RECURSIVE KERNEL DENSITY ESTIMATION FOR MODELING THE BACKGROUND AND SEGMENTING MOVING OBJECTS

Qingsong Zhu[1,3,4]*, *Ling Shao*[2], *Qi Li*[1,5], *Yaoqin Xie*[1,3,4]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, UK
[3]Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen, China
[4]School of Medicine, Stanford University, Stanford, California, USA
[5]University of Science and Technology of China, Hefei, China

ABSTRACT

Identifying moving objects in a video sequence is a fundamental and critical task in video surveillance, traffic monitoring, and gesture recognition in human-machine interface. In this paper, we present a novel recursive Kernel Density Estimation based background modeling method. First, local maximum in the density functions is recursively approximated using a mean shift method. Second, components and parameters in the mixture Gaussian distributions can be selected adaptively via a proposed thresholding mechanism, and finally converge to a stable background distribution model. In the scene segmentation, foreground is firstly separated by simple background subtraction approach. And then a local texture correlation operator is introduced to fill the vacancies and remove the fractional false foreground regions so as to obtain a better video segmentation quality. Experiments conducted on synthetic and video data demonstrate the superior performance of the proposed algorithms.★

Index Terms— video segmentation, background modeling, Recursive Kernel Density Estimation

1. INTRODUCTION

Visual surveillance and video segmentation is a very active research area in computer vision, intelligent surveillance and many other computer vision domains. A lot of segmentation algorithms have been proposed in previous works for various applications, such as Single Gaussian Model (SGM) [1], Gaussian Mixture Model (GMM) [2-3], non-parametric Kernel Density Estimation [4], and Sequential Kernel Density Approximation etc.

Many of the existing approaches compromise on the accuracy of the system, in favour of achieving fast processing speeds. In the classical GMM method, the model parameters for each Gaussian model are updated by using an online Expectation Maximization (EM) to represent the background changes [4-6]. And it has been proved effective to deal with dynamic scenes like swaying trees, water waving and ambient light changes.

However, GMM based methods are usually subject to its huge computation and low converging velocity and thus makes it impractical for real-time segmentation tasks. In addition, the detection of moving objects with fast or slowly speed is also unsatisfied. To overcome these disadvantages, non-parametric Kernel Density Estimation [4] 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.

In [7], Stauffer et al. have used a fast online k-means based approximation to update the parameters of the GMM (GMM-OK). While the approach is very effective when the contrast between foreground and background is high, it yields poor results when the contrast is low [8]. In [8], Singh et al. use a combination of the EM algorithm and the online k-means approximation (GMM-OKEM) for parameters update of the mixture model so as to obtain appreciable results in low contrast conditions. In [9], the distribution of temporal color variations is used to model the spectral feature of the background and its update. In [10], the motion information is used to model the dynamic scenes using adaptive Kernel Density Estimation.

Recently, a novel model which calls Sequential Kernel Density Approximation [11] is proposed. It utilizes Mixture Gaussian Distribution whose components and parameters can vary adaptively to approximate each peak value location in the density functions. The model can accurately represent complicated distribution functions. Moreover, the number of Mixture Gaussian component can also be adjusted adaptively as well as the corresponding parameters.

★**Acknowledgements:** This work was supported in part by the NSFC (61002040, 60903115, 50635030, 60932001), the NSFC Guangdong (10171782619-2000007), the grants of Introduced Innovative R&D Team of Guangdong Province：Image-Guided TherapyTechnology. **Corresponding Author:** qs.zhu@siat.ac.cn

In this paper, we propose a Recursive Kernel Density Estimation (r-KDE) [12] based video segmentation method. In the algorithm, mean shift method is used to approximate the local maximum values of the density function firstly. Using a proposed thresholding mechanism, components and parameters in the mixture Gaussian distributions can be determined adaptively, and finally converge to a relative stable background distribution model. In the Foreground detection, background subtraction is introduced to obtain a initial segmentation result, And then a local texture correlation operator is introduced to fill the vacancies and remove the fractional false foreground regions so as to refine the segmentation result.

2. PROPOSED METHOD

2.1. Background modeling

Denote $p_i(i=1,2,...m)$ 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 image point $\vec{p}$ is given by:

$$\vec{g}_t(\vec{p}) = \frac{1}{m} \cdot \frac{1}{(2\pi)^{d/2}} \sum_{i=1}^{m} \left\| \vec{\mathbf{C}}_i \right\|^{-1/2} \omega_i \exp\left(-\frac{1}{2} \left[M(\vec{p}, \vec{p}_i, \vec{\mathbf{C}}_i) \right]^2 \right) \quad (1)$$

where $M(\vec{p}, \vec{p}_i, \vec{\mathbf{C}}_i) = \sqrt{(\vec{p} - \vec{p}_i)^T C_i^{-1} (\vec{p} - \vec{p}_i)}$ indicates the Mahalanobis distance from $\vec{p}$ to $\vec{p}_i$. Probability density at $\vec{p}$ can be obtained as the sum of the average of weighted Mixture Gaussian densities, which are centered at $\vec{p}_i$ and having the common covariance matrix $\vec{\mathbf{C}}_i$.

Supposed the initial background Gaussian distribution have *m* Gaussian components. In order to find all *n* (*n*<<*m*) local maximum values in the distribution to be estimated, the classical variable-bandwidth mean shift algorithm is introduced as:

$$msv(\vec{p}) = \left(\sum_{i=1}^{m} \zeta_i(\vec{p}) \vec{\mathbf{C}}_i^{-1}(\vec{p}) \right)^{-1} \left(\sum_{i=1}^{m} \zeta_i(\vec{p}) \vec{\mathbf{C}}_i^{-1}(\vec{p}) \vec{p}_i \right) - \vec{p} \quad (2)$$

where

$$\vec{\mathbf{C}}_i^{-1}(\vec{p}) = \sum_{i=1}^{m} \omega_i(\vec{p}) \vec{\mathbf{C}}^{-1} \quad (3)$$

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

$$\vec{p} = \vec{p} + msv(\vec{p}) \quad (5)$$

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

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

The estimated covariance matrix is given by:

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

Repeating (2)-(4) 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 $\vec{p}_{t+1}^{new}$. 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 $\vec{p}_{t+1}^{new}$ in the density function:

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

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

$$S_{Middle} \leftarrow MSSA[\vec{g}_{t+1}^{new}(\vec{p}), \vec{p}_{t+1}^{new}] \quad (9)$$

$$S^{Right} \leftarrow MASA[\vec{g}_{t+1}(\vec{p}), S_{Middle}] \quad (10)$$

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 $\vec{p}_{t+1}^{new}$ and create a Gaussian for the merged mode. Otherwise, the next candidate can be determined by finding the next convergence point of $\vec{p}_{t+1}^{new}$ in the density function:

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

The covariance matrix and the weight of the merged mode should be also updated accordingly. If this condition is satisfied, all the *n* sample points (*n*<<*m*) which converge to that location should be approximated with a single Gaussian function $N(\omega_k, \mu_k, \sum_k)$ centered at the convergence location, where μ_k is a local maximum and $\sum_k$ 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 $N(\omega_k, \mu_k, \sum_k)_{k=1}^m$ at $\vec{p}_k, k=1,2,\cdots m$. Start from the second frame, the new frame is used to update the background distribution model. When the new sample $\vec{p}_{t+1}^{new}$ is available, probability density function of the sample point computed at image point $\vec{p}_{t+1}^{new}$ can be changed to:

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

where

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

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

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:

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

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 $N(\omega_{m+1}, \mu_{m+1}, \sum_{m+1})$ is produced.

2.2. Modeling the Foreground and refinement

Foreground can be obtained by subtracting the segmented background. While the background is obtained, primary foreground can be extracted by simple subtraction. However the vacancy phenomenon and fractional false foreground regions become seriously while the foreground and background are with similar color or gray levels. To improve the foreground quality, the post-processing procedure is usually implemented. Morphological operators such as "open," "close," and some specified template filters are often used to remove segmentation noise and fill the vacancies in the foreground image. In [14], an objective reconstruction function based on mathematical morphology is introduced as

$$F_g = M_{\text{sec}} \cap (M_{org} \oplus SE) \quad (16)$$

where F_g is the final refined segmentation result, M_{org} is the original binary image formed by segmentation methods such as GMM, GMM-OK, GMM-OKEM, etc., M_{sec} is a new binary image formed by a 3×3 morphological "open" operation on M_{org}, and SE is a structure element with size of 7×7 pixels. However, such post-processing algorithms are incorporated with the obtained foreground image only, but do not consider the information from the original images. In this paper, to improve the foreground segmentation quality, a local texture correlation method is introduced. It has been proved that pixels in a relative small image region between background and foreground have local texture correlation property [15]. Pixels in each neighborhood are reclassified according to the above cross-correlations to compensate small holes within foreground and remove the fractional false foreground regions. Assume I be the intensity at image point (x, y) in the t-frame, i.e., $\hbar(x, y) = \hbar(x, y \,|\, (t, I))$. The gradient vector at (x, y) can be expressed as

$$\operatorname{grad}\hbar(x, y \,|\, (t, I)) = \nabla_x \hbar(x, y \,|\, (t, I))\vec{i} + \nabla_y \hbar(x, y \,|\, (t, I))\vec{j} \quad (17)$$

Where $\vec{i}$ and $\vec{j}$ indicate the unit vector of the positive direction along $x-$ and $y-axis$, respectively. Given two adjacent pixels I_a, I_b in a small neighborhood of I, then gradient vector similarity for I_a and I_b can be expressed as

$$\text{Similar}(I_a, I_b) = \frac{\operatorname{grad}\hbar(x_a, y_a \,|\, (t, I_a)) \bullet \operatorname{grad}\hbar(x_a, y_a \,|\, (t, I_a))}{\left\|\operatorname{grad}\hbar(x_a, y_a \,|\, (t, I_a))\right\| \cdot \left\|\operatorname{grad}\hbar(x_a, y_a \,|\, (t, I_a))\right\|} \quad (18)$$

3. EXPERIMENTAL RESULTS

The algorithm starts with following initial parameters: $\xi = 0.05^6$. We conduct a series of experiments on two typical video clips: *Shopping Mall* (320*240) and *Lobby* (320*240).

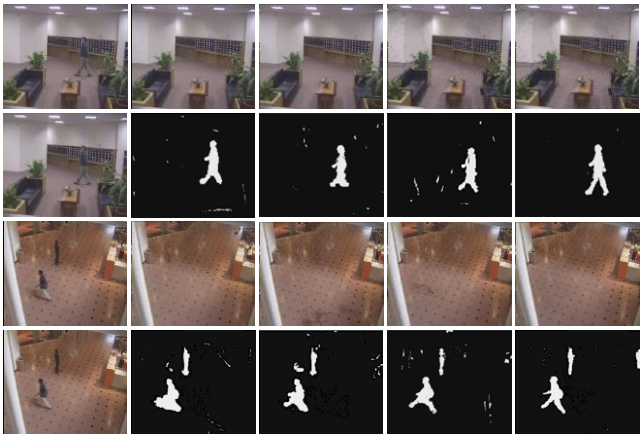

Fig. 2. The first (third) row, from left to right, shows original frame, background modeling results of GMM, GMM-OK, GMM-OKEM and r-KDE. The second (fourth) row, from left to right, shows original frame, foreground segmentation results of GMM, GMM-OK, GMM-OKEM and r-KDE.

The two test sequences all contain some changing background like the change of light intensity, the flashing lights etc. The system is running on a P4-2GHz desktop

with 1GB RAM. The algorithm is also compared with GMM, GMM-OK, GMM-OKEM. and the results are as shown in Fig. 2. In comparison, the proposed method outperforms in dynamic scenes (the change of light intensity, the flashing lights) and also gives better segmentation results. The computation time for video sequences *Shopping Mall* and *Lobby* are respectively shown in Table 1. It is clear that the computation time of our method is more efficient.

Table 1. Average execution time comparison of GMM, GMM-OK, GMM-OKEM and our proposed method on different datasets (in ms/frame).

Test video sequence	GMM [2]	GMM-OK[7]	GMM-OKEM[8]	**r-KDE**
Shopping Mall (320*240)	48.91	45.42	79.78	**31.86**
Lobby (320*240)	43.61	32.37	67.65	**20.98**

4. CONCLUSION AND FUTURE WORK

This paper presents a novel recursive Kernel Density Estimation method for dynamic video segmentation. Mean shift 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, foreground is separated by simple background subtraction method firstly. And then, Bayes classifier is proposed to eliminate the misclassification points to refine the segmentation result. Experiments with four typical video clips are used to demonstrate that the proposed method outperforms previous methods like 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.

5. REFERENCES

[1] C. Wren, A. Azarbayejani, T. Darrell and A. Pentland. Pfinder: "Real-time tacking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, vol. 19(1), pp.780-785, 1997.

[2] C. Stauffer and W. Grimson, "Learning Patterns of Activity using Real-time Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, vol. 22(8), pp. 747-757, 2000.

[3] Q. Zhu and Z. Song, "Dynamic video segmentation via a novel recursive Bayesian learning method," *IEEE International Conference on Image Processing(ICIP), HongKong*, pp. 2997-3000, 2010.

[4] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, "Background and Foreground Modeling using Nonparametric Kernel Density Estimation for Visual Surveillance," *Proceedings of IEEE*, vol. 90(7), pp. 1151-1163, 2002.

[5] Z. Zivkovic and F. Heijden, "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction," *Pattern Recognition Letters*, vol. 27(7), pp. 773-780, 2006.

[6] D.S. Lee, "Effective Gaussian Mixture Learning for Video Background Subtraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, vol. 27(5), pp. 827-832, 2005.

[7] C. Stauffer and W. Grimson, "Adaptive background mixture models for real time tracking," *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), Colorado, USA,* pp.599-608, 1999.

[8] A Singh, P Jaikumar, S.K. Mitra, M.V. Joshi and A. Banerjee. "Detection and Tracking of Objects in low contrast conditions," *In IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics. Gandhinagar, India,* pp.98-103, 2008.

[9] I. Haritaoglu, D. Harwood, and L. Davis, "W^4 :Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, vol. 22, pp. 809-830, 2000.

[10] A. Mittal and N. paragios, "Motion-based Background Subtraction using Adaptive Kernel Density Estimation," *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)* , vol .2, pp. 302-309, 2004.

[11] B. Han, D. Comaniciu, Y. Zhu, and L.S. Davis, "Sequential Kernel Density Approximation and Its Applications to Real-Time Visual Tracking," *IEEE Transaction Pattern Analysis and Machine Intelligence(PAMI)*, vol. 30(7), pp. 1186-1197, 2008.

[12] A.E. Brockwell, "Recursive Kernel Density Estimation of the Likelihood for Generalized State-Space Models," CMU Statistics Dept. *Tech. Report* , #816.

[13] T. Aach and A. Kaup, "Bayesian Algorithms For Adaptive Change Detection In Image Sequences Using Markov Random Fields," *IEEE Transaction on Image Processing(TIP)*, vol. 7(2), pp. 147-160, 1995.

[14] L. Xu and J. landabaso, "Segmentation and tracking of Multiple moving objects for intelligent video analysis," *BT Technol. J.,* vol. 22(3), pp. 140-149, 2004.

[15] J. Rittscher, J. Kato, S. Joga and A. Blake, "A probabilistic background model for tracking," *IEEE International Conference on Computer Vision(ICCV)* , vol, 2(1), pp. 336-350, 2000.