ROBUST VIDEO OBJECT SEGMENTATION BASED ON K-MEANS BACKGROUND CLUSTERING AND WATERSHED IN ILL-CONDITIONED SURVEILLANCE SYSTEMS

Tse-Wei Chen, Shou-Chieh Hsu, and Shao-Yi Chien

Media IC and System Lab
Graduate Institute of Electronics Engineering and Department of Electrical Engineering
National Taiwan University
BL-421, No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
{variant, sychien}@media.ee.ntu.edu.tw

ABSTRACT

A robust video object segmentation algorithm for complex conditions in surveillance systems is proposed in this paper. This algorithm contains an unsupervised K-Means background clustering technique to model the temporal distribution in RGB domain for each spatial position. Based on the proposed background model, the object mask generation process integrates noise reduction, cast shadow cancellation, and improved watershed transform to obtain satisfying object masks. Experiments show that it can be applied on low-fame-rate and noisy video sequences in surveillance systems in which temporal tracking becomes impractical, and achieve better segmentation results than the previous works for complex lighting conditions and outdoor scenes.

1. INTRODUCTION

Intelligent video surveillance systems, which discover semantics in captured video scenes, can achieve unsupervised monitoring and provide indexing and retrieval functions. The enabling technology for such systems is video object segmentation, with which the moving objects can be extracted from video sequences. Many segmentation algorithms have been proposed. Among them, algorithms with background modeling usually show superior performance. Representative algorithms of them include an efficient and fast method using background registration [1] and non-parametric model for background subtraction [2]. Although these algorithms can derive satisfying segmentation results for simple conditions, for complex conditions, such as varied lighting in spatial and temporal domain, poor illumination situation, background transition [3], and outdoor scenes, a robust segmentation algorithm is required. Sometimes temporal tracking technique can also be employed to alleviate these problems, but fast-moving objects cannot be easily tracked in low-frame-rate video sequences. Besides, to deal with complex lighting situations, both luminance and chrominance components should be considered in the robust segmentation algorithm.

Based on these concepts, we propose a new video segmentation algorithm with an unsupervised background training technique based on K-Means clustering in RGB domain, whose advantages include the capability of handling background transition and flexibility in different lighting condition. Besides, the shadow cancellation technique is integrated to eliminate the effect of cast shadow areas [4]. Finally, an improved watershed transform is applied to obtain smooth and satisfying results in the object masks [5] [6] by utilizing spatial information.

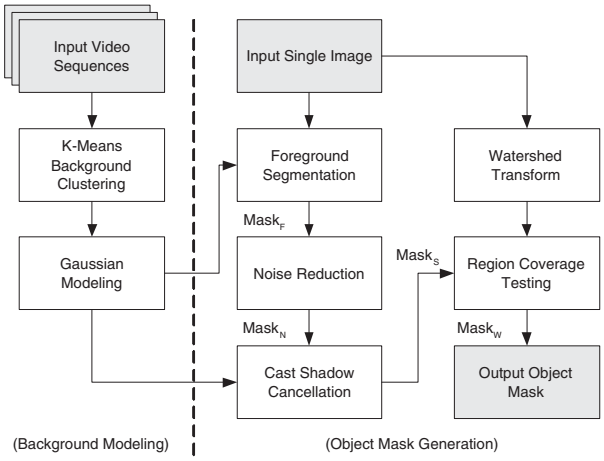

Fig. 1. Overview of the proposed algorithm.

The paper is organized as follows. The proposed algorithm is first described in Sec. 2. Next, in Sec. 3, the experimental results will be shown. Finally, a short conclusion is given in Sec. 4.

2. PROPOSED ALGORITHM

An overview of the proposed algorithm can be illustrated in Fig. 1. The left side of the figure is background modeling, whereas the right side is object mask generation.

2.1. Background Modeling

The background decision boundary is determined by K-Means background clustering and Gaussian modeling. For pixels at the same fixed position in a video sequence, background pixels usually belong to a static scene, such as Fig. 2(a), and foreground pixels usually belong to different objects such as vehicles and human beings, such as Fig. 2(b)(c). The temporal pixel value distribution in RGB domain at this fixed position can be shown in Fig. 2(d), where foreground and background clusters can be roughly separated by human eyes.

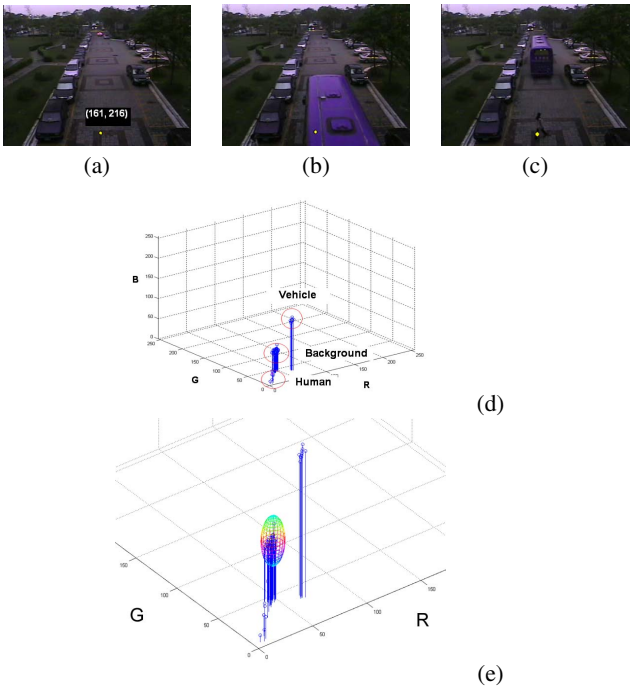

Fig. 2. An example of different clusters of pixel (161, 216). (a) Background pixel. (b) Foreground pixel: vehicles. (c) Foreground pixel: human beings. (d) Three main clusters which can be easily observed in the RGB space. (e) The background ellipsoid in the RGB space of the proposed Gaussian model.

2.1.1. K-Means Background Clustering

The iterative steps of the proposed K-Means background clustering algorithm are performed for each spatial position in the temporal domain, stated as follows:

Step 1: Determine a suitable value of K from prior information or optimization of Bayesian Information Criterion (BIC) [7] for certain amount of pixels in the temporal domain.

Step 2: Classify each pixel in the temporal domain at the fixed position into K clusters with an initial mean vector of each cluster.

Step 3: Assign each pixel to its corresponding cluster according to the nearest mean vector, and the distortion D_n in the n-th iteration is defined as

$$D_n = \sum_{i=1}^{K} \sum_{j \in S_{n,i}} \|\boldsymbol{I}_j - \boldsymbol{\mu}_{n,i}\|^2, \tag{1}$$

where $\boldsymbol{I}_j$ is the RGB vector of an input pixel in the temporal domain, $\boldsymbol{\mu}_{n,i}$ and $S_{n,i}$ representing the mean vector and the set of pixels of the i-th cluster in the n-th iteration, respectively.

Step 4: When $n > 1$, calculate the ratio of the distortion difference. This ratio is defined as

$$R_n = \frac{|D_{n-1} - D_n|}{D_n}. \tag{2}$$

The iteration stops when $R_n < Th_k$, where Th_k is a small threshold close to zero. Otherwise, compute the mean vector of each cluster from previous assignments of pixels, return to step 3, and n is incremented.

2.1.2. Gaussian Modeling

Obtained from K-Means clustering, the largest cluster is considered as the background cluster for each position (x, y). The distribution of background pixels is modeled as a Gaussian distribution in the temporal domain. It roughly determines the decision boundary between background and foreground. For simplicity and efficiency, we assume that each channel in the RGB domain is independent, and the Gaussian model includes the following parameters:

$$\boldsymbol{I}(x,y) = \begin{pmatrix} I_R(x,y) \\ I_G(x,y) \\ I_B(x,y) \end{pmatrix}, \boldsymbol{\mu}(x,y) = \begin{pmatrix} \mu_R(x,y) \\ \mu_G(x,y) \\ \mu_B(x,y) \end{pmatrix}, \tag{3}$$

$$\boldsymbol{\Sigma}(x,y) = \begin{pmatrix} \sigma_R^2(x,y) & 0 & 0 \\ 0 & \sigma_G^2(x,y) & 0 \\ 0 & 0 & \sigma_B^2(x,y) \end{pmatrix}, \tag{4}$$

where $\boldsymbol{I}(x,y)$ is the RGB vector of an input pixel, $\boldsymbol{\mu}(x,y)$ is the mean vector, and $\boldsymbol{\Sigma}(x,y)$ is the covariance matrix. Thus, the background Gaussian model can be denoted as $N[\boldsymbol{\mu}(x,y), \boldsymbol{\Sigma}(x,y)]$. An example of the Gaussian model of a background cluster is shown in Fig. 2(e). Note that, different model parameters are maintained for different positions respectively.

2.2. Object Mask Generation

The proposed method for object mask generation consists of two parallel processes shown in Fig. 1.

2.2.1. Foreground Segmentation

Based on the Mahalanobis distance measurement, the discriminant function is an ellipsoid in RGB space and is defined as

$$g(x,y) = [\boldsymbol{I}(x,y) - \boldsymbol{\mu}(x,y)]^T [\alpha \boldsymbol{\Sigma}(x,y)]^{-1} [\boldsymbol{I}(x,y) - \boldsymbol{\mu}(x,y)], \tag{5}$$

where α is a scalar to determine the size of the ellipsoid. The higher the value of α, the more background pixels are covered. In this stage, the output mask is given by

$$Mask_F(x,y) = \begin{cases} 1, & \text{if } g(x,y) > 1 \\ 0, & \text{otherwise.} \end{cases} . \tag{6}$$

2.2.2. Noise Reduction

The noise reduction scheme is based on area filtering [1] and is shown in Fig. 3(a). Four-connectivity connected component operation is used, and the area thresholds Th_{White} and Th_{Black} can be adjusted according to the noise properties. This process can effectively reduce noise pixels in the output mask, also preserving more high frequency details than binary morphological opening. The output mask in this stage is denoted as $Mask_N(x,y)$.

2.2.3. Cast Shadow Cancellation

The output mask after shadow cancellation is given by:

$$Mask_S(x,y) = Mask_N(x,y) \cdot f_{Shadow}(x,y), \tag{7}$$

where $f_{Shadow}(x,y)$ is a binary shadow function in YCbCr domain, simplified and improved from [4], and thresholds of chrominance and gradient variation can be obtained from the background cluster.

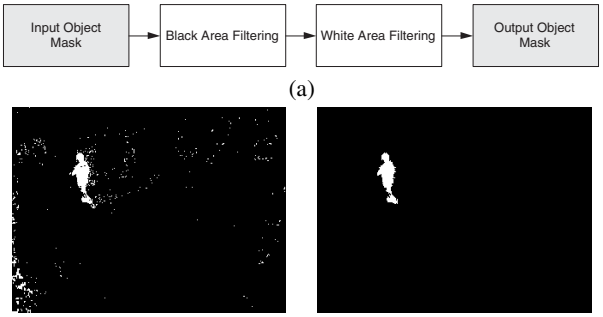

Fig. 3. (a) Illustration of noise reduction process. (b) The mask before noise reduction. (c) The mask after noise reduction.

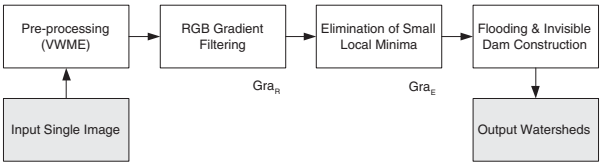

Fig. 4. Illustration of the proposed watershed transform.

2.2.4. *Watershed Transform*

Watershed transform is a common technique frequently applied in image and video segmentation [5]. The proposed watershed transform, which consists of four steps, is illustrated in Fig. 4.

In *Pre-processing* step, the image is processed by an edge-oriented low pass filter. It is found that Variance-Weighted-Mean-Estimator (VWME) filter [8] has a good property to preserve edges and to reduce noise. The result can be seen in Fig. 5.

In *RGB Gradient Filtering* step, since simply using luminance component is not enough, the proposed gradient filter is extended to RGB space based on the concept of morphological gradient filter. The maximum RGB distance is calculated in a window, and the gradient image of the input image is defined as

$$Gra_R(x,y) = \max_{(m,n)\in S} \|\boldsymbol{I}(m,n) - \boldsymbol{I}(x,y)\|, \qquad (8)$$

where S is the set of pixels in a $N \times N$ window centered at (x, y).

In *Elimination of Small Local Minima* step, small local minima in the gradient image are eliminated to avoid the problem of over-segmentation. The output gradient image in this step is

$$Gra_E(x,y) = \begin{cases} Gra_R(x,y), & \text{if } Gra_R(x,y) > Th_M \\ 0, & \text{otherwise.} \end{cases}, \qquad (9)$$

where Th_M is the threshold to determine the segmentation density. The result can be shown in Fig. 6.

In *Flooding and Invisible Dam Construction* step, dams among regions are labeled with the largest region in a window. Thus, every pixel in the image belongs to a single region.

Fig. 5. Illustration of *Pre-processing* step. (a) The original image. (b) The image after VWME filtering.

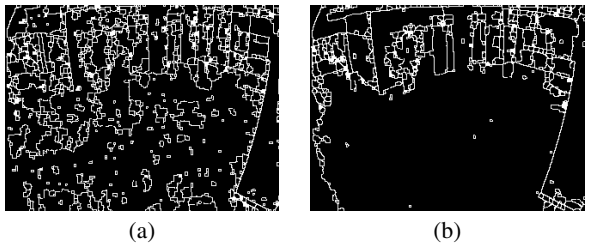

Fig. 6. Illustration of *Elimination of Small Local Minima* step. (a) Watersheds before local minima elimination. (b) Watersheds after local minima elimination.

2.2.5. *Region Coverage Testing*

In the final stage of the proposed algorithm, the mask is refined by utilizing spatial information of watersheds. For each region W_i derived by the proposed watershed transform,

$$Mask_W(x,y) = \begin{cases} 1, & \text{if } \frac{\sum_{(m,n)\in W_i} Mask_S(m,n)}{\sum_{(m,n)\in W_i}} \geq \beta \\ 0, & \text{otherwise.} \end{cases}, \qquad (10)$$

where β is a constant determining the coverage ratio.

3. EXPERIMENTAL RESULTS

Three surveillance sequences provided by 2006 IPPR contest [9], which is held in Taiwan, are employed in the experiments. These video sequences are captured in different places, corrupted by noise, and the frame rates are 5fps, 15fps, and 5fps, respectively as shown in Fig. 7(a)(d)(g). The parameter α is set to 9 and β is set to 0.5 for these sequences. The extracted masks with the algorithm proposed in [1] are used for comparison, shown in Fig. 7(b)(e)(h), and the extracted masks with the proposed method are shown in Fig. 7(c)(f)(i). For *Scene1* in Fig. 7(c), people in both left side and right side are segmented correctly by the proposed algorithm, not affected by the environment with poor illumination. For *Scene2* in Fig. 7(f), the shadow on the wall is successfully cancelled, and the person whose pants color is similar to background is segmented. Also, the person in the dark region is captured. For *Scene3* in Fig. 7(i), which is an outdoor scene, although sizes of moving objects are quite different, vehicles and human beings are both extracted in the object mask. In addition, the shapes of the vehicles are clear and complete. The ground truth masks $Mask_{GD}(x,y)$ are produced by human and offered by 2006 IPPR contest [9]. The error rate is defined as

$$ER = \frac{\sum_{(x,y)\in I} Mask_W(x,y) \oplus Mask_{GD}(x,y)}{\sum_{(x,y)\in I}}, \qquad (11)$$

where I is the set containing the whole image, and $\oplus$ denotes exclusive OR operation for binary masks. With ground truth masks of 150 frames for each sequence, the error rates are 0.2614%, 0.4640%, and 0.4843% for *Scene1*, *Scene2*, and *Scene3*, respectively. The average error rate is only about 0.4%. Furthermore, the proposed algorithm wins the first-place award in 2006 IPPR contest.

4. CONCLUSION

In this paper, a robust video object segmentation algorithm is proposed for fixed cameras. It includes a background modeling technique based on K-Means clustering, which can effectively classify the foreground and background. The object mask generation process with region coverage testing and cast shadow cancellation is capable of extracting clear shapes and small moving objects in different lighting conditions. Experiments show that this algorithm works well even on low-frame-rate video sequences, sequences with complex lighting conditions, and outdoor video sequences, which cannot be correctly processed by the previous works.

5. REFERENCES

[1] S.-Y. Chien, S.-Y. Ma, and L.-G. Chen, "Efficient moving object segmentation algorithm using background registration technique," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 7, pp. 577–586, July 2002.

[2] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proceedings of the 6th European Conference on Computer Vision*, 2000.

[3] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1999, pp. 23–25.

[4] G. Fung, N. Yung, G. Pang, and A. Lai, "Towards detection of moving cast shadows for visual traffic surveillance," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2001, pp. 2505–2510.

[5] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 539–546, Sept. 1998.

[6] Y.-P. Tsai, C.-C. Lai, Y.-P. Hung, and Z.-C. Shih, "A bayesian approach to video object segmentation via merging 3-D watershed volumes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 175–180, Jan. 2005.

[7] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 2000, pp. 727–734.

[8] F. Sattar, L. Floreby, G. Salomonsson, and B. Lovstrom, "Image enhancement based on a nonlinear multiscale method," *IEEE Transactions on Image Processing*, vol. 6, no. 6, pp. 888–895, June 1997.

[9] The Chinese Image Processing and Pattern Recognition Society (IPPR), "http://www.ippr.org.tw/."

Fig. 7. The original image of (a) *Scene1*, (d) *Scene2*, and (g) *Scene3*. The masks of (b) *Scene1*, (e) *Scene2*, and (h) *Scene3* obtained by [1], and the masks of of (c) *Scene1*, (f) *Scene2*, and (i) *Scene3* obtained by the proposed algorithm.