In our approach, spatial saliency based on Gaussian mixture model in luminance space and temporal saliency obtained by background subtraction, are calculated as an auxiliary knowledge to adjust the pixel-wise learning rate of object detection model adaptively, which makes foreground detection more robust when objects move with different speeds and under illumination changes.

The rest of the paper is organized as follows. In Sec. 2 we analyze the drawback of traditional GMM based object detection method. Sec. 3 presents the proposed algorithm, and Sec. 4 shows the experimental results on the CAVIAR's dataset. Finally, Sec. 5 gives conclusions.

2. The Analysis of Traditional Model

Zivkovic [6] used a mixture model consisting of Gaussian distributions to estimate a density distribution from a sequence for a pixel at a location x, which can be denoted by (1)

$$P(I_{t,x}) = \sum_{n=1}^{N} w_{t-1,x,n} * \frac{1}{\sqrt{2\pi\sigma^2_{t-1,x,n}}} \exp\left(-\frac{\left(I_{t,x} - \mu_{t-1,x,n}\right)^2}{2\sigma^2_{t-1,x,n}} \right) \quad (1)$$

where $w_{t-1,x,n}$ is mixture weight for the n th Gaussian model, $\mu_{t-1,x,n}$ and $\sigma^2_{t-1,x,n}$ are the mean and variance of Gaussian model, and N is the number of Gaussian models.

In GMM based object detection method, $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ must be updated when new data $I_{t,x}$ comes. A learning rate α is used to limit the influence of past data and absorb coming information. When $I_{t,x}$ matches the n th Gaussian model, the $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ can be update as follow:

$$w_{t,x,n} = w_{t-1,x,n} + \alpha(1 - w_{t-1,x,n}) \quad (2)$$

$$\mu_{t,x,n} = \mu_{t-1,x,n} + (\alpha / w_{t-1,x,n}) * (I_{t,x} - \mu_{t-1,x,n}) \quad (3)$$

$$\sigma^2_{t,x,n} = \sigma^2_{t-1,x,n} + (\alpha / w_{t-1,x,n}) * ((I_{t,x} - \mu_{t-1,x,n})^2 - \sigma^2_{t-1,x,n}) \quad (4)$$

In [6] α is set to 0.001 to guarantee that the objects that move slowly can be correctly detected. But with illumination changes, the adjustment of $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ cannot keep up with the change of a scene due to the low learning rate, and some parts of background may be considered as foreground. Meanwhile, because the learning rate is low, the misclassification event will last for a long time. On the contrary, if we set α larger, the objects or parts of objects with slow motion speed may be seen as background. Therefore, the performance of object detection based on GMM depends on the selection of learning rate.

3. The Proposed Method

In our opinion, the learning rate of foreground in GMM should be different from that of background. In this paper, we use the semantic information to adjust the pixel-wise learning rate adaptively for more robust detection performance, which are obtained by spatial saliency map based on Gaussian mixture model (GMM) in luma space and temporal saliency map obtained by background subtraction. In addition, we design a two-pass background estimation framework, in which the initial estimation is used for temporal saliency estimation, and the other is to detect foreground and update model parameters. The proposed algorithm consists of the following three steps.

3.1 Spatial Saliency Analysis

In this section, we calculate the spatial saliency based on Gaussian mixture model. Firstly, we separate a frame into several clusters based on GMM [9], and then generate spatial saliency map by calculating the weighted distance from each point to the cluster centers. We use algorithm 1 to obtain the image clusters.

Algorithm 1: **Image frame cluster based on GMM**

Input: the current video frame

1. Obtain luma histogram of the current frame, and find the local maximums p_i of the histogram
2. Remove small peaks of the histogram. If $\|p_i - p_j\|_2 < \varepsilon$, the bigger peak will be preserved.
3. Use the remaining peaks as initial mean $\mu_i = p_i$ $i = 1,...,m$, m is the number of remaining peaks.
4. Use the midpoint between two peaks which can identify interval of an initial Gaussian cluster, denoted as $[bl\ br]$, to calculate initial covariance matrices of each Gaussian cluster.
$$\sigma_i^2 = \sum_{q=bl}^{q=br} \left(h_q - \mu_i\right)^2 * c(h_q) \Big/ \sum_{q=bl}^{q=br} c(h_q)$$
where $c(h_q)$ is the histogram value at h_q
5. Normalize the heights of peaks and set them as the initial weights of Gaussian clusters, denoted as w_i
$$w_i = c(\mu_i) \Big/ \sum_{i=1}^{m} c(\mu_i)$$
6. Run EM algorithm to find more accurate w_i, μ_i, σ_i^2.

Output: w_i, μ_i, σ_i^2.

Then we use the Gaussian clusters to calculate spatial saliency. We suppose the luminance intensities of background centralize around several peaks in the histogram, and the outstanding objects in picture have different luminance intensities from background. Furthermore, we believe that the areas covering the outstanding objects are much smaller than the background area. Based on the hypothesis, we divide Gaussian clusters into background clusters and foreground clusters. We sort

w_i by descending order, find the first k weights which satisfy $\sum_{j=1}^{k} w_j > \eta$, and set the corresponding Gaussian clusters to background clusters. The left clusters are regarded as foreground clusters.

Based on the labels of cluster, the spatial saliency map is obtained by calculating weighted distance from each point to the cluster centers, which is shown as follow:

$$ss_{t,x} = \rho * \sum_{i=1}^{k} w_i^2 * (I_{t,x} - \mu_i^B)^2 + (1-\rho) \sum_{j=1}^{m-k} w_j^2 * (I_{t,x} - \mu_j^F)^2 \quad (5)$$

where μ_i^B is the mean of background cluster, μ_j^F is the mean of foreground cluster, and ρ is set to 0.6 in this paper. Because background is much bigger than the moving objects, the mixing weights of background clusters are usually larger. So the outstanding objects which are much different from background clusters can be marked salient.

3.2 Spatio-temporal Saliency Map Generation

Since spatial saliency may mark some outstanding objects which are part of background, in this section, we combine spatial saliency and temporal saliency to get more accurate saliency information. For moving object detection, the moving foreground can be regarded as temporal saliency, so we use GMM to generate temporal saliency map.

First, we use the recent GMM based background subtraction method [6] to obtain a preliminary binary object map $fg_{t,x}$, where $fg_{t,x}$ equals 0 in background region and equals 1 in object region.

After obtaining the binary foreground map, temporal saliency map can be obtained by (6) and we find θ equals 0.3 can get good performance.

$$st_{t,x} = \theta * (1 - fg_{t,x}) + (1-\theta) * fg_{t,x} \quad (6)$$

Because the salient areas in spatial saliency map include some other static objects, we do not use the sum of temporal and spatial saliency maps, and we find the product of temporal and spatial saliency maps can eliminate some error. So a single saliency map is generated by.

$$STs_{t,x} = ss_{t,x} * st_{t,x} \quad (7)$$

It is observed that the spatio-temporal saliency contains the preliminary semantic information of the frame which describe the probability of a pixel belong to foreground.

(a) (b) (c)

Fig. 1 An example of Spatio-temporal saliency and learning rate scale function (a) Frame 90 from"Walk1" (b) Spatio-temporal saliency extraction result (c) learning rate scale function

3.3 Learning Rate Control Scheme

In this section, we use GMM again to detect moving objects. After obtaining the spatio-temporal saliency map, the learning rate of GMM model can be adaptively adjusted according to the saliency semantic information. We choose logistic function as learning rate scale function, as shown as Fig 1. Supposed that μ_{st} is the mean of saliency map, and σ_{st} is the variance of saliency map. We use the threshold $\tau = \mu_{st} + \sigma_{st}$ to distinguish salient region and un-salient region. The learning rate scale function can be defined as follow:

$$SF_{t,x} = a / \left(1 + \exp(b * (STs_{t,x} - \tau))\right) + c \quad (8)$$

where a is 999, and c is 1.0 in this paper. b is a parameter to control the size of distinguish belt, which is set to 0.1 in this paper. So the learning rate of GMM can be adjusted as follow:

$$\alpha_{t,x} = \gamma * SF_{t,x} \quad (9)$$

where γ is 0.0001. It can be found that the range of $\alpha_{t,x}$ is [0.0001, 0.1], and if the regions have higher saliency, the lower learning rate $\alpha_{t,x}$ will be assigned to the regions.

In addition, the overall proposal can be described as follow:

Algorithm 2: **Foreground Detection Scheme**

Input: video sequence,
For $t = 1, ..., L$

1. Compute $ss_{t,x}$ by algorithm 1 and $fg_{t,x}$ by GMM [6]
2. Compute $STs_{t,x}$ by (6) and (7)
3. Calculate $\alpha_{t,x}$ using (7) and (8)
4. Based new learning rate $\alpha_{t,x}$, perform GMM [6] again to get the final foreground map. At meanwhile, update $w_{t-1,x,n}$, $\mu_{t-1,x,n}$, $\sigma^2_{t-1,x,n}$ for next frame as follow:

   If $I_{t,x}$ matches the n th Gaussian model

   Use (2), (3), (4) to update the model parameters

   else

   A model replacement is performed to

incorporate $I_{t,x}$ into the GMM.

$k = \arg\min_{n=1,.....N} w_{t-1,x,n}$

$\mu_{t,x,k} = I_{t,x}, \sigma^2_{t,x,k} = \sigma^2_0, w_{t,x,k} = w_0$

$w_{t,x,n} = w_{t-1,x,n} - \alpha w_{t-1,x,n}$ *when* $n \neq k$

end

Output: the final foreground map of each frame.

4. Experiment and Result

To evaluate the performance of our proposed model, we applied it to extract moving objects on CAVIAR's datasets [10]. In the experiment, Zivkovic's model [6] was used as an anchor, and the constant learning rate α in [6] was set to 0.001. For a more comprehensive comparison, background subtraction with a higher learning rate (α = 0.01) was also compared with proposal.

Fig 2 illustrates the subject results of sequence "Walk1". Fig 2 (a) includes the 60th, 90th and 120th frame from "Walk1". Pictures in Fig 2 (b) are the results of GMM with α = 0.001. It can be seen that a lot of background noises are classified into foreground, although the man can be detected. When we increase the learning rate, the noise sensitivity is reduced significantly as shown in Fig 2 (c). But when the man stop and keep static for a while, GMM with higher learning rate may classify some parts of the man into background. However, the proposed spatio-temporal saliency can help us use different model parameters in different regions. Because spatial saliency is robust to illumination changes, the saliency values of background areas are low, while the area covering the person shows a high degree of saliency no matter the person is moving or not as shown in Fig 2 (d). So the higher learning rate α is set to the background region, while the lower learning rate is assigned to foreground. From Fig 2(e), we can see that the proposal can detect the man, and eliminate the background noise.

For quantitative analysis of our proposed method, ROC measure was applied, where true positive (TP), false positive (FP) were used. We firstly used image segmentation to split images into objects, and then marked the foregrounds manually. Let Rgt and Rd be the ground truth region and the detected region respectively. The region $Rgt \cap Rd$ is defined as TP, and the region $\overline{Rgt} \cap Rd$ is considered as FP.

So detection rate and false alarm rate can be obtained as follow:

$$DR = n(TP)/n(Rgt) \quad (9)$$

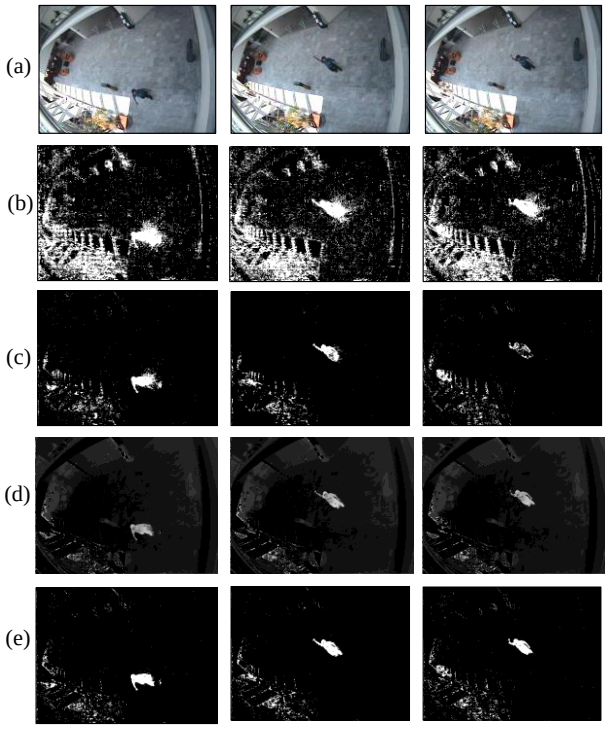
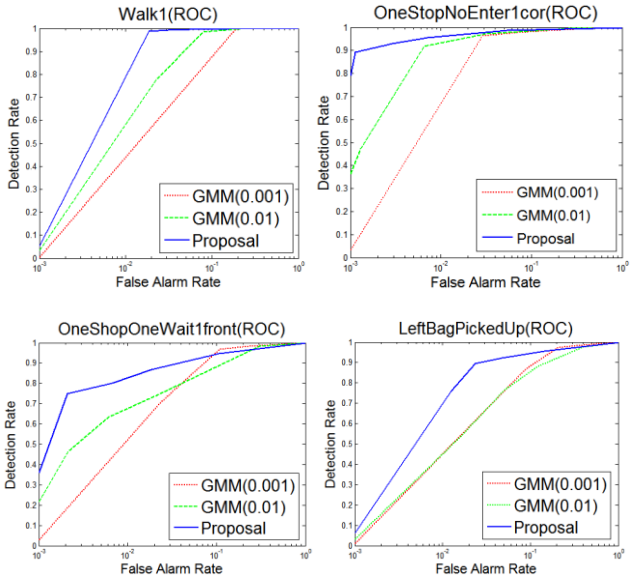


Fig 2 (a) Test frames;(b) GMM(α = 0.001); (c) GMM(α = 0.01); (d) Spatio-temporal saliency; (e) Proposal

$$FAR = n(FP)/n(\overline{Rgt}) \quad (10)$$

where $n(.)$ is an operator to count the pixel number in a region. Using detection rate and false alarm rate, we can obtain the ROC curves.



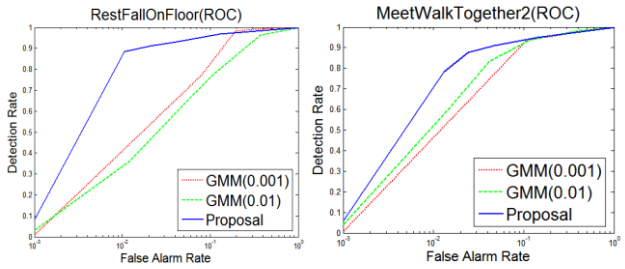


Fig3 ROC curve of six videos in CAVIAR's datasets.

The ROC curves in Fig 3 illustrates that the proposal achieves better ROC performance than anchors for six videos in the CAVIAR's datasets. Due to space limitation, we don't provide all the results here. Compared with anchors, the proposal can maintain a high detect rate while the false alarm rate decreases.

The above experimental results show that the proposed method is robust to illumination changes and movement with different moving speeds, which can achieve good balance between detecting moving objects and eliminate noises

5. Conclusion

In this paper, a more robust object detection algorithm based on spatio-temporal saliency is proposed. Spatial saliency based on Gaussian mixture model in luma space and temporal saliency obtained by background subtraction, are calculated as an auxiliary semantic knowledge to adjust pixel-wise learning rate of object detection model adaptively, which achieves good balance between detecting moving objects and eliminate noises. Experiment results show our proposal can achieve better performance than the existing background subtraction method based on GMM.

Acknowledgments

This work is supported by the major national science and technology special projects (2010ZX03004-003-03); the National Natural Science Foundation of China under Grant No. 61172173, 60970160, 61070080, 61003184, 60832002; the National Grand Fundamental Research 973 Program of China under Grant No.2009CB320906;

References

[1] C. Lakshmi Devasena, R. Revathi, M. Hemalatha,"Video Sur-veillance Systems - A Survey", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011, pp:635-642

[2] Soumaya Ghorbel, Maher Ben Jemaa and Mohamed Chtourou, "Object-based Video compression using neural networks", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011,pp:139-148

[3] C. Stauffer and W. Grimson. "Adaptive background mixture models for real-time tracking". In IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 246-252.

[4] A. Elgammal, D. Harwood, L. Davis, "Non-parametric model for background subtraction", in: Proceedings of the 6th European Conference on Computer Vision, 2000, pp. 751–767.

[5] A. Mittal, N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation", in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 302–309.

[4] Z. Zivkovic, and Ferdinand van der Heijden "Efficient adaptive density estimation per image pixel for the task of background subtraction", Pattern Recognition Letters, 2006, pp 773–780.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Machine Intell, vol. 20, no. 11, Nov. 1998, pp. 1254-1259

[8] L. Itti, N. Dhavale, F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in: Proceedings of SPIE 48th Annual International Symposium on Optical Science and Technology, vol. 5200, Aug 2003, pp. 64-78.

[9] Heng-Da Cheng , and Ying Sun, "A Hierarchical Approach to Color Image Segmentation Using Homogeneity", IEEE Trans on Image Processing, 2000, pp. 2071-2082.

[10]http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

Yang Xia received the B.S degrees from Wuhan University of technology in 2005, Wuhan, China. He is currently pursuing the Ph.D. degree in National Engineering Research Center For Multimedia Software, Wuhan University, Wuhan, China.His research interests include image/video processing, video coding and artificial intelligence.

Ruimin Hu received the B.S and M.S degrees from Nanjing.University of Posts and Telecommunications, Nanjing China,in 1984 and in 1990 respectively. and Ph.D degree in Communication and Electronic System from Huazhong University of Science and Technology, Wuhan, China in 1994. Dr. Hu is the director of National Engineering Research Center For Multimedia Software, Wuhan University and Key Laboratory of Multimedia Network Communication Engineering in Hubei province. He is Executive Chairman of the Audio Video coding Standard (AVS) workgroup of China in Audio Section. He has published two books and over 100 scientific papers. His research interests include audio/video coding and decoding, video surveillance and multimedia data processing.

Zhongyuan Wang received the B.S. degree and M.S degree in computer science from Wuhan University, Wuhan, China, in 1995 and 2001, and he received the Ph.D. degree in Communication and Information System in Wuhan University in 2008. From 2001, he worked as a Member of Research Staff in National Multimedia Software Engineering Research Center of Wuhan University. His

research interests include video compression, multimedia communications.

Tao Lu received the B.S and M.S degrees from Computer Science and Engineering Department, Wuhan institute of technology, Wuhan, China. He is currently pursuing the Ph.D. degree in National Engineering Research Center For Multimedia Software, Wuhan University, Wuhan, China. His research interests include image/video processing, computer vision and artificial intelligence.