HUMAN OBJECT TRACKING ALGORITHM WITH HUMAN COLOR STRUCTURE DESCRIPTOR FOR VIDEO SURVEILLANCE SYSTEMS

Shao-Yi Chien, Wei-Kai Chan, Der-Chun Cherng, and Jing-Ying Chang

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
sychien@cc.ee.ntu.edu.tw

ABSTRACT

Segmentation, tracking, and description extraction are important operations in smart camera surveillance systems. In this paper, a robust segmentation-and-descriptor based tracking algorithm is proposed. Segmentation is applied first, and description for each connected component is extracted for object classification to generate the video object masks. It can do segmentation, tracking, and description extraction with a single algorithm without redundant computation. In addition, a new descriptor for human objects, Human Color Structure Descriptor (HCSD), is also proposed for this algorithm. Experimental results show that the proposed algorithm can provide precise video object masks and trajectories. It is also shown that the proposed descriptor, HCSD, can achieve better performance than Scalable Color Descriptor and Color Structure Descriptor of MPEG-7 for human objects.

1. INTRODUCTION

Smart camera surveillance systems [1] [2] [3] [4], which integrate computer vision algorithms, can analyze the video contents on-line or off-line and find suspicious events automatically. They have been used in many military applications and will be commercialized as consumer products in the near future. For each smart camera, segmentation and tracking are the most important operations since the results are the fundamentals for other analysis tools.

Many object tracking algorithms have been proposed [4], which can be summarized with Fig. 1. The conventional tracking algorithms derive the trajectory of each object. As shown in Fig. 1(a), the input data of the tracking system could be image data, segmentation results, or feature points. The core of the system is the *Hypothesis Generation and Prediction* block. It maintains a motion model and predicts the next state of the object by single hypothesis or multiple hypothesis tracking algorithm with Kalman filter. The state is usually the position of the object. Around the predicted position, *Local Refinement* searches the real position in input data. Template matching or mean-shift [5] algorithm can be exploited. The new position and some measurement data are then feedbacked to *Hypothesis Generation and Prediction*. These kinds of tracking algorithms can only generate the trajectory of each object; however, in order to achieve the requirements of object-based coding, analysis, and indexing, the video objects should be identified and tracked with precise contours. The conventional algorithm can be combined with video segmentation to put the segmentation results as the input data in Fig. 1(a) and label the object masks according to the trajectory of each object. On the other hand, contour based tracking algorithms with deformable model (active contour) can be applied, as shown in Fig. 1(b) [6] [7] [8]. After segmenting video objects, the contours are tracked with active contours. Both these two kinds of algorithms may fail when objects are occluded or uncovered, or objects enter or leave the scene. Besides, in order to solve error propagation problems, complex algorithms or reset schemes are usually required.

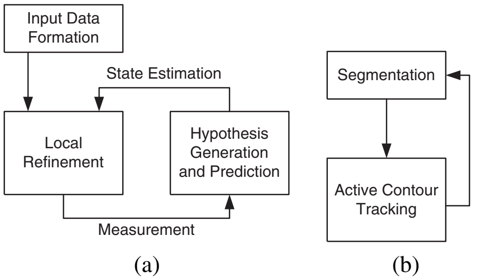


Fig. 1. Block diagram of the conventional tracking algorithms. (a) Region and feature based algorithms. (b) Contour based algorithms.

Since surveillance cameras record video continuously, the duration of the video clips are long. Users need to traverse in the whole video sequence to find some special events of interest, which is very time-consuming. Video object indexing and representation techniques, which can generate some metadata or descriptions for video objects, can help to provide quick content retrieval ability. To integrate description generation in the smart camera systems, with conventional approaches, the description extraction should be applied after segmentation and tracking. It is not efficient since many operations of segmentation, tracking, and description extraction are similar.

Therefore, in our opinion, segmentation, tracking, and description generation should be the same task, and can be done in one highly-integrated algorithm. Bases on our previously-developed segmentation algorithm [9] with change detection and background registration, in this paper, we will develop our tracking algorithm with a new proposed descriptor, Human Color Structure Descriptor (HCSD), as a segmentation-and-descriptor based tracking algorithm.

1-4244-0367-7/06/$20.00 ©2006 IEEE

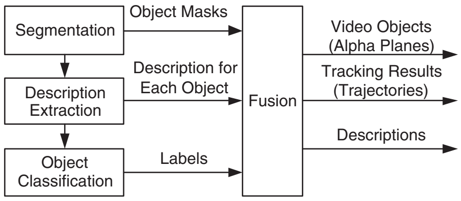


Fig. 2. Block diagram of the proposed segmentation-and-descriptor based tracking algorithm.

2. PROPOSED ALGORITHM

2.1. Segmentation-and-Descriptor Based Tracking Algorithm

The target of the proposed algorithm is to generate segmentation masks, tracking results, and descriptions at the same time with a single algorithm. The overview of the proposed algorithm is shown in Fig. 2. First of all, the segmentation algorithm [9] is applied to generate object masks, which can be used to indicate where the foreground objects are. Then we extract the associated description for each object. With the description of each object in every frame, the objects can be classified into semantic video objects corresponding to different physical objects. Each video object is then given a special label. With object masks, description for each object, and the video object labels, the *Fusion* block can generate the video object information (alpha plane), tracking results (trajectories), and descriptions for all objects at the same time.

2.2. Human Color Structure Descriptor

The objects-of-interest in the target surveillance systems are human objects. Here, we propose a compact descriptor for these objects, which is named as Human Color Structure Descriptor (HCSD). It can be defined by the following equation. For an object i,

$$HCSD_i = \{(\mathbf{c}_{ib}, \mathbf{p}_{ib}), (\mathbf{c}_{il}, \mathbf{p}_{il}), (\mathbf{c}_{is})\}, \tag{1}$$

where $\mathbf{c}_{ib}$, $\mathbf{c}_{il}$, and $\mathbf{c}_{is}$, are the colors of body, legs, and shoes of human object i, respectively, and $\mathbf{p}_{ib}$ and $\mathbf{p}_{il}$ are the positions of body and legs of the object. Each color needs 24 bits, where each position information needs 20 bits. Therefore, the total bit number required for HCSD of each object is 112 bits.

2.3. Descriptor Extraction and Tracking Algorithm

With the concept and the descriptor described in Section 2.1 and Section 2.2, the detailed algorithm is described in this section. The flowchart of the proposed algorithm is shown in Fig. 3, where some example illustrations are shown in Fig. 4. First of all, the segmentation algorithm is applied to get the silhouettes, or object masks $OM(x, y)$, of foreground objects in each frame. That is to say, when a pixel (x, y) belongs to foreground objects, $OM(x, y) = 1$; otherwise, $OM(x, y) = 0$, as shown in Fig. 4(b). Next, the connected component operation is exploited to separate each object. Each connected component is given a special label and is recorded in $CC(x, y)$, as shown in Fig. 4(c). For each connected component, morphology skeleton operation is then applied, which can be described with the following equations. The pixels of connected component i are viewed as a set of pixels of object i,

$$A_i = \{(x, y)|CC(x, y) = i\}. \tag{2}$$

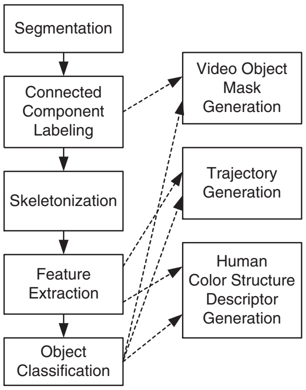

Segmentation
Connected Component Labeling
Skeletonization
Feature Extraction
Object Classification
Video Object Mask Generation
Trajectory Generation
Human Color Structure Descriptor Generation

Fig. 3. Flowchart of the proposed algorithm.

The skeleton of object i is calculated with the following equation.

$$S(A_i) = \bigcup_{k=1}^{K} S_k(A_i), \tag{3}$$

where

$$S_k(A) = (A \ominus kB) - (A \ominus kB) \circ B, \tag{4}$$

$$(A \ominus kB) = (\ldots((\underbrace{A \ominus B) \ominus B) \ominus \ldots) \ominus B}_{k \text{ times}}, \tag{5}$$

$$K = \max\{k|(A \ominus kB) \neq \emptyset\}, \tag{6}$$

$\ominus$ and $\circ$ are morphology erosion and opening operations, respectively, and B is the structuring element of morphology operations. The skeleton of object i is then separated to body skeleton BS and limb skeleton LS with the following equations.

$$BS(A_i) = S_K(A_i), \tag{7}$$

$$LS(A_i) = S(A_i) - BS(A_i). \tag{8}$$

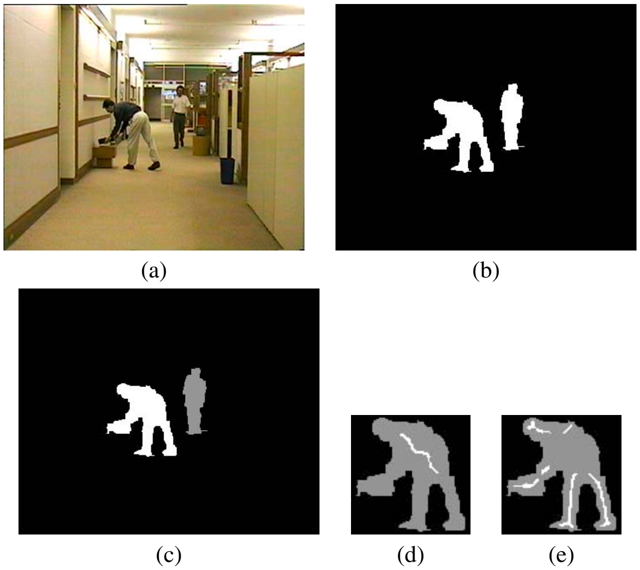
Examples of body and limb skeleton are shown in Fig. 4(d) and Fig. 4(e), respectively. After that, the leg skeleton $LEGS(A_i)$ and shoes skeleton $SHOES(A_i)$ are extracted from limb skeleton by keeping the lower parts and lower ends. The reason to extract skeleton is that it can provide rich structure information and is robust from noise [3]. In the fourth step in Fig. 3, HCSD is extracted. The color information $\mathbf{c}_{ib}$, $\mathbf{c}_{il}$, and $\mathbf{c}_{is}$, are calculated as the average colors of pixels on BS, $LEGS$, and $SHOES$; the position information are the center-of-gravity of BS and $LEGS$. Finally, in the last step, the extracted descriptor, HCSD, is used to classify objects into associated video objects. The object i in the frame t should belong to the same video object as object j in the frame $t - 1$ if

$$d(HCSD_i^t, HCSD_j^{t-1}) < d(HCSD_i^t, HCSD_k^{t-1}|k \neq j),$$

where

$$d(HCSD_i, HCSD_j) = \sqrt{d(\mathbf{c}_{ib}, \mathbf{c}_{jb})^2 + d(\mathbf{c}_{il}, \mathbf{c}_{jl})^2 + d(\mathbf{c}_{is}, \mathbf{c}_{js})^2}. \tag{9}$$

In the fusion stage, the information generated in the flow are integrated. Combining the results of *Connected Component Labeling* and *Object Classification*, it can generate the video object mask



(a) (b)

(c) (d) (e)

Fig. 4. An example of the proposed algorithm, where the test sequence is *Hall Monitor*. (a) *Hall Monitor* #116. (b) Segmentation result. (c) Connected component labeling result. (d) Body skeleton BS of the left object. (e) Limb skeleton LS of the left object.

(alpha plane) for each video object, which can be used for object-based coding or other video analysis tools. The trajectory of each object can be generated from HCSD and the results of *Object Classification*. Similarly, HCSD for each object can also be generated by combining the results of *Feature Extraction* and *Object Classification*.

The proposed tracking algorithm can deal with all the situations including object occlusion/uncovering and the change of number of objects, since each frame is segmented separately, and object masks are then linked with descriptor-based object classification. For the same reason, this algorithm has no error propagation problem. Therefore, it is also robust from noise of the input sequences and noise introduced from the segmentation algorithm. On the other hand, it is obvious that the proposed algorithm is efficient, since segmentation, tracking, and description extraction can be done with a single system without redundant computation.

3. EXPERIMENTAL RESULTS

3.1. Human Object Tracking

The test sequence *Hall Monitor* of MPEG is used to show the ability of the proposed algorithm. In Fig. 5, the extracted information of the left object are shown. The first row shows the video object planes extracted by the proposed segmentation and tracking algorithm. It is shown that the segmentation results are precise. Note that, since the shadow cancellation mode is turned on [9], the shadow parts of the object are not included. In the second and third rows of Fig. 5, the body skeletons and limb skeletons of this human object are also shown. The positions of the skeletons are correct and can give good representation of the object.

Next, some experimental results with sequences captured by real surveillance cameras are demonstrated. Figure 6 shows the tracking results of sequence *BL1F*, which is captured on the first floor of our building, where Fig. 7 shows sequence *BL4F* captured on the fourth floor. Different objects are marked with contours in different colors. It is shown that the proposed algorithm can track these objects correctly and generate precise contour information.

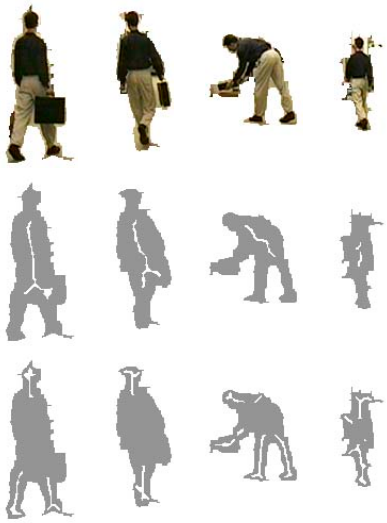


Fig. 5. Experimental results of the proposed algorithm. The test sequence is *Hall Monitor*.

3.2. Human Object Indexing

In this subsection, the proposed descriptor, Human Color Structure Descriptor (HCSD) is compared with two MPEG-7 color descriptors: Color Structure Descriptor (CSD) and Scalable Color Descriptor (SCD) [10]. The color spaces of CSD and SCD are selected as HMMD since the performance is better than other color spaces [10]. These two descriptors are both based on color histogram. Among them, like HCSD, structure of color is also considered in CSD. Some video object planes are picked and used as queries to the database of all the video object planes. In these experiments, the number of detected video object planes is set the same as the ground truth. Therefore, the precision and recall rate are the same, and the miss rate is $(1 - Precision)$. The average results are shown in Table 1. It is shown that the precision orders of all the three sequences are: HCSD $>$ CSD $>$ SCD. Note that, SCD and CSD are scalable descriptors. That is, the number of bits for these descriptors can be adjusted to achieve tradeoff between precision and file size. The precision is adjusted to the highest level in this comparison.

The comparison of the precision-size curve is shown in Fig. 8, where HCSD is represented with a point because it is a descriptor with fixed size. It is proved that the proposed descriptor, HCSD, is effective for human objects since it can achieve higher precision with much fewer bits.

4. CONCLUSION

In this paper, we propose a segmentation-and-descriptor based tracking algorithm with a proposed Human Color Structure Descriptor (HCSD). This algorithm can deal with almost all the situations, has no error propagation problem, and is robust from noise. In addition, the proposed algorithm can accomplish segmentation, tracking, and

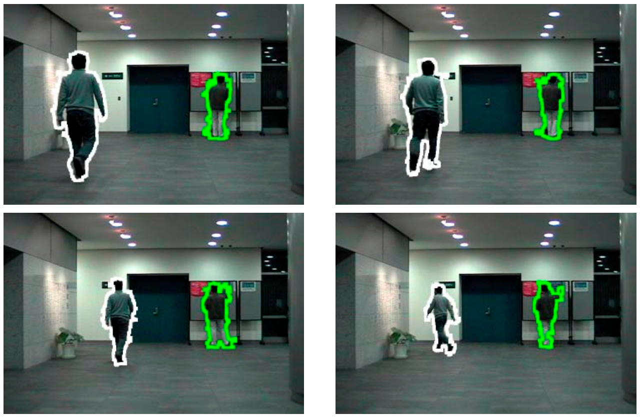


Fig. 6. Experimental results of the proposed algorithm. The test sequence is *BL1F*.

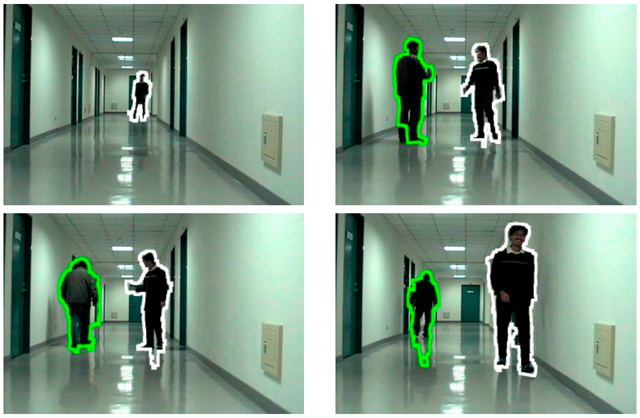


Fig. 7. Experimental results of the proposed algorithm. The test sequence is *BL4F*.

description extraction with a single system without redundant computation. Experiments shows that this algorithm can generate precise object masks and tracking results. It is also shown that the proposed descriptor, HCSD, can perform well for human objects.

5. REFERENCES

[1] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, "Urban surveillance systems: from the laboratory to the commercial world," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1478–1497, Oct. 2001.

[2] A. Del Bue, D. Comaniciu, V. Ramesh, and C. Regazzoni, "Smart cameras with real-time video object generation," in *Proceedings of 2002 International Conference on Image Processing*, 2002, pp. III–429 – III–432.

[3] G. L. Foresti, "Object recognition and tracking for remote video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1045–1062, Oct. 1999.

Table 1. Precision Comparison Between the Proposed Descriptor and MPEG-7 Descriptors.

	Precision(%)		
Sequence	SCD	CSD	HCSD (Proposed)
Hall Monitor	70%	97%	100%
BL1F	97%	98%	99%
BL4F	96%	96%	99%

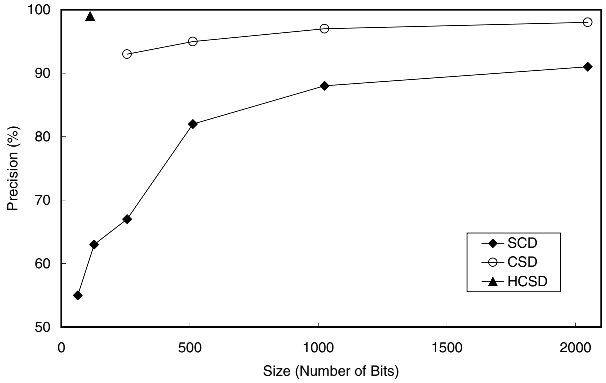


Fig. 8. Precision-size (number of bit) curve of the proposed descriptor and MPEG-7 descriptors. The test sequence is *Hall Monitor*.

[4] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, Aug. 2004.

[5] D. Comaniciu and V. Ramesh, "Mean shift and optimal prediction for efficient object tracking," in *Proceedings of 2000 International Conference on Image Processing*, 2000, pp. III–70–III–73.

[6] M. Kass, A. Witkin, and D. Terzopoulos, "Snake: active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1987.

[7] S. Sun, D. R. Haynor, and Y. Kim, "Semiautomatic video object segmentation using VSnakes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 75–82, Jan. 2003.

[8] A. Yilmaz, X. Li, and M. Shan, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1531–1536, Nov. 2004.

[9] S.-Y. Chien, Y.-W. Huang, B.-Y. Hsieh, S.-Y. Ma, , and L.-G. Chen, "Fast video segmentation algorithm with shadow cancellation, global motion compensation, and adaptive threshold techniques," *IEEE Transactions on Multimedia*, vol. 6, no. 5, pp. 732–748, Oct. 2004.

[10] B. S. Manjunath, P. Salembier, and T. Sikora, Eds., *Introduction to MPEG-7*, John Wiley & Sons, 2002.