Video Foreground/Background Segmentation using Spatially Distributed Model and Edge-based Shadow Cancellation

Shian-De Tsai, Jin-Jang Leou and Han-Hui Hsiao
Department of Computer Science and Information Engineering, National Chung Cheng University,
621 Chiayi, Taiwan

Keywords: Video Foreground/Background Segmentation, Spatially Distributed Model, Edge-based Shadow Cancellation.

Abstract: Video foreground/background segmentation is to extract relevant objects (the foreground) from the background of a video sequence, which is an important step in many computer vision applications. In this study, the spatially distributed model is built by a splitting process using Gaussian probability distribution functions in spatial and color spaces. Then, edge-based shadow cancellation is employed to obtain more robust segmentation results. The proposed approach can well handle illumination variations, shadow effect, and dynamic scenes in video sequences. Based on experimental results obtained in this study, as compared with two comparison approaches, the proposed approach provides the better video segmentation results.

1 INTRODUCTION

Video foreground/background segmentation is to extract relevant objects (the foreground) from the background of a video sequence, which is the important step in many computer vision applications. Because a video sequence may contain illumination variations, shadow effect, dynamic scenes, …, video foreground/background segmentation is a challenging task.

Existing video foreground/background segmentation approaches include three categories, namely, thresholding, background subtraction, and motion-based. The first category of approaches is based on thresholding pixel differences between two related frames (two consecutive frames or the current frame and a background frame). Because segmentation results are sensitive to thresholding values, various adaptive thresholding approaches were proposed (Tsaig and Averbuch, 2002); (Kim and Hwang, 2002).

For the second category of approaches, Heikkila and Pietikainen (2006) proposed an efficient texture-based method for background modeling. The local binary pattern (LBP) texture operator is employed, which has several good properties for background + modeling. Zhang et al. (2008) proposed a novel dynamic background subtraction approach based on the covariance matrix descriptor. The covariance matrix integrates the pixel-level and region-level features together and efficiently represents the correlation between features. Wang et al. (2008) presented three algorithms (running average, median, mixture of Gaussian) for modeling the background directly from the compressed video. Their approach utilizes DCT coefficients at block level to represent background, and adapts the background by updating DCT coefficients. Li et al. (2004) proposed a Bayesian framework that incorporates spectral, spatial, and temporal features to characterize the background appearance. A Bayes decision rule is derived for classification based on the statistics of principal features.

For the third category of approaches, motion-based foreground/background segmentation can be treated as fitting a collection of motion models to spatiotemporal image data. Mezaris et al. (2004) proposed a model-based foreground/background segmentation approach including three stages: initial segmentation of the first frame using color, motion, and position features, a temporal tracking algorithm, and a trajectory-based region merging procedure. Wang et al. (2005) proposed a Bayesian network to model interactions among the motion vector field, the intensity segmentation field, and the video segmentation field. The Markov random field is then

This work was supported in part by National Science Council, Taiwan, Republic of China under Grants NSC 98-2221-E-194-034-MY3 and NSC 99-2221-E-194-032-MY3.

Tsai S., Leou J. and Hsiao H..
Video Foreground/Background Segmentation using Spatially Distributed Model and Edge-based Shadow Cancellation.
DOI: 10.5220/0004069600890095
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications and Wireless Information Networks and Systems* (SIGMAP-2012), pages 89-95
ISBN: 978-989-8565-25-9

used to encourage the formation of continuous regions.

2 PROPOSED APPROACH

In proposed approach, spatial and color information are used as frame features and the camera is assumed to be stationary. Each pixel in frame *t* is described as a 5-dimensional feature vector $\bar{x}_t = [x, y, Y, U, V]^T$, where (x,y) is the pixel coordinate, color is encoded by the YUV format, and *T* denotes transpose. The probability distribution function of $\bar{x}_t$ for model component *j* is given by:

$$p(\bar{x}_t \mid \theta_{(j,t)}) = \omega_{(j,t)} \frac{e^{-\frac{1}{2}(\bar{x}_t-\mu_{(j,t)})(\Sigma_{(j,t)})^{-1}(\bar{x}_t-\mu_{(j,t)})}}{\sqrt{(2\pi)^d \left|\Sigma_{(j,t)}\right|}}, \qquad (1)$$

where the parameters $\theta_{(j,\,t)} = \{\omega_{(j,\,t)}, \mu_{(j,\,t)}, \Sigma_{(j,\,t)}\}$ are the weight, mean, and covariance matrix of model component *j* in frame *t*, and the dimensionality *d* is 5. As shown in Figure 1, a spatially distributed background model, a Gaussian foreground detection process, and edge-based shadow cancellation are employed in the proposed approach.

2.1 Building the Background Model

To remove noise and tune the boundaries of foreground objects, the Gaussian smoothing filter is applied on each video frame, which is described as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \qquad (2)$$

The background model is initialized, based on the first video frame of a video sequence. It is assumed that no foreground objects appear in the first frame. Then, an iterative splitting procedure, namely, principal direction divisive partitioning (PDDP) (Boley, 1998) is employed to initialize background component(s).

Initially, the whole background frame is treated as a single component, and the support map value of each background pixel is set to 1. Then, a single iteration of the splitting procedure divides each existing component into two new components. Given the background frame and the current components, each component having the highest spatial variance will be determined. Here, it is assumed that the spatial and color distributions are independent. Using the spatial covariance matrix of each component, the eigenvalue λ_j^s and the corresponding eigenvector v_j^s of component *j* with spatial distribution *s* can be computed. The splitting component C_{sp}^s is given by

$$C_{sp}^s = \arg\max_j \{\lambda_j^s\}. \qquad (3)$$

If $\lambda_{sp}^s > T_{sp}^s$, where T_{sp}^s is a predefined threshold, then the splitting component is split and a new component is generated and reassigned as the pixels $\bar{x}$ of the splitting component which satisfy

$$(\bar{x}^s - \mu_{sp}^s) \cdot v_{sp}^s > 0, \qquad (4)$$

where μ_{sp}^s is the spatial mean of the splitting component C_{sp}^s. That is to place a separating plane through the spatial mean, perpendicular to v_{sp}^s. The parameter sets of two corresponding components are then re-estimated based on their newly assigned pixels, respectively, and the support map value of each pixel is updated correspondingly. The proposed approach applies the splitting procedure on both the spatial and color component frames. The splitting procedure will be iterated until the largest eigenvalue of a splitting component is smaller than a threshold.

Because a component may contain two or more spatially disconnected regions, which should be split. Here, the connected component algorithm (Haralick and Shapiro, 1992) is employed to further split a component containing two or more spatially disconnected regions. The initial background model is thus completely constructed for the background.

For a subsequent frame, each pixel $\bar{x}_t$ is assigned to the most likely model component, i.e., each pixel can be assigned to the component with the maximum posterior probability C_{map} defined as

$$C_{map} = \arg\max_j \{\log(p(\bar{x}_t \mid \theta_j))\}. \qquad (5)$$

Because the spatial and color distributions are assumed to be independent, the distribution function in Eq. (5) can be re-expressed as the function of a 2-D spatial Gaussian and a 3-D color Gaussian with parameter sets θ_j^s and θ_j^c for the spatial vector $\bar{x}_t^s = [x, y]^T$ and the color vector $\bar{x}_t^c = [Y, U, V]^T$, respectively. Hence, C_{map} can be modified as

$$C_{map} = \arg\max_j \{\log(p(\bar{x}_j^s \mid \theta_j^s)) + \log(p(\bar{x}_j^c \mid \theta_j^c))\}. \qquad (6)$$

The support map value of each pixel of the new frame is updated to respond the new assignment. Additionally, to obtain stable assignments, a pixel can only be assigned to a background component if

$$\log(p(\bar{x}_t^s \mid \theta_j^s)) > T_{lik}^s, \qquad (7)$$

where T_{lik}^s is a predefined threshold.

Next, to detect new foreground object(s) in the new frame, the "unassigned" pixels in the support map should be detected. Here, if pixel $\vec{x}_t$ of a component satisfies

$$\log(p(\vec{x}_t \mid \theta_{C_{map}})) \le T_{unassign} \ , \tag{8}$$

where $T_{unassign}$ denotes a minimum probability threshold, then $\vec{x}_t$ is determined as an "unassigned" pixel.

2.2 Foreground Detection and Introducing Foreground Model

Initially, all the initial model components of a frame are labeled as the background. A foreground component is detected when some region of pixels having a low probability under the mixture model. Such a region appears in the support map as a region having high density of "unassigned" pixels. The support map is divided into nonoverlapping blocks of size 16×16 pixels. For a block, if the number of "unassigned" pixels exceeds a threshold T_d, it is detected as a "foreground" block. Initially, a single foreground component is built for all unassigned pixels in these "foreground" blocks. Then, similar to the background model, the splitting procedure is recursively applied to build the foreground model.

After pixel assignment and foreground detection, the parameters of both background and foreground components of a frame are re-estimated. Given the parameters of the previous frame $\theta_{(j,t-1)}$, the new parameters $\theta_{(j,t)}$ of the current frame can be computed (using an adaptive learning rate) as

$$\theta_{(j,t)} = \alpha_j \theta_{(j,Cmap)} + (1 - \alpha_j)\theta_{(j,t-1)}, \tag{9}$$

where α_j is a vector of learning rates updated by a variable factor α_j^w, which is proportional to the area fraction of each foreground/background component within a frame.

2.3 Edge-based Shadow Cancellation

After foreground detection, the initial foreground mask FM_t of frame t usually contains both moving objects and some shadows. A shadow often appears in an area where the pixel (gray-level) values change "gradually" from the background to the shadow region. Here, the Canny edge detector (Canny, 1986) is used to generate a binary edge map CE_t of frame t, where "1" denotes the edge and "0" denotes otherwise.

Using the binary edge maps of 5 successive frames, the integrated Canny edge map ICE_t is defined as

$$ICE_t(\vec{x}) = \begin{cases} \text{not edge}, & \text{if } CE_t(\vec{x}) = 0, \\ \text{static edge}, & \text{if } CE_t(\vec{x}) = CE_{t-3}(\vec{x}) = CE_{t-5}(\vec{x}) = 1, \\ \text{moving edge}, & \text{otherwise}, \end{cases} \quad . \tag{10}$$

where $\vec{x}$ denotes a pixel.

Based on ICE_t and FM_t of frame t, the moving Canny edge MCE_t can be defined as

$$MCE_t = \{\vec{x} \mid ICE_t(\vec{x}) = \text{moving Canny edge}, \ \vec{x} \in FM_t\}, \tag{11}$$

whereas the edges of the foreground mask EFM_t can be defined as

$$EFM_t = \{\vec{x} \mid \vec{x} \in FM_t, \vec{x}' \notin FM_t, \ \mathrm{NG}(\vec{x}, \vec{x}') = \mathit{true}\}. \tag{12}$$

where NG($\vec{x}$, $\vec{x}'$) is a logic function that returns *true* when $\vec{x}$ and $\vec{x}'$ are 4-connected neighbors. Note that MCE_t provides important information for seed point selection in region growing (as an illustrated example shown in Figure 2). Because some gaps exist in moving Canny edges MCE_t, in this study, morphologic dilation with a 3×3 structure element is applied on MCE_t, resulting in $DMCE_t$.

Because region growing is used to detect shadow regions, some seed points are required. The shadow region edges SRE_t in FM_t may be employed as seed points, which can be defined as

$$SRE_t = \{\vec{x} \mid \vec{x} \in EFM_t, \ \min\|\vec{x} - \vec{x}'\| > T_{sre}, \vec{x}' \in MCE_t\}, \tag{13}$$

where T_{sre} is a threshold confining the searching neighbor and ||•|| denotes the Chebyshev distance. Additionally, SRE_t also contains some sporadic pixels (the edges of moving objects). To remove sporadic pixels, the connected component algorithm (Haralick and Shapiro, 1992) is also used to connect initial seed points. Each connected region with its number of pixels more than a threshold T_{seed} is included in the final shadow region edges SRE_t^{final}. The pixels in SRE_t^{final} serve as seed points of the region growing algorithm (Adam and Bischof, 1994) used for shadow detection, which are expanded pixel by pixel in FM_t, resulting in the detected shadow FM_t^{shadow}. Note that the pixels in $DMCE_t$ should not be included in FM_t^{shadow} (as an illustrated example shown in Figure 3).

Based on FM_t and FM_t^{shadow}, the initial moving object MO_t (obtained as $MO_t = FM_t - FM_t^{shadow}$)

contains some noisy pixels and holes, which will be processed by a post-processing procedure. In the post-processing procedure, to remove noisy pixels and holes, morphologic erosion with a 3×3 structure element is first applied on MO_t, then the connected component algorithm (Haralick and Shapiro, 1992) is used to remove small connected regions with threshold T_{sre}, and finally morphologic dilation with a 3×3 structure element is applied to obtain the final foreground/background segmentation results (as an illustrated example shown in Figure 4).

3 EXPERIMENTAL RESULTS

In this study, 12 test video sequences and the corresponding ground truth hand segmentations are employed. They are "*Office*," "*Outdoor*," "*Browse1*," "*LightSwitch*," "*NightCar*," "*IntelligenRoom*," "*ParkingLot*," "*OneLeaveShopReenter*," "*WavingTree1*," "*WavingTree2*," "*Raining*," and "*Boat*." Here, sequences 1-3 contain some gradual illumination variations, whereas sequence 4 contains great illumination variations. Sequences 5-8 contain both gradual illumination variations and shadow effect. Finally, sequences 9-12 contain some dynamic scenes, such as waving tree, raining, and moving water. To evaluate the effectiveness of the proposed approach, two comparison methods, namely, self-organizing background subtraction (SOBS) (Maddalena and Petrosino, 2008) and spatially distributed model (SDM) (Dickinson et al., 2009) are implemented in this study. The parameter values and thresholds used in the proposed approach are listed in Table 1, which are empirically determined in this study.

To evaluate the performance of the three comparison approaches, the Jaccard coefficient J_c by Rosin and Ioannidism (2003) and the total error (E_{tot}) by Toyama et al. (1999) are employed. A pixel being classified as "foreground" by both the approach and the ground truth is denoted as "true positive" (*TP*). If it is classified as "foreground" by only the approach, it is denoted as "false positive" (*FP*). If it is classified as "foreground" by only the ground truth, it is denoted as "false negative" (*FN*). If *TP*, *FP*, and *FN* denote the numbers of "true positive," "false positive," and "false negative" pixels in a video sequence, respectively, then

$$J_c = \frac{TP}{(TP + FP + FN)}, \tag{14}$$

$$E_{tot} = FP + FN. \tag{15}$$

In Figure 5, as compared with two comparison approaches, the proposed approach can handle video sequences containing shadow effect and gradual illumination variations, whereas in Figure 6, as compared with two comparison approaches, the proposed approach can handle video sequences containing dynamic scenes, such as waving tree, raining, and moving water.

Additionally, in terms of two performance indexes, namely, Jaccard coefficients and total errors listed in Table 2, the performance of the proposed approach is better than those of two comparison approaches.

4 CONCLUDING REMARKS

In this study, a video foreground/background segmentation approach using spatially distributed model and edge-based shadow cancellation is proposed to deal with video sequences containing illumination variations, shadow effect, and dynamic scenes. Based on the experimental results obtained in this study, as compared with two comparison methods, the proposed approach provides the better video segmentation results.

REFERENCES

Adam, R. and Bischof, L., 1994. Seeded region growing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(6), 641-647.

Boley, D., 1998. Principle direction devisive partitioning. *Data Mining and Knowledge Discovery*, 2(4), 325-344.

Canny, J. F., 1986. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(11), 679-698.

Dickinson, P., Hunter, A., and Appiah, K., 2009. A spatially distributed model for foreground segmentation. *Image and Vision Computing*, 27(9), 1326-1335.

Haralick, R. M. and Shapiro, L. G., 1992. Reading, MA: Addision-Wesley. *Computer and Robot Vision*, 28-48.

Heikkila, M., Pietikainen, M., and Member, S., 2006. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(4), 657-662.

Kim, C. and Hwang, J. N., 2002. Fast and automatic video object segmentation and tracking for content-based applications. *IEEE Trans. on Circuits and Systems for Video Technol.*, 12(2), 122-129.

Li, L. et al., 2004. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. on Image Process.*, 13(11), 1459-1472.

Maddalena, L. and Petrosino, A., 2008. A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Trans. on Image Process.*, 17(7), 1168-1177.

Mezaris, V., Kompatsiaris, I., and Strintzis, M. G., 2004. Video object segmentation using Bayes-based temporal tracking and trajectory-based region merging. *IEEE Trans. on Circuits and Systems for Video Technol.*, 14(6), 782-795.

Rosin, P. and Ioannidism, E., 2003. Evaluation of global image thresholding for change detection. *Pattern Recognition Letters*, 24(14), 2345-2356.

Toyama, K., Krumm, J., Brumitt, B., and Meyers, B., 1999. Wallflower: principles and practice of background maintenance. in *Proc. of IEEE Int. Conf. on Computer Vision*, 255-261.

Tsaig, Y. and Averbuch, A., 2002. Automatic segmentation of moving objects in video sequences: a region labeling approach. *IEEE Trans. on Circuits and Systems for Video Technol.*, 12(7), 597-612.

Wang, W., Yang, J., and Gao, W., 2008. Modeling background and segmenting moving objects from compressed video. *IEEE Trans. on Circuits and Systems for Video Technol.*, 18(5), 670-681.

Wang, Y. et al., 2005. Spatiotemporal video segmentation based on graphical models. *IEEE Trans. on Image Process.*, 14(7), 937-947.

Zhang, S. et al., 2008. A covariance-based method for dynamic background subtraction. in *Proc. of IEEE Int. Conf. on Pattern Recognition*, 1-4.

APPENDIX

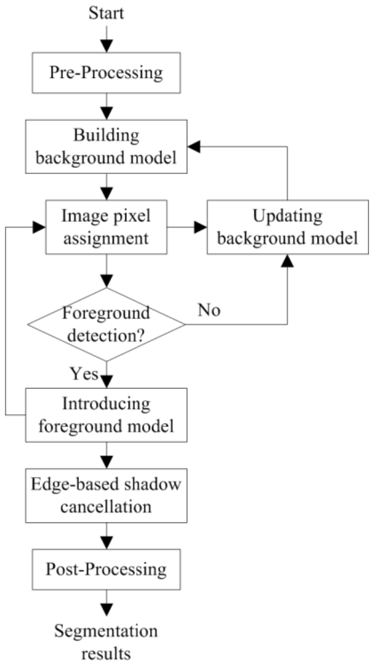
Start
Pre-Processing
Building background model
Image pixel assignment
Updating background model
Foreground detection?
No
Yes
Introducing foreground model
Edge-based shadow cancellation
Post-Processing
Segmentation results

Figure 1: The proposed video foreground/background segmentation approach.

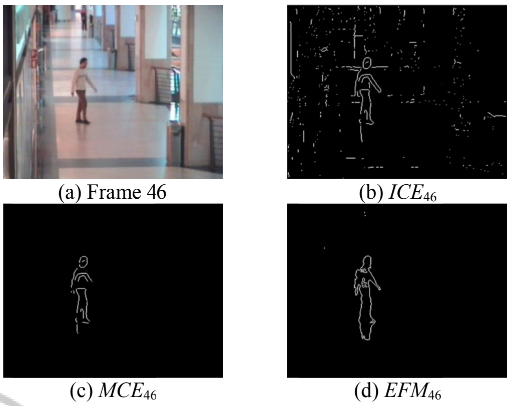
(a) Frame 46 (b) ICE_{46}
(c) MCE_{46} (d) EFM_{46}

Figure 2: (a) Frame 46, (b) ICE_{46}, (c) MCE_{46}, and (d) EFM_{46}.

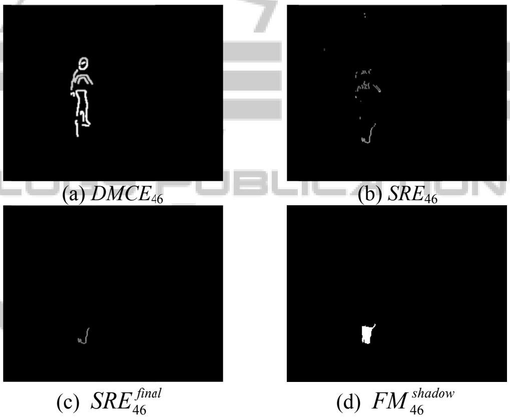
(a) $DMCE_{46}$ (b) SRE_{46}
(c) SRE_{46}^{final} (d) FM_{46}^{shadow}

Figure 3: An illustrated example of region growing.

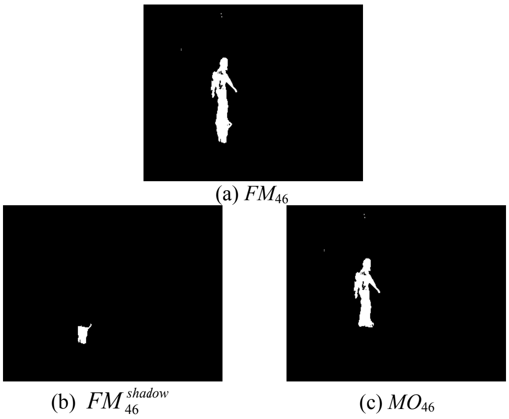
(a) FM_{46}
(b) FM_{46}^{shadow} (c) MO_{46}

Figure 4: An illustrated example of shadow cancellation.

Table 1: Parameter values and thresholds used in the proposed approach.

Parameters and thresholds	Values
Max background components	300
Max foreground components	250
T_{sp}^{s} for background split (spatial)	800
T_{sp}^{c} for background split (color)	50
T_{lik}^{s} for background spatial likelihood	2.4×10^{-8}
T_{d} for unassigned pixel density	0.8
T_{sre} for searching neighbor	2
T_{seed} for minimum seed point	30

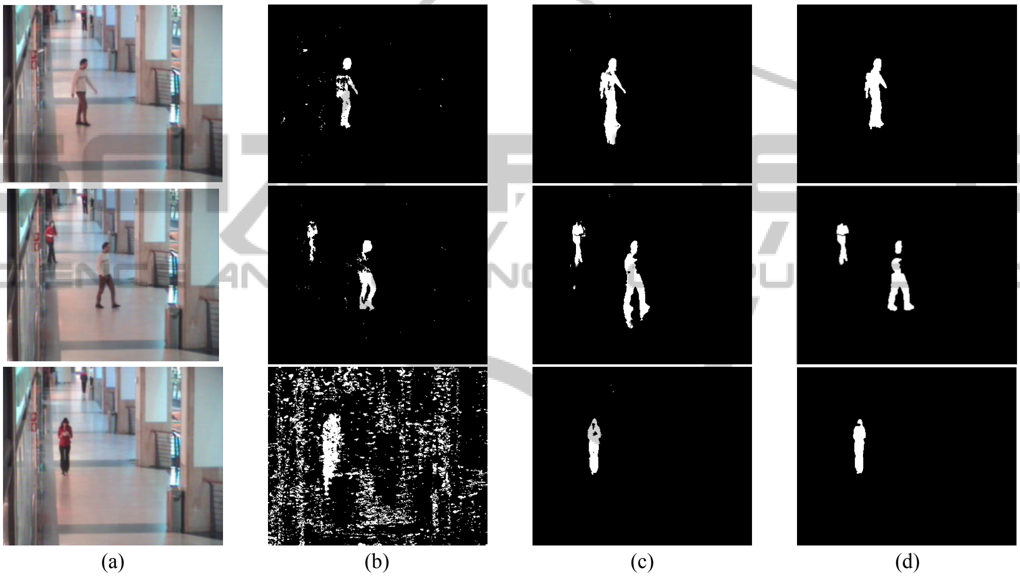

Figure 5: Experimental results of sequence "*OneLeaveShopReenter.*" (a) Original video frames 47, 61, 201; (b)-(d) segmentation results by SOBS, SDM, and the proposed approach, respectively.

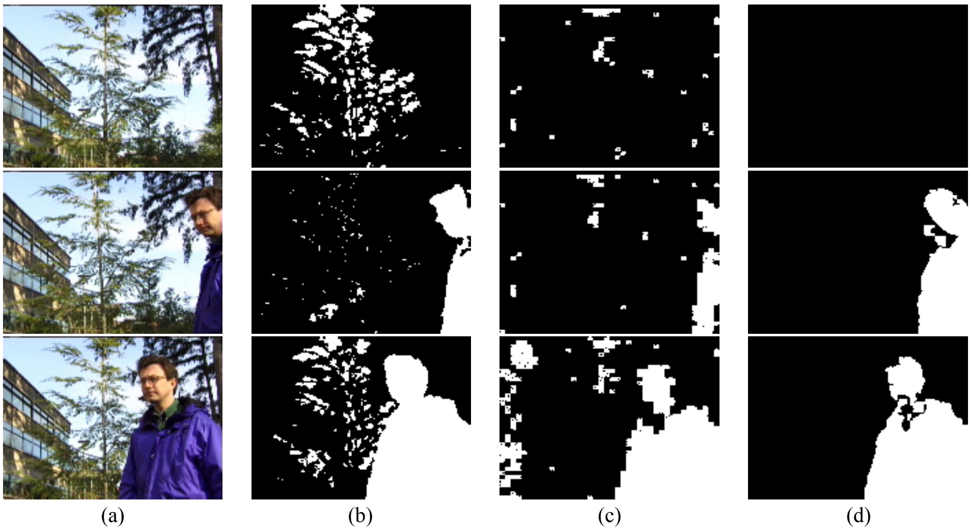

Figure 6: Experimental results of sequence "*WavingTree1.*" (a) Original video frames 10, 19, 21; (b)-(d) segmentation results by SOBS, SDM, and the proposed approach, respectively.

Table 2: Jaccard coefficients and total errors of 12 test sequences by SOBS, SDM, and the proposed approach (Proposed).

Sequence	Jaccard coefficient			Total errors ($\times 10^3$)		
	SOBS	SDM	Proposed	SOBS	SDM	Proposed
Office	0.43	0.41	0.47	13.4	15.7	13.6
Outdoor	0.41	0.46	0.43	18.6	15.0	15.2
Browse1	0.47	0.43	0.54	17.3	18.6	14.3
LightSwitch	0.28	0.23	0.41	21.6	25.5	13.5
NightCar	0.34	0.39	0.41	15.4	13.2	11.2
IntelligenRoom	0.68	0.71	0.78	10.3	8.6	5.5
ParkingLot	0.56	0.47	0.54	15.4	18.6	16.7
OneLeaveShopReenter	0.32	0.41	0.48	16.8	12.3	8.5
WavingTree1	0.23	0.52	0.68	18.3	10.4	7.3
WavingTree2	0.23	0.54	0.62	19.7	11.9	9.8
Raining	0.34	0.42	0.47	18.6	15.4	11.9
Boat	0.29	0.43	0.53	15.5	14.3	11.3
Average	0.38	0.45	**0.53**	16.7	14.9	**11.6**