Applied Mathematics & Information Sciences
An International Journal

A Recursive Kernel Density Learning Framework for Robust Foreground Object Segmentation

Qingsong Zhu[1,2,3,4,5], *Zhanpeng Zhang*[1,2,3,4,5,6], *Yaoqin Xie*[1,2,3,4,5*]

[1] Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen 518055, China
[2] Shenzhen Key Lab for Low-Cost Healthcare, Chinese Academy of Sciences, Shenzhen 518055, China
[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[4] Research Center for Medical Robotics, Chinese Academy of Sciences, Shenzhen 518055, China
[5] School of Medicine, Stanford University, Stanford, California, USA
[6] Sun Yat-Sen University, Guangzhou, China

Received: Oct. 13, 2011; Revised Feb. 15, 2012; Accepted Mar. 4, 2012
Published online: 1 May 2012

Abstract: Dynamic video segmentation is an important research topic in computer vision. In this paper, we present a novel recursive Kernel Density Learning framework based video segmentation method. In the algorithm, local maximum in the density functions is approximated recursively via a mean shift method firstly. Via a proposed thresholding scheme, components and parameters in the mixture Gaussian distributions can be selected adaptively, and finally converge to a relative stable background distribution mode. In the segmentation, foreground is firstly separated by simple background subtraction method. And then, the Bayes classifier is introduced to eliminate the misclassifications points to improve the segmentation quality. Experiments on a series of typical video clips are used to compare with some previous algorithms.

Keywords: Image Segmentation, Scene Modeling, Recursive Kernel Density Learning.

1. Introduction

Motion segmentation is an important processing step in many computer vision and video processing tasks, such as object tracking, video surveillance, information retrieval, and video coding. In video surveillance, detection of moving objects from a video sequence usually play a very crucial role for the success of object tracking, video segmentation, and behavior understanding. Motion detection aims at segmenting foreground regions corresponding to moving objects from the background. Although a lot of classic moving object segmentation algorithms such as [1-6] have been proposed for different applications, a popular technology in the existing video surveillance systems is background subtraction, which extracts moving objects in an image sequence captured from a static camera by comparing each coming frame with a background model. A crucial step of this technique is to obtain a stable and accurate background model. A lot of literature about video segmentation has been directed to the issue of constructing a robust and accurate background model. The usual algorithm to moving object detection is through background subtraction, which consists in maintaining an up-to-date model of the background and detecting moving objects as those that deviate from such a model.

There is no unique classification of proposed methods, existing algorithms for background modeling may be classified as a lot of categories from different viewpoints and different application layers. Some usually referred dichotomies, here cited in order to highlight advantages and tradeoffs of most existing methods, include the following. *

- ***Pixel-based vs Region-based:***

Pixel-based methods [7-9] assume that the time series of observations is independent at each pixel, while region-based methods [10-12] take advantage of interpixel relations, segmenting the images into regions or refining the low-level classification obtained at the pixel level. This step obviously increases the overall complexity.

[*] Corresponding author: e-mail: yq.xie@siat.ac.cn

- ***Unimodal-based vs Multimodal-based:***

  Basic background models assume that the intensity values of a pixel can be modeled by a single unimodal distribution [7]. Such models usually have low complexity. But cannot deal with moving backgrounds, while this is possible with multimodal models [8-9], which are at the price of higher complexity.

- ***Predictive-based vs Nonpredictive-based:***

  Predictive-based methods model the scene as a time series and develop a dynamical model to recover the current input based on past observations. The magnitude of the deviation between the predicted and actual observations can then be used as a measure of change. Predictive mechanisms of varying complexity have been considered in many literatures [13-14], which have used a Kalman filter based approach for modeling the dynamics of the state at a particular pixel. While nonpredictive-based methods neglect the order of the input observations and build a probabilistic representation of the observations at a particular pixel. there are also a number of literatures [7,15-16],[9,17-18,24] are proposed in recent years. in [7,15-16], a Single Gaussian Model (SGM) is considered to model the statistical distribution of a background pixel. In [9,17-20,24], Gaussian Mixture Model (GMM) is used to model some visual properties in traffic surveillance applications. The Expectation Maximization (EM) algorithm is usually used, which although optimal, is computationally quite expensive. GMM approach is capable of dealing with multiple hypothesis for the background and can be useful in scenes such as waving trees, escalators, rain or snow.

- ***Parametric-based vs Nonparametric-based:***

  Parametric models [7] are tightly coupled with underlying assumptions, not always perfectly corresponding to the real data, and the choice of parameters can be cumbersome, thus reducing automation. On the other hand, nonparametric models [8], [12] are more flexible but heavily dependent.

- ***Recursive-based vs Nonrecursive-based:***

  Nonrecursive-based algorithms [11, 21-22] store a buffer of a certain number of previous sequence frames and estimate the background model based on the temporal variation of each pixel within the buffer. While recursive-based approaches [7], [9] recursively update a single background model based on each new-coming frame, in [7], the background well adapts to eventual variations, but memory requirements can be significant. In [9], although space complexity is lower, the input frames from instant past can have an effect upon the current background, and, therefore, any error in the background model is carried out for a long time period.

Based on the previous research work of the academic community, some classic and representative video segmentation algorithms have been summarized in previous works and published literatures for various applications, such as Gaussian Mixture Model (GMM) [9, 24], non-parametric Kernel Density Estimation (NKDE) [25], and Sequential Kernel Density Approximation (SKDA) [29] etc. Segmentation of video has been a classical research topic in intelligent surveillance and many other computer vision domains. The aim of the proposed method in this paper is to obtain a robust background distribution by a new background modeling method called Recursive Kernel Density Learning. In the algorithm, mean shift method is used to approximate the local maximum values of the density function firstly. Via a proposed thresholding scheme, components and parameters in the mixture Gaussian distributions can be determined adaptively, and finally converge to a relative stable background distribution mode. Firstly foreground is separated by background subtraction method. And Bayes classifier is used to eliminate the misclassification image points and refine the segmentation result.

The paper is organized as follows. In section 2, we present a fairly compact overview of existing approaches adopted for background subtraction. Section 3 describes the proposed video segmentation algorithm. Experimental results on a series of real video chips and the comparison with traditional algorithms are presented in section 4. Conclusions and future research directions can be found in section 5.

2. Related work

In the classical GMM method, the model parameters for each Gaussian model are updated using an online EM to represent the background changes [24-26]. And it has been proved effective to deal with dynamic scenes like swaying trees, water waving and ambient light changes. In [24], a kernel-based function is employed to represent the color distribution of each background pixel. The kernel based distribution is a generalization of GMM which requires no assumptions to the underling distribution as well as the parameter estimation. In [27], the distribution of temporal color variations is used to model the spectral feature of the background and its update. In [28], the motion information is used to model the dynamic scenes. However, GMM based methods are usually subject to their huge computation and low convergence speed and thus makes them impractical for real-time segmentation tasks. In addition, the detection of moving objects with fast or slow speed is also unsatisfied. To overcome these disadvantages, nonparametric Kernel Density Estimation (NKDE) [25] is proposed. It utilizes the nearest historical samples and kernel density prediction to obtain background density function estimation. And then the function is used to compute probability values of the new observed samples and decide it is background or foreground. The NKDE method has the advantage of demanding no assumption of Gaussian model and flexible to dynamic variation of complicated density function. However, the NKDE method demands huge computation cost and storage space for historical samples. In recent years, a new model which calls

Sequential Kernel Density Approximation (SKDA) [29] is proposed. It utilizes mixture Gaussian Distribution whose components and parameters can vary adaptively to approximate each peak value location in the density functions. Compared with GMM and NKDE models, SKDA model can accurately represent complicated distribution functions. Moreover, the number of mixture Gaussian components can also be adjusted adaptively as well as the corresponding parameters. In addition, this model also outperforms in computation complexity and memory requirement vs. GMM and NKDE.

3. Proposed method

3.1. Modeling and update of background

Denote $p_i(i = 1, 2, \dots, m)$ to a set of means of Gaussians in R^l and C_i refers to a symmetric positive definite $l \times l$ covariance matrix which is associated with corresponding Gaussian functions. Each Gaussian function is associated with a weighted ω_i and $\sum_{i=1}^{m} \omega_i = 1$. The probability density function of each pixel point $\mathbf{p}$ is given by:

$$\mathbf{g}_t(\mathbf{p}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \|\mathbf{C}_i\|^{-1/2} \omega_i exp\Big(-\frac{1}{2}[M(\mathbf{p}, \mathbf{p}_i, \mathbf{C}_i)]^2\Big) \tag{1}$$

where

$$M(\mathbf{p}, \mathbf{p}_i, \mathbf{C}_i) = \sqrt{(\mathbf{p} - \mathbf{p}_i)^T C_i^{-1} (\mathbf{p} - \mathbf{p}_i)} \tag{2}$$

indicates the Mahalanobis distance from $\mathbf{p}$ to $\mathbf{p}_i$. Probability density at $\mathbf{p}$ can be obtained as the sum of the average of weighted mixture Gaussian densities, which are centered at $\mathbf{p}_i$ and having the common covariance matrix $\mathbf{C}_i$.

Suppose the initial background Gaussian distribution has m Gaussian components. In order to find all $n(n \ll m)$ local maximum values in the distribution to be estimated, the classical variable-bandwidth mean shift algorithm is introduced as:

$$msv(\mathbf{p}) = \Big(\sum_{i=1}^{m} \zeta_i(\mathbf{p}) \mathbf{C}_i^{-1}(\mathbf{p})\Big)^{-1} \Big(\sum_{i=1}^{m} \zeta_i(\mathbf{p}) \mathbf{C}_i^{-1}(\mathbf{p}) \mathbf{p}_i\Big) - \mathbf{p} \tag{3}$$

where $\mathbf{C}_i^{-1}(\mathbf{p}) = \sum_{i=1}^{m} \omega_i(\mathbf{p}) \mathbf{C}^{-1}$.

$$\zeta_i(\mathbf{p}) = \frac{\omega_i \cdot \|\mathbf{C}_i\|^{-1/2} \cdot exp\left(-\frac{1}{2} D^2(\mathbf{p}, \mathbf{p}_i, C_i)\right)}{\sum_{i=1}^{m} \omega_i \cdot \|\mathbf{C}_i\|^{-1/2} \cdot exp\left(-\frac{1}{2} D^2(\mathbf{p}, \mathbf{p}_i, C_i)\right)} \tag{4}$$

$$\mathbf{p} = \mathbf{p} + msv(\mathbf{p}) \tag{5}$$

And $\zeta_i(\mathbf{p})$ satisfies $\sum_{i=1}^{m} \zeta_i(\mathbf{p}) = 1$. Hessian matrix is used to as the stop function as:

$$\mathbf{H}(\mathbf{p}) = (\nabla^T \nabla) \hat{g}_t(\mathbf{p}) = C_i^{-1} \left((\mathbf{p}_i - \mathbf{p})(\mathbf{p}_i - \mathbf{p})^T - \mathbf{C}_i\right) C_i^{-1} \tilde{g}_t(\mathbf{p}) \times \frac{1}{(2\pi)^{d/2}} \sum_{i=1}^{m} \|\mathbf{C}_i\|^{-1/2} \omega_i exp\left(-\frac{1}{2} M^2(\mathbf{p}, \mathbf{p}_i, \mathbf{C}_i)\right) \tag{6}$$

The estimated covariance matrix is given by:

$$\hat{C}_i = \frac{\hat{\omega}_i^{2/(d+2)}}{|2\pi(-\mathbf{H}_i^{-1}(\hat{p}_i))|^{1/(d+2)}} \mathbf{H}_i^{-1}(\hat{p}_i) \tag{7}$$

Repeating (3)-(5) until background initial mode converged to the only stable point in Equation (1). Now, it must be determined which other modes converge to S^{left} and should be merged with $\mathbf{p}_{t+1}^{new}$ The candidates that converge to S^{left} are determined by mean-shift algorithm. And this procedure is repeated until no additional candidate converges to S^{left}. The first candidate mode is the convergence point S_{middle} of $\mathbf{p}_{t+1}^{new}$ in the density function:

$$\mathbf{g}_{t+1}^{N} \leftarrow \mathbf{g}_{t+1}(\mathbf{p}) - N(\omega_i^*, S_{middle}, C_{t+1}^{new}) \tag{8}$$

Note that all the candidates are one of the components in previous density function $\hat{g}_t(\mathbf{p})$. The Mean-Shift Search Algorithm ($MSSA$) is performed for S_{middle} in $\mathbf{g}_{t+1}^{new}$ and for S^{right} in $\mathbf{g}_{t+1}(\mathbf{p})$:

$$S_{middle} \leftarrow MSSA\left[\mathbf{g}_{t+1}^{new}(\mathbf{p}), \mathbf{p}_{t+1}^{new}\right] \tag{9}$$

$$S^{right} \leftarrow MSSA\left[\mathbf{g}_{t+1}(\mathbf{p}), S_{middle}\right] \tag{10}$$

If the convergence point of S_{middle} and S^{left} are not equal, we can draw a conclusion that there are no further mergence with $\mathbf{p}_{t+1}^{new}$ and create a Gaussian for the merged mode. Otherwise, the next candidate can be determined by finding the next convergence point of $\mathbf{p}_{t+1}^{new}$ in the density function:

$$\mathbf{g}_{t+1}^{N} \leftarrow \mathbf{g}_{t+1}(\mathbf{p}) - N(\omega_i^*, S_{middle}, C_{t+1}^{new}) \tag{11}$$

The covariance matrix and the weight of the merged mode should also be updated accordingly. If this condition is satisfied, all the n sample points $(n \ll m)$ which converge to that location should be approximated with a single Gaussian function $N(\omega_k, \mu_k, \varSigma_k)$ centered at the convergence location, where μ_k is a local maximum and $\varSigma_k$ is obtained by the curvature in the location which approximates the peak value. The weight ω_k of each Gaussian is equal to the sum of the kernel weights of the data points that converge to the maxima of background initial mode. Suppose background probability density distribution model consisted with m Gaussian distributions $N(\omega_k, \mu_k, \varSigma_k)_{k=1}^{m}$ at $\mathbf{p}_k(k = 1, 2, \dots, m)$. Start from the second frame, the new frame is used to update the background distribution model. When the new sample $\mathbf{p}_{t+1}^{new}$ is available, probability density function of the sample point computed at image point $\mathbf{p}_{t+1}^{new}$ can be changed to:

$$\mathbf{g}_{t+1}(\mathbf{p}) = \frac{\alpha}{(2\pi)^{d/2}} \times \sum_{i=1}^{m} \|\mathbf{C}_t^i\|^{-1/2}\omega_t^i \times exp\left(-\frac{1}{2}\left[M_t(\mathbf{p}_t, \mathbf{p}_t^i, \mathbf{C}_t^i)\right]^2\right) + \frac{(1-\alpha)}{(2\pi)^{d/2}} \times \sum_{i=1}^{m} \|\mathbf{C}_{t+1}^i\|^{-1/2}\omega_{t+1}^i \times exp\left(-\frac{1}{2}\left[M_{t+1}(\mathbf{p}_{t+1}, \mathbf{p}_{t+1}^i, \mathbf{C}_{t+1}^i)\right]^2\right) \quad (12)$$

where

$$M_t(\mathbf{p}_t, \mathbf{p}_t^i, \mathbf{C}_t^i) = \sqrt{(\mathbf{p}-\mathbf{p}_i)^T C_i^{-1}(\mathbf{p}-\mathbf{p}_i)} \quad (13)$$

$$M_{t+1}(\mathbf{p}_{t+1}, \mathbf{p}_{t+1}^i, \mathbf{C}_{t+1}^i) = \sqrt{(\mathbf{p}-\mathbf{p}_{t+1}^{new})^T (C^{-1})_{t+1}^{new}(\mathbf{p}-\mathbf{p}_{t+1}^{new})} \quad (14)$$

If the new sample successfully matches with the jth distribution of the m background density distribution, then we can merge the new sample into the jth distribution. The matching criterion is described as:

$$|p_{t+1}^{new} - \mu_{j,t}| \le \varepsilon \cdot \sigma_{j,t} \quad (15)$$

where ε is a decision factor. And then this distribution will perform corresponding update including mean, variance and weight. The rest of background density distributions remain unchanged. If the matching fails, a new Gaussian distribution $N(\varepsilon_{m+1}, \mu_{m+1}, \Sigma_{m+1})$ is produced. The flow chart of the background modeling and update algorithm for individual pixel is as shown in Fig. 1.

3.2. Segmentation and refinement of foreground

Foreground can be obtained by subtracting the segmented background firstly. However there still exist a lot of misclassification points. In order to improve the final segmentation quality, Bayes classifier [9] is introduced in this work. It has been proved that, in natural scenes, a pixel is more likely to be foreground if most of its neighboring pixels belong to foreground, and vice versa. Consider 8-neighbors around image point $\mathbf{X} = (x, y)$ and a n-dimensional feature vector $\mathbf{v}_t$ extracted from the position $\mathbf{X}$ at time t from the image sequence, prior probability of $\mathbf{X}$ can be formulated as:

$$P_{\mathbf{X}}(fg) = \frac{1}{N} \cdot exp\{-\left[\phi_{nv}B + \phi_{dg}C\right]\} \quad (16)$$

where N is a normalization term, ϕ_{nv} and ϕ_{dg} indicate the number of background pixels in the horizontal/vertical and diagonal neighborhood respectively, B and C are the corresponding penalty coefficients. With $P_{\mathbf{X}}(fg)$ and the foreground likelihood $P_{\mathbf{X}}(\mathbf{v}_t|fg)$, using the Bayes classifier theory, pixel $\mathbf{X}$ will be classified as foreground if :

$$2P_{\mathbf{X}}(\mathbf{v}_t|fg)P_{\mathbf{X}}(fg) > T \quad (17)$$

where T is a fixed threshold. Otherwise $\mathbf{X}$ is classified as background. Let T_a denote $T/P_{\mathbf{X}}(fg)$, then the inequality (17) can be rewritten as:

$$2P_{\mathbf{X}}(\mathbf{v}_t|fg) > T_a \quad (18)$$

Then T_a becomes an adaptive threshold determined by T and the prior probability $P_{\mathbf{X}}(fg)$. It can be further expressed as the following equation:

$$T_a = T \times \left\lceil P_{\mathbf{X}}(fg) + \left(1 - \underset{0\le\phi_{hv}\le4, 0\le\phi_{dg}\le4}{median}\left[P_{\mathbf{X}}(fg)\right]\right)\right\rceil^{-\varepsilon} \quad (19)$$

where ε is a balance factor. Note that, in order to measure the number of ϕ_{nv} and ϕ_{dg}, the four neighbors in the four corners of the 8-neighbors are determined using the fixed threshold T, while the remaining neighboring pixels are determined using their adaptive threshold T_a.

4. Experiment results

The algorithm starts with following initial parameters: We conduct a series of experiments on two typical video clips: Waving Trees (160*120) and Walking Man (384*288). The two test sequences all contain some changing background like swaying trees and much worse lighting conditions etc. The system is running on a P4-2GHz desktop with 1GB RAM. The algorithm is also compared with GMM, NKDE and SKDE and the results are as shown in Fig. 2.

In Fig. 2, the first column is the original video frames, the 2nd column shows the background modeling result via our proposed method, the 3rd, 4th, 5th and 6th columns displayed the foreground segmentation results by GMM, NDKE, SKDE and Our proposed method respectively. In comparison, the proposed method outperforms in dynamic scenes (swaying trees and much worse lighting conditions) and also gives better segmentation results. The computation time for Waving Trees video sequences is 0.48, 0.52, 0.36 and 0.29 s/frame, respectively; for Walking Man video sequences it is 0.32, 0.37, 0.21 and 0.15 s/frame. It is clear that the computation time of our proposed method is more efficient. In addition, the convergence speed of background modeling changes with respect to the frame is provided in Fig. 3. We can see that the extracted background by our method converges quickly to a constant (around the 280th, 310th frame for Waving Trees and Walking Man, respectively) compared with GMM, NKDE and SKDE.

5. Conclusion and future work

This paper presents a novel recursive kernel density learning framework for dynamic video segmentation. Mean shift

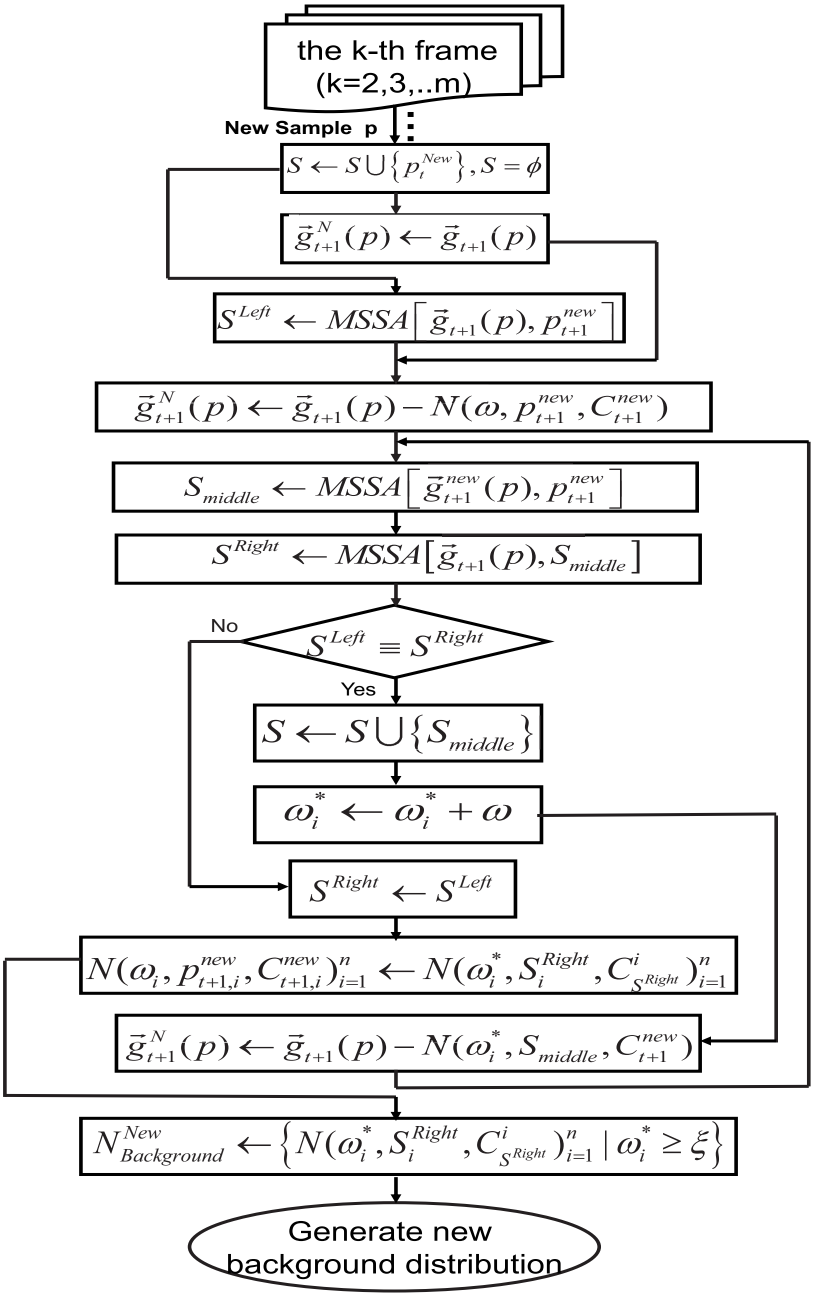
the k-th frame (k=2,3,..m)
New Sample p
$S \leftarrow S \cup \{p_t^{New}\}, S = \phi$
$\vec{g}_{t+1}^{N}(p) \leftarrow \vec{g}_{t+1}(p)$
$S^{Left} \leftarrow MSSA[\vec{g}_{t+1}(p), p_{t+1}^{new}]$
$\vec{g}_{t+1}^{N}(p) \leftarrow \vec{g}_{t+1}(p) - N(\omega, p_{t+1}^{new}, C_{t+1}^{new})$
$S_{middle} \leftarrow MSSA[\vec{g}_{t+1}^{new}(p), p_{t+1}^{new}]$
$S^{Right} \leftarrow MSSA[\vec{g}_{t+1}(p), S_{middle}]$
No
$S^{Left} \equiv S^{Right}$
Yes
$S \leftarrow S \cup \{S_{middle}\}$
$\omega_i^{*} \leftarrow \omega_i^{*} + \omega$
$S^{Right} \leftarrow S^{Left}$
$N(\omega_i, p_{t+1,i}^{new}, C_{t+1,i}^{new})_{i=1}^{n} \leftarrow N(\omega_i^{*}, S_i^{Right}, C_{S^{Right}}^{i})_{i=1}^{n}$
$\vec{g}_{t+1}^{N}(p) \leftarrow \vec{g}_{t+1}(p) - N(\omega_i^{*}, S_{middle}, C_{t+1}^{new})$
$N_{Background}^{New} \leftarrow \{N(\omega_i^{*}, S_i^{Right}, C_{S^{Right}}^{i})_{i=1}^{n} \mid \omega_i^{*} \ge \xi\}$
Generate new background distribution

Figure 1 Workflow of the background modeling algorithm.

method is used to approximate the peak values of the density function recursively. Components and parameters of mixture Gaussian distribution are adaptively selected via a proposed scheme and finally converge to a relative stable background distribution. In the segmentation, firstly foreground is separated by simple background subtraction method. And then, Bayes classifier is proposed to eliminate the misclassification points to refine the segmentation result. Experiments with two typical video clips are used to demonstrate that the proposed method outperforms previous methods like GMM, NKDE and SKDE in both segmentation result and converging speed. Future work can address how to deal with more challenging scenarios and how to improve algorithm converging and the system running speed further.

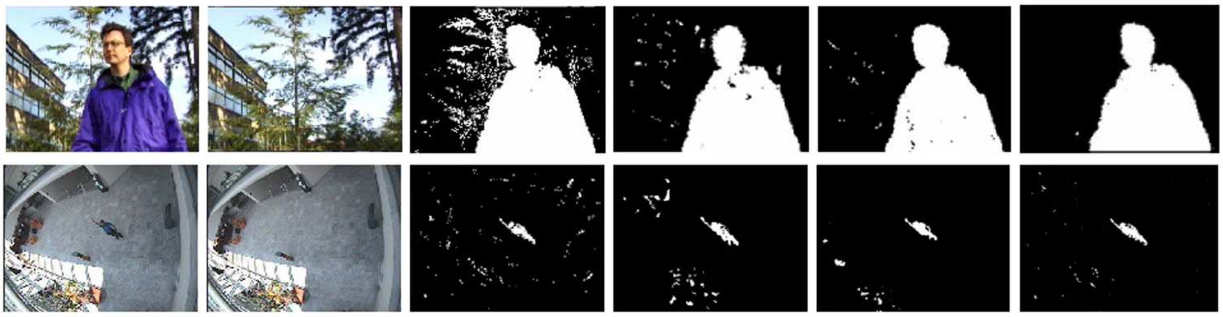

Figure 2 Comparison of GMM, NKDE, SKDE and Our proposed method.

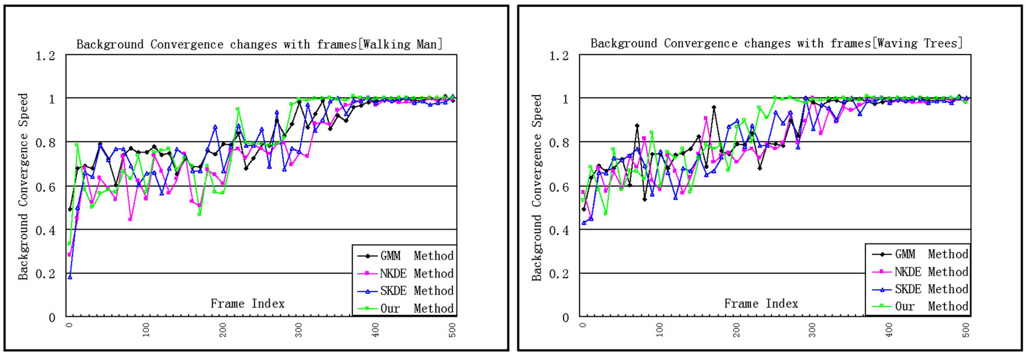

Figure 3 Background stability comparison between GMM, NKDE, SKDE and Our proposed method.

Acknowledgment

This study has been financed partially by the Projects of National Natural Science Foundation of China (Grant No. 50635030, 60932001, 61072031, 61002040), the National Basic Research (973) Program of China (Sub-grant 6 of Grant No. 2010CB732606) and the Knowledge Innovation Program of the Chinese Academy of Sciences, and was also supported by the China Scholarship Council (CSC) and China Postdoctoral Project.

References

[1] T. Meier and K. N. Ngan, IEEE Transactions on Circuits System and Video Technology (CSVT) **8**, 525(1998).

[2] D. Wang, IEEE Transactions on circuits System and Video Technology (CSVT) **8**, 539(1998).

[3] P. Salembier and F. Marques, IEEE Transactions on Circuits System and Video Technology (CSVT) **9**, 1147(1999).

[4] Y. N. Deng and B. S. Manjunath, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **23**, 800(2001).

[5] I. Patras, E. A. Hendriks and R. L. Lagendijk, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **23**, 326(2001).

[6] S. Y. Chien, S. Y. Ma and L. G. Chen, IEEE Transactions on Circuits System and Video Technology (CSVT) **12**, 577(2002).

[7] C. Wren, A. Azarbayejani, T. Darrell and A. Pentland, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **19**, 780(1997).

[8] K. Kim, T. H. Chalidabhongse, D. Harwood and L. S. Davis, Real-Time Imaging **11**, 172(2005).

[9] C. Stauffer and W. E. L. Grimson, Proc. International Conference on Computer Vision and Pattern Recognition (CVPR), 246(1999).

[10] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomono, O. Hasegawa, P. Burt and L. Wixson, A system for video surveillance and monitoring (Carnegie Mellon University, Pittsburgh, 2000).

[11] K. Toyama, J. Krumm, B. Brumitt and B. Meyers, Proc. International Conference on Computer Vision (ICCV), 255(1999).

[12] A. Elgammal, D. Hanvood, and L. S. Davis, Proc. European Conference on Computer Vision (ECCV), 751(2000).

[13] K. P. Karmann and A. V. Brandt, Time Varying Image Processing and Moving Object Recognition **2**, 278(1990).

[14] D. Koller, J. Weber and J. Malik, Proc. European Conference on Computer Vision (ECCV), 189(1994).

[15] C. Eveland, K. Konolige and R. C. Bolles, Proc. International conference on Computer Vision and Pattern Recognition (CVPR), 266(1998).
[16] A. Cavallaro and T. Ebrahimi, Proc. SPIE Visual Communications and Image Processing (VCIP), 465(2001).
[17] P. Kaewtrakulpong and R. Bowden, Proc. European Workshop on Advanced Video Based Surveillance Systems, 149(2001).
[18] D. S. Lee, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **27**, 827(2005).
[19] L. Li, W. Huang, I. Y. H. Gu and Q. Tian, IEEE Transactions on Image Proceeing (IP) **13**, 1459(2004).
[20] M.Harville, Proc. European Conference on Computer Vision (ECCV), 543(2002).
[21] R. Cucchiara, M. Piccardi and A. Prati, IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI) **25**, 1(2003).
[22] B. P. L. Lo and S. A. Velastin, Proc. International Symposium on Intelligent Multimedia, Video, and Speech Processing (ISIMP), 158(2001).
[23] M. King, B. Zhu and S. Tang, **8**, 520(2001).
[24] C. Stauffer and W. Grimson, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **22**, 747(2000).
[25] A. Elgammal, R. Duraiswami, D. Harwood and L. S. Davis, Proceedings of IEEE **90**, 1151(2002).
[26] Z. Zivkovic and F. Heijden, Pattern Recognition Letters **27**, 773(2006).
[27] D. S. Lee, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **27**, 827(2005).
[28] I. Haritaoglu, D. Harwood and L. Davis, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **22**, 809(2000).
[29] A. Mittal and N. paragios, Proc. International Conference on Computer Vision and Pattern Recognition (CVPR), 302(2004).
[30] B. Han, D. Comaniciu, Y. Zhu, and L. S. Davis, IEEE Transaction Pattern Analysis and Machine Intelligence (PAMI) **30**, 1186(2008).
[31] A. E. Brockwell, Recursive Kernel Density Estimation of the Likelihood for Generalized State-Space Models (CMU Statistics Dept, 2005).
[32] T. Aach and A. Kaup, IEEE Transaction on Image Processing **7**, 147(1995).

Qingsong Zhu received the BS and MS degrees in computer science from University of Science and Technology of China (USTC), Hefei, China. He is currently an assistant professor in Research Centre for Medical Robotics and Minimally Invasive Surgical Devices, Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His current research interests focus on computer vision, statistical pattern recognition, machine learning and robotics. He is a member of the IEEE.

Zhanpeng Zhang received the BS degree in 2010 from Sun Yat-Sen University, Guangzhou, China. He is currently a postgraduate student in School of Information Science and Technology, Sun Yat-Sen University. He is also working as a visiting student in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include computer vision and machine learning.

Yaoqin Xie Associate Professor, Overseas High-Caliber Personnel in Shenzhen, China. He received his B.Eng, M.Eng, and Ph.D. from Tsinghua University of China in 1995, 1998, and 2002, respectively, and won the excellent doctor degree dissertation of Tsinghua University. He joined Stanford University as a postdoctoral fellow from 2006 to 2008. Dr. Xie's research areas have been focused on image-guided surgery. He has published more than 90 papers. Dr. Xie was working in Peking University of China from 2002, and joined Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, on 2010. He is a committee member of Chinese Society of Molecular Imaging, committee member of Chinese Society of Medical Imaging Physics, editorial board member of Chinese Medical Imaging Technology, member of American Association of Physicist in Medicine (AAPM), Chinese Society of Biomedical Engineering. He is also the referee of IEEE Transactions on Information Technology in BioMedicine.