Object-Based Video Abstraction for Video Surveillance Systems

Changick Kim, *Member, IEEE*, and Jenq-Neng Hwang, *Fellow, IEEE*

***Abstract*—Key frames are the subset of still images which best represent the content of a video sequence in an abstracted manner. In other words, video abstraction transforms an entire video clip to a small number of representative images. In this paper, we present a scheme for object-based video abstraction facilitated by an efficient video-object segmentation (VOS) system. In such a framework, the concept of a "key frame" is replaced by that of a "key video-object plane (VOP)." In order to achieve an online object-based framework such as object-based video surveillance system, it becomes essential that semantically meaningful video objects are directly accessed from video sequences. Moreover, the extraction of key VOPs needs to be automated and context dependent so that they maintain the important contents of the video while remove all redundancies. Once a VOP is extracted, the shape of the VOP needs to be well described. To this end, both region-based and contour-based shape descriptors are investigated, and the region-based descriptor is selected for the proposed system. The key VOPs are extracted in a sequential manner by successive comparison with the previously declared key VOP. Experimental results on the proposed online processing scheme combined with efficient VOS show the proposed integrated scheme generates desirable summarizations of surveillance videos.**

***Index Terms*—Object-based key frame extraction, shape descriptor, video abstraction, video analysis, video surveillance, video-object segmentation (VOS).**

I. Introduction

TRADITIONAL video-coding standards, such as MPEG-1/MPEG-2 and H.261/H.263, lack high-level interpretation of video contents. The MPEG-4 [1] video standard introduces the concept of a video-object layer (VOL) to support content-based functionality. Its primary objective is to support the coding of video sequences, which are segmented based on video contents, and to allow separate and flexible reconstruction and manipulation of contents at the decoder. Thus, video-object segmentation (VOS), which emphasizes partitioning the video frames into semantically meaningful video objects (VOs) and background, becomes an important issue for successful use of MPEG-4/MPEG-7, or object-based video analysis applications. As an example in MPEG-7, segmented results based on the frame-to-frame motion information or abrupt shape change can be utilized for a semantic-level (object-level) description.

The shape of the VO is a binary image representing the extent of the object. There are many applications where image analysis can be reduced to the analysis of shapes. Shape-analysis methods play an important role in systems for object recognition, matching, registration, and analysis. For example, an intelligent video security system may use shape techniques to determine the identity (or unwanted action) of the intruder. Shape descriptors can be used in e-commerce, where it is difficult to use an annotated text index to specify the required shape of the object, and query-by-example is simpler and faster. In this paper, we present an object-based key-frame selection method which can be deployed in an intelligent video surveillance system, where changes in content are detected through observations made on the objects in the video sequences.

Most of the traditional key-frame extraction (KFE) algorithms are rectangle frame based. Popularly used visual criteria to extract key frames are shot-based criteria, color-feature-based criteria, and motion-based criteria [2]. However, they are limited to rely on low-level image features and other readily available information instead of using semantic primitives of video, such as interesting objects, actions, and events. The early attempt to object-based KFE has been reported in [24], in which the key objects are defined as regions of coherent motions since the extraction of semantic object was not available in the paper. The following KFE works based on the semantic object have been proposed [3], [4]. In [3], the ratio of the number of intra-coded macroblocks (I-MBs) to the total number of (encoded) MBs in a video-object plane (VOP) in intra mode was used as the key frame selection criteria. When the ratio exceeds a certain threshold the frame is labeled as a key frame. However, the MPEG-4 encoder is required and the accuracy of using the ratio of I-MBs is too low to be effective. Erol and Kossentini proposed an automatic key VOP selection framework based on the shape content of VOs [4]. Significant changes in the shape of VOs are detected in the MPEG-4 compressed domain, for which a MPEG-4 decoder is used for parsing and partial decoding of the MPEG-4 bitstream to obtain the shape information of VOs. Unlike previous works, which use pre-segmented VOPs and MPEG-4 codec, our goal is to develop an object-based framework for online video abstraction, which is an extension of our previous work [5]. Naturally, the efficient VOS scheme is integrated for access to VOPs in a scene.

The rest of the paper is organized as follows. Section II presents general issues associated to the proposed system. In Section III, an efficient VOS algorithm is described as the first step of the object-based video abstraction system. The object-based KFE algorithm by sequential selection is introduced in Section IV, promising simulation results are also reported in

Manuscript received January 1, 2001; revised September 18, 2002. This paper was recommended by Associate Editor H. Zhang.

C. Kim is with Epson Research and Development Inc., Palo Alto, CA 94304 USA.

J.-N. Hwang is with the Information Processing Laboratory, University of Washington, Seattle, WA 95119 USA.

Digital Object Identifier 10.1109/TCSVT.2002.806813

this section. Experimental results from the integrated system and the Conclusion follow in Sections V and VI, respectively.

II. Overview of The System

Our goal is to develop core techniques for an intelligent video surveillance system that can detect significant instances based on VO in a scene. Our approach to an integrated scheme for object-based key frame selection is composed of two parts, namely: 1) VOS and 2) object-based key frame selection. In this section, general issues associated to the proposed system are discussed.

A. VOS

There are some special scenarios for which automatic detection of the appearance of a VO is crucially desirable. For instance, it is inevitable in developing object-based video surveillance system, which needs to be implemented online combined with some event detection schemes. A desirable VOS scheme for online object-based applications should meet the following criteria: 1) semantically meaningful objects should be segmented; 2) segmentation algorithm should be efficient; and 3) human intervention for initialization should be minimized.

Change detection for inter-frame difference is one of the most feasible solutions [6]–[8], [23] because it enables automatic detection of new appearance. While the algorithms enable automatic detection of objects and allow larger nonrigid motion compared to object tracking methods [9], object boundaries tend to be irregular in some critical image areas due to the lack of spatial edge information. This drawback can be overcome by using spatial edge information to smooth and adapt the object boundaries in the post-processing stage [7]. We believe that the spatial edge information can be incorporated in the motion-detection stage to simplify algorithm and generate noise resistant results. Though the segmentation of common objects in arbitrary scenes is still beyond the capabilities of an artificial system, we have recently presented an automatic segmentation algorithm that succeeded under constrained conditions [10], [11]. In Section III, we address the VOS algorithm, which is a moving-edge (ME) detection scheme that is a combination of an edge-change detection method and nonlinear filtering to fill in the regions inside MEs. Our approach is restricted by the assumption that one or more VOs are extracted from the video sequences taken by a fixed camera, e.g., surveillance video.

B. Object-Based Key-Frame Selection

Section IV addresses an efficient solution for key-frame selection in presence of single or multiple VOs in a scene. We show the complete video abstraction based on the VOs can be achieved online by describing shapes of VOPs, extracted by an efficient VOS scheme [10], [11]. Broadly, there are two types of shape descriptors: region- and contour-based shape descriptors. Since the region-based shape descriptors, such as moments, make use of all pixels constituting the shape, it can describe any shapes, i.e., not only a simple shape with a single connected region, but also a complex shape that consists of holes in the object or several disjoint regions. The region-based shape descriptor not only can describe such diverse shapes efficiently in a single descriptor, but is also robust to minor deformation along the boundary of the object. The traditionally well-known descriptor is Hu's seven moments, which is invariant to translation, rotation, and scale change [12]. Recently, region-based descriptors using Zernike moments or angular radial transformation have been employed for efficient image retrieval from the large image database [13]–[15].

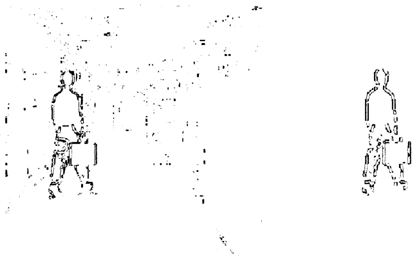


Fig. 1. Edge maps resulting from (a) Eq. (1) and (b) Eq. (2)

Contour-based methods express shape properties of the object contour, including the turning angle method, curvature scale space, and Fourier descriptors [15]. These methods provide a more complete description of shape than region-based descriptors. However, it results in a totally different description if it is a complex object split into multiple disjoint regions, while such situations can be treated well with region-based descriptors. For more details, readers are referred to [15]–[16].

For the proposed system, a desirable shape representation should satisfy the following conditions.

1) *Robustness to small deformation*–unlike the conventional requirement for a shape-based image retrieval system, the shape descriptor for the proposed scheme should be robust to minor changes of the boundaries so that there are not too many key objects selected. For instance, the representation must not be sensitive to the gait of a human in a scene.
2) *Robustness to noise*—since VOs are extracted online, object boundaries may contain undesirable irregularities due to image noise. The representation must be robust to these types of noise.
3) *Feature extraction efficiency*—the feature vector should be computed efficiently.

III. VOS

A. Extraction of a ME Map

Our segmentation algorithm [10], [11] starts with edge detection, which is the first and most important stage of human visual processing, as discovered by Marr and *et al.* [17]. While edge information plays a key role in extracting the physical change of the corresponding surface in a real scene, exploiting simple difference of edges for extracting shape information of moving objects in video sequence suffers from great deal of noise even in stationary background [see Fig. 1(a)]. This is due to the fact that the random noise created in one frame is different from the

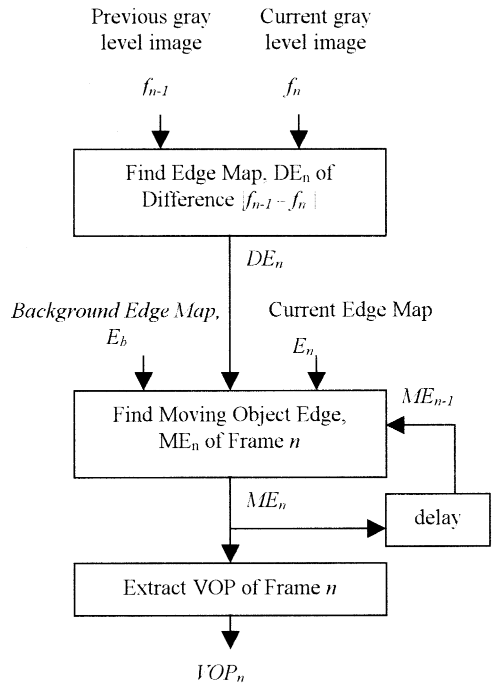

Previous gray level image
Current gray level image
f_{n-1}
f_n
Find Edge Map, DE_n of Difference $|f_{n-1} - f_n|$
DE_n
Background Edge Map, E_b
Current Edge Map E_n
ME_{n-1}
Find Moving Object Edge, ME_n of Frame n
ME_n
delay
Extract VOP of Frame n
VOP_n

Fig. 2. Block diagram of the segmentation

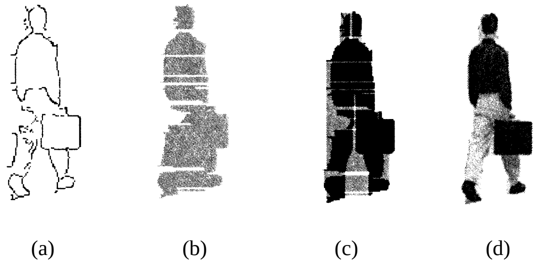


Fig. 3. VOP extraction process. (a) ME map, ME45. (b) Horizontal candidates. (c) Logical AND (black areas) of horizontal and vertical candidates. (d) Extracted VOP after morphological operations.

noise created in the successive frame. The difference of edges is defined as

$$|\Phi(f_{n-1}) - \Phi(f_n)| = |\theta(\nabla G * f_{n-1}) - \theta(\nabla G * f_n)| \quad (1)$$

where the edge maps $\Phi(f)$ are obtained by the Canny edge detector [18], which is accomplished by performing a gradient operation ∇ on the Gaussian convoluted image $G * f$, followed by applying the nonmaximum suppression to the gradient magnitude to thin the edge and the thresholding operation with hysteresis to detect and link edges. On the other hand, edge extraction from difference image in successive frames results in a noise-robust difference edge map DE_n because Gaussian convolution included in the Canny operator suppresses the noise in the luminance difference [see Fig. 1(b)]

$$\mathrm{DE}_n = \Phi(|f_{n-1} - f_n|) = \theta(\nabla G * |f_{n-1} - f_n|). \quad (2)$$

Fig. 2 shows the block diagram of our segmentation algorithm. After calculating edge map of difference of images as shown in Fig. 1(b), we extract the ME ME_n of the current frame f_n based on the edge map DE_n of difference $|f_{n-1} - f_n|$, the current frame's edge map $E_n = \Phi(f_n)$, and the background edge map E_b. Note that E_b contains absolute background edges adopted to increase the extraction performance. For video surveillance sequences, such as “Hall Monitor”, which contains no moving objects in the beginning of the video clip, the edge map of the first frame is used as a background edge map. For sequences which have temporarily still objects from the beginning, such as “Miss America” or “Akiyo”, the background edge map E_b can be created by manually deleting MEs of target objects [10]. We define the edge model $E_n = \{e_1, \ldots e_k\}$ as a set of all edge points detected by the Canny operator in the current frame n. Similarly, we denote $\mathrm{ME}_n = \{m_1, \ldots m_l\}$ the set of l ME points, where $l \le k$. The edge points in ME_n are not restricted to the object boundary, but can also be in the interior of the object boundary. If DE_n denotes the set of all pixels belonging to the edge map from the difference image, then the ME model generated by edge change is given by selecting all edge pixels within a small distance T_{change} of DE_n, i.e.,

$$\mathrm{ME}_n^{\text{change}} = \left\{ e \in E_n \middle| \min_{x \in \mathrm{DE}_n} \|e - x\| \le T_{\text{change}} \right\}. \quad (3)$$

Some ME_n might have scattered noise, which need to be removed before proceeding to the next steps. In addition, the previous frame's MEs can be referenced to detect temporarily still MEs, i.e.,

$$\mathrm{ME}_n^{\text{still}} = \left\{ e \in E_n | e \notin E_b,\ \min_{x \in \mathrm{ME}_{n-1}} \|e - x\| \le T_{\text{still}} \right\}. \quad (4)$$

The final ME map for the current frame f_n is expressed by combining the two maps

$$\mathrm{ME}_n = \mathrm{ME}_n^{\text{change}} \cup \mathrm{ME}_n^{\text{still}}. \quad (5)$$

For the initial ME map ME_0, only a blank image is required in the case of a surveillance video such as “Hall Monitor”. Note that any manual initialization is not required for the surveillance type of sequences, because moving objects will not appear in the very beginning of the sequences. For head-and-shoulder type of video, such as “Miss America” or ‘”Akiyo”, we need to manually delineate the moving object (e.g., by outlining the outer contour of the objects of interest) in the first frame. For details, readers are referred to [10] and [11].

B. Extraction of VOP

With ME map ME_n, as shown in Fig. 3(a), detected from DE_n, VOPs are ready to be extracted. The horizontal candidates are declared to be the region inside the first and last edge points in each row [see Fig. 3(b)] and the vertical candidates for each column. After finding both horizontal and vertical VOP candidates, the intersection regions [black areas in Fig. 3(c)] obtained by logical AND operation are further processed by morphological operations. We regard the AND areas as absolute areas for VOP. Nonlinear filtering using cascaded morphological operations is conducted to achieve spatial continuity and to remove small holes inside moving objects in the segmentation map. This is achieved thanks to the spatial regularization properties of morphological operations, like opening and closing. Some of the extracted VOs are shown in Fig. 4. Fig. 5 shows the VOP extraction system implemented for online demonstration.

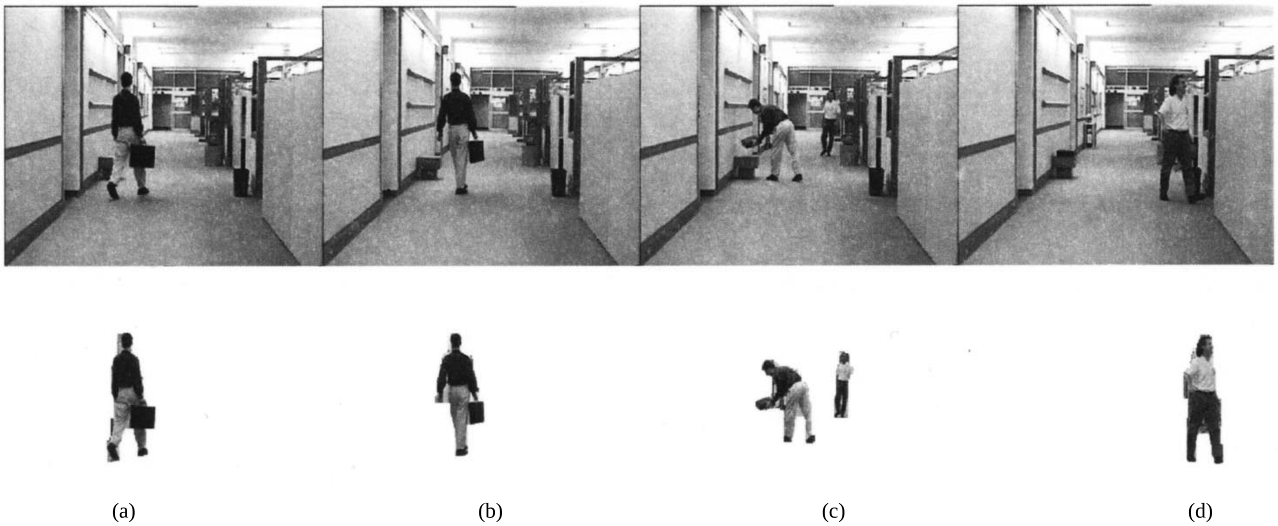
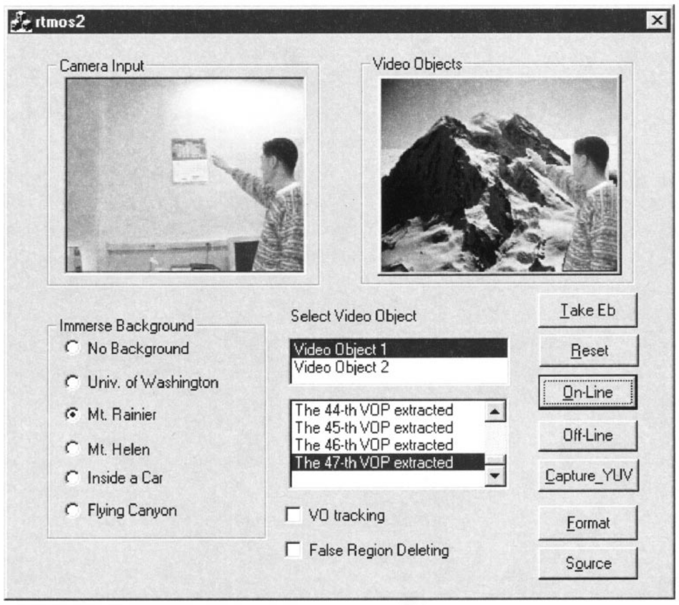


Fig. 4. Extracted VOPs from “Hall Monitor.” Frames (a) 46, (b) 55, (c) 105, and (d) 294.



The 44-th VOP extracted
The 45-th VOP extracted
The 46-th VOP extracted
The 47-th VOP extracted

Fig. 5. Implementation of online VO extraction. Extracted VOP can be easily immersed into various backgrounds in encoder or decoder side.

IV. Object-Based Key-Frame Extraction

A. Problem Formulation

We present a VO-based scheme where key frames are extracted in a sequential manner which is similar to that used for video parsing proposed by Zhang *et al.* [19], [20]. Here, as a starting point, we assume every frame has only one VO throughout the sequence, and thus, a frame can be represented as an object mask image. It is assumed that we have a data set D of K frames; each frame is represented in terms of a d-dimensional feature vector that belongs to either of the two different classes $+1$ (=key frames) or -1 (=nonkey frames)

$$D = \{(x_k, y_k) | k \in \{1, \ldots, K\}, x_k \in \Re^d, y_k \in \{+1, -1\}\}. \tag{6}$$

The ultimate objective of KFE is to construct a binary classifier, which is a function f that maps the points from their feature space to their label space, i.e.,

$$\begin{aligned} f : \Re^d &\rightarrow \{+1, -1\} \\ x_k &\mapsto y_k. \end{aligned} \tag{7}$$

We need a shape descriptor or feature vector x to describe each VOP (or object mask). As mentioned in Section II, two representative shape descriptors are investigated for possible use in the proposed system: one is a moment-based descriptor which belongs to region-based descriptors, the other is Fourier descriptor which belongs to contour-based descriptors.

1) Moment-Based Descriptor: Moment-based descriptors are among the most popular region-based descriptors. Moments were first used in mechanics for purposes other than shape description. A set of moments has been shown to be invariant to translation, rotation, and scale change by Hu [12]. Based on the results from [12], a moment set can be computed to uniquely describe the shape information contained in the object mask image. The advantage of this method is that it is mathematically concise. The disadvantage is that it is difficult to correlate high-order moments with shape features. This method is also known to have low discriminatory power in the case of shape-based image retrieval system since the descriptor tends to return too many false positives.

2) Fourier Descriptor: To investigate the usefulness of the contour-based shape descriptor for the proposed system, we employed a Fourier descriptor, which is well known as a contour-based descriptor. It is computed by first resampling the contour of the VO to the fixed number of contour points, representing each contour coordinate as a complex number, and then computing the Fourier coefficients (FCs) of these complex numbers. As a result, the first FC gives the mass center of the closed contour. The second coefficient gives the radius of the circle with an area equal to that of the shape. Scaling the remaining FCs with this coefficient results in the scale-invariant descriptor. Since the orientation and starting point of the contour affect only the

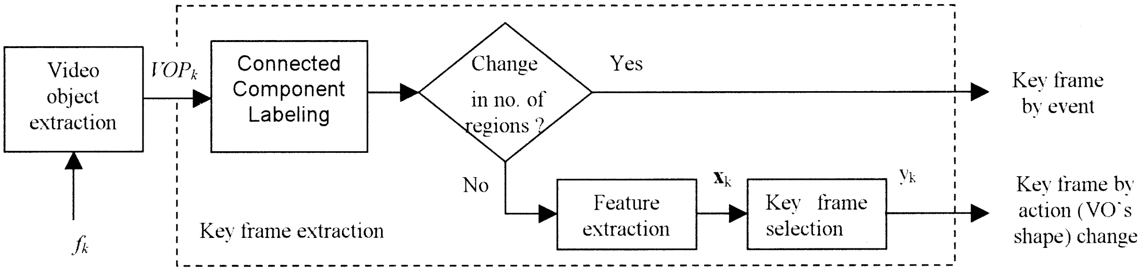


Fig. 6. Block diagram for integrated system for object-based KFE.

phases of the FCs, only the magnitude portion of these coefficients are used in the experiments. In our test, we employ the scaled magnitudes of the first eight FCs as our scale- and rotation-invariant Fourier descriptor.

In the proposed system, the first frame is always chosen as a key frame. The distance $D(F_{\text{last}}, F_k)$ needs to be defined between the current frame, F_k and the last extracted key frame F_{last}. The dissimilarity index of two frames is F_i; F_j is defined as

$$D(F_i, F_j) = d(O_i, O_j) \tag{8}$$

where $d(\cdot, \cdot)$ measures the dissimilarity between two object masks (or VOPs) O_i and O_j, which are extracted from frame F_i and F_j, respectively. We compute the distance (weighted L1 norm) by comparing the shape similarity between two object masks (VOPs) more specifically

$$d(O_i, O_j) = \sum_l w_l |x_i(l) - x_j(l)| \tag{9}$$

where the weighting constant w_l is an inverse of standard deviation σ_l. The weighting constants are computed from each feature element of manually generated test data set (object masks), in which various object shapes are contained. If this difference exceeds a given threshold T_d, the current frame is selected as a new key frame. This is shown as follows.

Step1: $\forall k \in [1, K], y_k = -1$
Step2: $y_1 = +1, \text{last} = 1, k = 1$
Step3: $k = k + 1$, *if* $D(F_{\text{last}}, F_k) > T_d$ **and** $k - \text{last} > T_f$
  then $y_k = +1, \text{last} = k$
Step4: Repeat step3 until $k = K$.

Note that K is the number of frames within a shot and y_k denotes a class label of F_k. We employ T_f to avoid successive selection in highly active frames. In our formulation, T_f is set to be a third of the frame rate.

Here, it is assumed that only a single VO exists in a scene. In Section IV-B, the algorithm is further extended and generalized to include the case that multiple objects exist in a scene.

B. Extension to the Case of a Multiple VO Sequence

When *online* processing is required, such as those encountered in object-based video surveillance, multiple VOs may exist in a scene. Therefore, a VOS algorithm that can detect multiple objects and an efficient key VOP selection method should be provided to deal with this scenario. To select key frames from a shot which contains one or more VOs, two criteria are exploited: one is event driven and the other is action change driven. Fig. 6 illustrates the proposed framework integrated with an efficient VOS algorithm.

1) Key Frames by Event: The first criterion is based on the change of the number of regions between the last declared key frame and the current frame. There are two cases for change in the number of regions. If the number increases, it implies either a new appearance of one or more objects or segregation of two or more overlapped objects. If the number decreases, this implies either disappearance of one or more objects, or overlap of two or more objects. In either case, we declare the frame as a new key frame, assuming an important event occurs. For this decision, connected components labeling [21] are first conducted to label separate regions. Specifically, we adopt a row-by-row labeling algorithm, which makes two passes over the image: 1) the first pass, to record equivalence and assign temporary labels and 2) the second pass, to replace each temporary label by the label of its equivalence class. In between the two passes, the recorded set of equivalence, stored as a binary relation, is processed to determine the equivalence classes of the relation. To avoid false regions due to noise, the regions smaller than fixed size is not labeled and ignored in the following steps. These labeled regions are also used in the KFE by action change, which will be explained in the following.

2) Key Frames by Action (or Shape) Change: As long as the number of labeled regions in the last selected key frame and that in the current frame are the same, the KFE problem is modeled as choosing a compact set of samples (key frames) given feature vectors from sequential frames. We assume that we have a data set δ of K frames in an n-dimensional feature space belonging to two different classes $+1$ (=key frames) or -1 (=nonkey frames)

$$\Delta = \{(\mathbf{x}_{k,m}, y_k) | k \in \{1, \ldots, K\}, m \in \{1, \ldots, M\} \\ \mathbf{x}_{k,m} \in \Re^n, y_k \in \{+1, -1\}\} \tag{10}$$

where K denotes the number of sequential frames in which the number of labeled regions maintains M.

Our classification is based on the distance measure between two frames. If two frames are denoted as F_i and F_j, and they contain the same number of labeled regions as $R_i = \{r_{i,m}, m = 1, \ldots M\}$ and $R_j = \{r_{j,m}, m = 1, \ldots M\}$, then the distance between these two frames can be defined as

$$D(F_i, F_j) = \max[d(r_{i,1}, r_{j,\text{match}_{ij}(1)}), d(r_{i,2}, r_{j,\text{match}_{ij}(2)}), \ldots \\ d(r_{i,M}, r_{j,\text{match}_{ij}(M)})] \tag{11}$$

where $\mathrm{match}_{ij}(m)$ denotes the spatially closest labeled region in F_j for the mth labeled region in F_i. We take city block distance [22] between center points of two regions to measure spatial closeness. Note that the assumption that two corresponding regions are spatially closest holds as long as the number of regions retains same. We define a distance (Mahalanobis distance) between two regions as

$$d(r_{i,m}, r_{j,n}) = \sum_{l} w_l |x_{i,m}(l) - x_{j,n}(l)| \tag{12}$$

where w_l is the weighting constant and $\mathbf{x}_{i,m}$ is a shape feature vector extracted from $r_{i,m}$. Two representative shape descriptors are tested in the next section. If the distance $D(F_{\mathrm{last}}, F_k)$ between the last selected key frame and the current frame is greater than predefined threshold value, we recognize the existence of a different action or shape change from the last key frame.

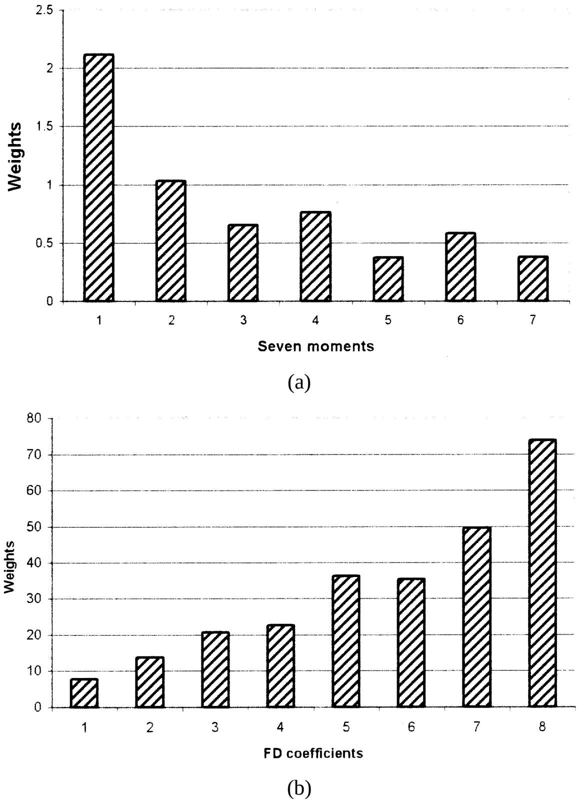
Details of our object-based key-frame selection method using two criteria abovementioned is given as follows. The first frame in a shot is always chosen as a key frame. Then, the numbers of objects are computed for the current frame F_k and the last extracted key frame F_{last}. If the numbers are different from each other, the frame F_k is declared as a key frame, assuming an event occurs; otherwise, the distance $D(F_{\mathrm{last}}, F_k)$ is computed between the current frame F_k and the last extracted key frame F_{last}. If this difference exceeds a given threshold T_d, the current frame is selected as a new key frame, as follows.

```
Step1: $\forall k \in [1, K], y_k = -1$
Step2: $y_1 = +1, \mathrm{last} = 1, k = 1$
Step3: $k = k + 1,$
         if $n_k \neq n_{k-1}$ and $k - \mathrm{last} > T_f$ then $y_k = +1, \mathrm{last} = k$
         otherwise
             if $D(F_{\mathrm{last}}, F_k) > T_d$ and $k - \mathrm{last} > T_f$
             then $y_k = +1, \mathrm{last} = k$
Step4: Repeat step3 until $k = K$.
```

Here, K is the number of frames within a shot and n_k is the number of labeled regions in F_k. Note that the scheme introduced in Section IV-A becomes a subset of this generalized algorithm with slight modification.

C. Experiments for Selection of Appropriate Shape Descriptor

Here, we present experimental results on the proposed object-based key frame selection scheme, investigated with two different shape descriptors. Two MPEG-4 test sequences were used: one is "Hall Monitor", a surveillance video sequence containing small moving objects and a complex background in CIF format, and the other is "Bream", a video sequence which contains a fish swimming and turning in QCIF format. The fish's shape change is quite large due to its drastic nonrigid motion. In order to evaluate the performance of the proposed KFE scheme, the ground-truth object masks were used. Two VOs (VO_A and VO_B) of the "Hall Monitor" sequence were manually extracted, whereas the alpha planes were used for the "Bream" sequence. In our experiments, regions smaller than 300 pixels are regarded as noise, and are ignored.



(a)

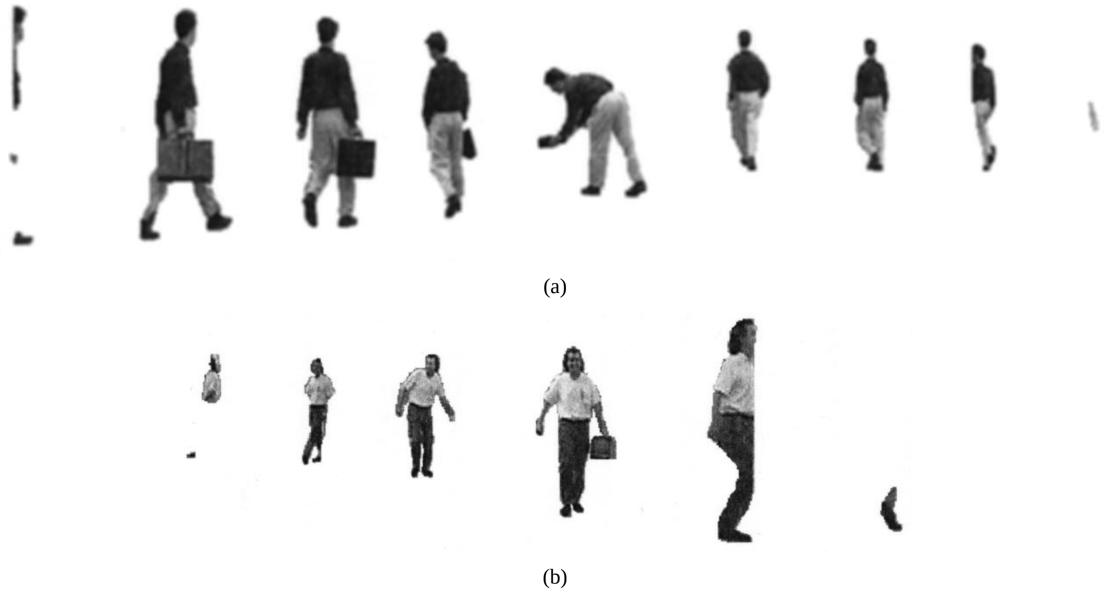
(b)

Fig. 7. Weight vectors (inverse of SD) used for (9) and (12) in these experiments. Weight vector for (a) seven moments and (b) FCs.

For the experiments using a region-based descriptor, the seven-dimensional feature vector for each labeled region is generated using Hu's seven moments [12], [22]. In order to reduce the range of values, the ***log*** of seven moments was used. The threshold values for the distance measure were set to be 5.0 for T_d and 10 for T_f. When using the Fourier descriptor as a feature vector, two threshold values were set to be 8.0 and 10 for T_d and T_f, respectively. The object's boundaries are resampled by 64 points. As a feature vector, the scaled magnitudes of the first eight FCs are calculated from the boundaries of the object masks.

Weight vectors for both seven moments and FCs [see Fig. 7(a) and (b)] are selected by taking inverse of standard deviation for each moment or coefficient. These are calculated from an arbitrarily generated data set where various object shapes are contained, and are used in every experiment without change.

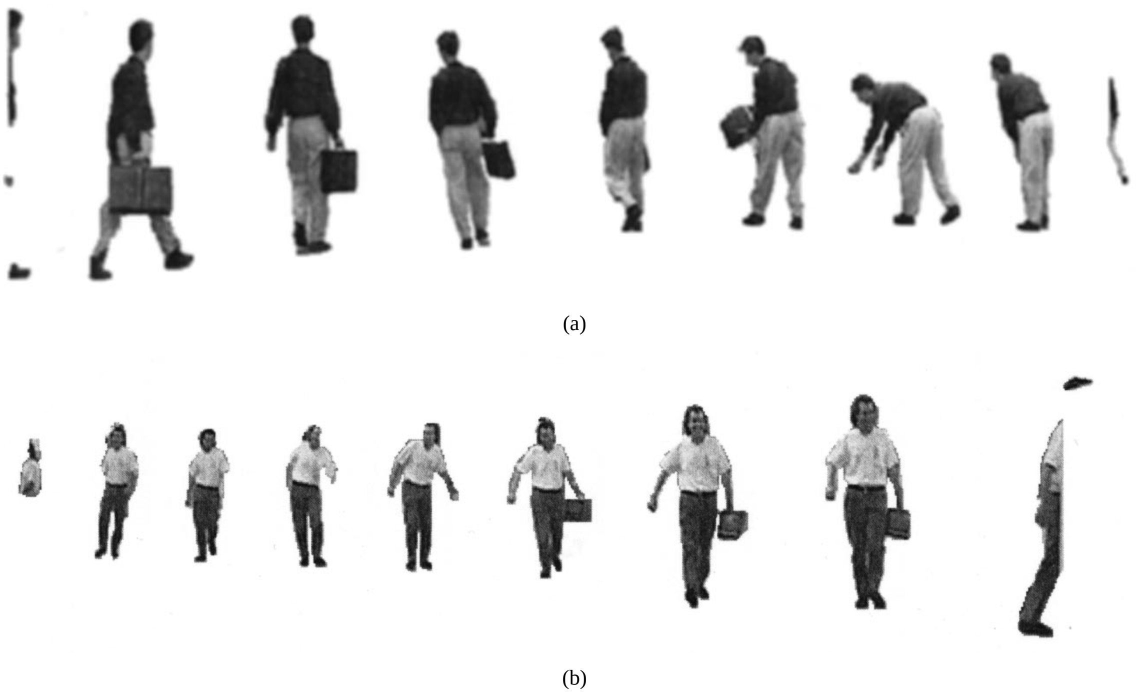
Figs. 8(a) and (b) show the experimental results applied to VO_A and VO_B, repectively, for Hall Monitor using region-based shape descriptor. As shown in Fig. 8, the proposed algorithm using the region-based shape descriptor provides a concise and effective summarization of each VO. For example, it reports important instances of each VO, such as appearance, walking, turning, bending forward, standing back, walking, and disappearance. Fig. 9 shows the results using contour-based descriptors with parameters that generate the best summarization results. As shown in Fig. 9(a) and (b), the abstraction results tends to be sensitive to minor deformation along the boundary of the object, such as the gait of the moving person and segregation of two arms, generating some redundant VOPs, whereas it misses some instances such as walking instances after the



(a)

(b)

Fig. 8. Selected key VOPs from the ground truth VOPs in "Hall Monitor" using region-based descriptor $(T_d = 5,\ T_f = 10)$. (a) Selected key VOPs for VO_A. Frame number: 16, 27, 49, 80, 112, 197, 221, 239, and 250 (left to right). (b) Selected key VOPs for VO_B. Frame number: 81, 92, 173, 242, 307, and 318 (left to right).



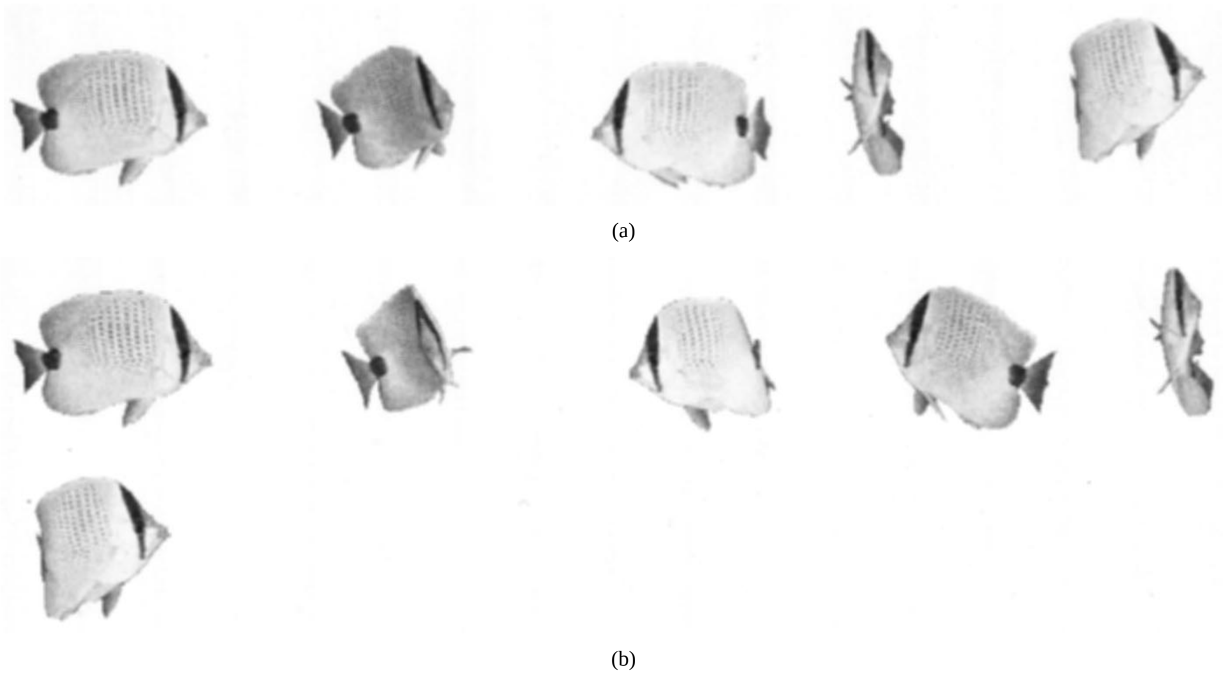
(a)

(b)

Fig. 9. Selected key VOPs from the ground truth VOPs in "Hall Monitor" using contour-based descriptor $(T_d = 8, T_f = 10)$. (a) Selected key VOPs for VO_A. Frame number: 16, 27, 49, 80, 112, 197, 221, 239, and 250 (left to right). (b) Selected key VOPs for VO_B. Frame number: 81, 92, 173, 242, 307, and 318 (left to right).

VO_A bends over. Another test results using the region-based descriptor on the sequence "Bream"' is also shown in Fig. 10. Note that the same parameters used in "Hall Monitor" were applied in Fig. 10(a). The abstract represents crucial VOPs denoting turning instances, as well as steady states before/after turning actions. Unlike the region-based descriptor case, only three key VOPs, i.e., frames 0, 116, and 144, were selected in the case of using a contour-based descriptor, where the same parameters used in the test on Hall Monitor were employed. In order to have better abstract of the sequence, the threshold Td was changed to be five, and the results are shown in Fig. 10(b).

Therefore, we observe that the region-based descriptor is more suitable for the proposed system in terms of its robustness to minor shape deformation and parameter setting. To support our observation, we further conducted a simple test on four VOPs from VO_A of Hall Monitor (see Fig. 11). We define two distance ratios $A = d(O_{109}, O_{126})/d(O_{109}, O_{113})$ and $B = d(O_{109}, O_{140})/d(O_{109}, O_{113})$. Since VOP$_{109}$, VOP$_{113}$,


(a)

(b)

Fig. 10. Selected key VOPs from the ground truth VOPs in "Bream" sequence. (a) Key VOPs using region-based descriptor $(T_d = 5, T_f = 10)$. Frame number: 0, 110, 133, 223, 234. (b) Key VOPs using contour-based descriptor $(T_d = 5, T_f = 10)$. Frame number: 0, 113, 124, 193, 223, 234.

Fig. 11. VOPs used on a test (VOP$_{109}$, VOP$_{113}$, VOP$_{126}$, VOP$_{140}$ from left to right). VOP$_{109}$, VOP$_{113}$, and VOP$_{126}$ represent bending instances, while VOP$_{140}$ shows an almost standing body. Note that VOP$_{126}$, which has separate arms, results in a large distance from VOP 109 in the case of the Fourier descriptor.

TABLE I
DISTANCE MEASURES FOR DISTANCE RATIO TEST

Distance measure	$d(O_{109}, O_{113})$	$d(O_{109}, O_{126})$	$d(O_{109}, O_{140})$	A	B	B/A
FC	1.9909	10.2289	12.4765	5.1378	6.2668	1.2197
7M	0.3418	0.6457	1.4601	1.8891	4.2718	2.2261

and VOP$_{126}$ represent bending instances, while VOP$_{140}$ denotes almost standing up, it is expected to select a descriptor that has a large ratio between A and B, or B/A. Due to separated arms in VOP$_{126}$, the distance $d(O_{109}, O_{126})$ using FC generates fairly large value compared to $d(O_{109}, O_{113})$, resulting in $A_{\rm FC} = 5.1378$, whereas $A_{\rm 7M} = 1.8891$ in the case of seven moments (7 M) is used as a descriptor. As a result, $B_{\rm 7M}/A_{\rm 7M}$ is larger than $B_{\rm FC}/A_{\rm FC}$.

From the above experiments, we concluded that the region-based descriptor is more suitable for the proposed system, since a contour-based descriptor tends to be too sensitive for slight changes along the object boundary. On the contrary, this can be a good aspect in a shape-based image retrieval system.

V. SYSTEM INTEGRATION

With the region-based descriptor selected for the proposed system, we implemented the whole integrated system as shown in Fig. 6. We present experimental results on the MPEG-4 test sequence "Hall Monitor". In order to evaluate the performance of the generalized algorithm proposed in Section IV-B, first, the ground-truth object images are tested, which are binary images containing both VOs. In the feature-extraction stage, Hu's seven moments are calculated from each ground-truth object. Weighting vector $\mathbf{w}$ and threshold values were the same as those used in Section IV-C. The experimental results are shown in Fig. 12, which reports important instances in the sequence, such as birth (appearance) and death (disappearance) of two objects, as well as distinguishable action changes, such as different walking or turning scenes, and bending scenes to put/take something. Note that, in the VOP labeling stage, the regions smaller than a size of 300 pixels are regarded as noise and are ignored; those are also shown in Fig. 12.

Now that we are convinced that the generalized algorithm for multiple VOs can perform well, the online system is integrated and tested. Fig. 13 shows the experimental results applied to our integrated system. Note that the proposed automatic segmentation scheme has been integrated for this experiment. For VOP extraction, we used morphological closing with a rectangular structuring element 9×23 followed by 5×7 size of opening.

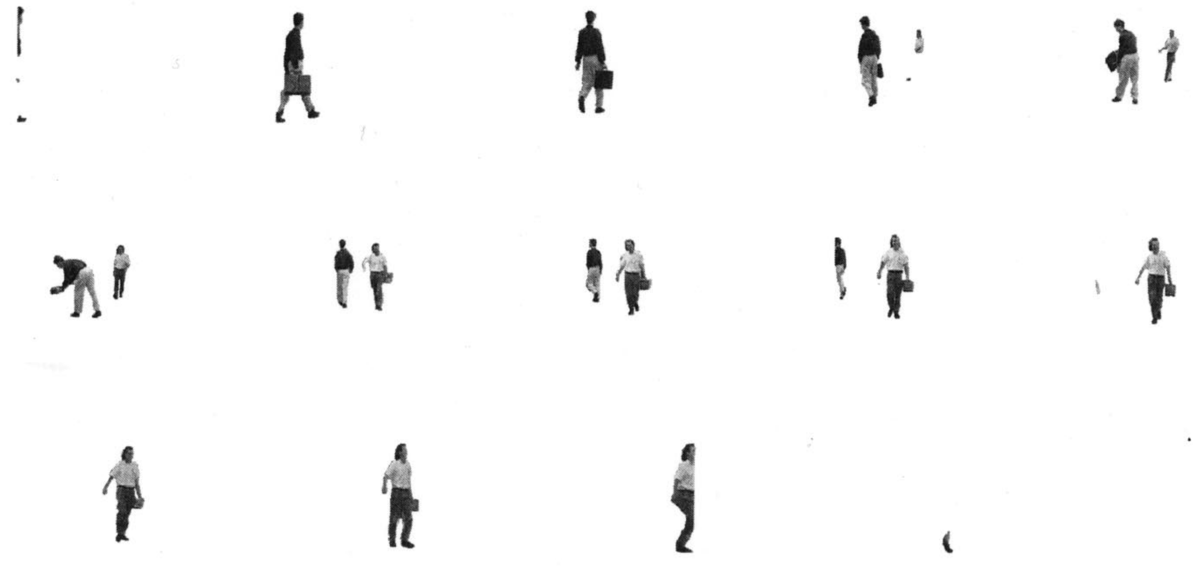


Fig. 12. Extracted key frames from ground truths. Numbers denote the key frame numbers in raster scan order: $\mathbf{16}, 27, 49, 80, \mathbf{91}, 111, 193, 220, 239, \mathbf{250}, 274, 292, 307, \mathbf{318}$. Numbers in bold denote key frames by event. Note that regions smaller than a predefined size (300 pixels) are ignored in the KFE stage, but are shown in the images.

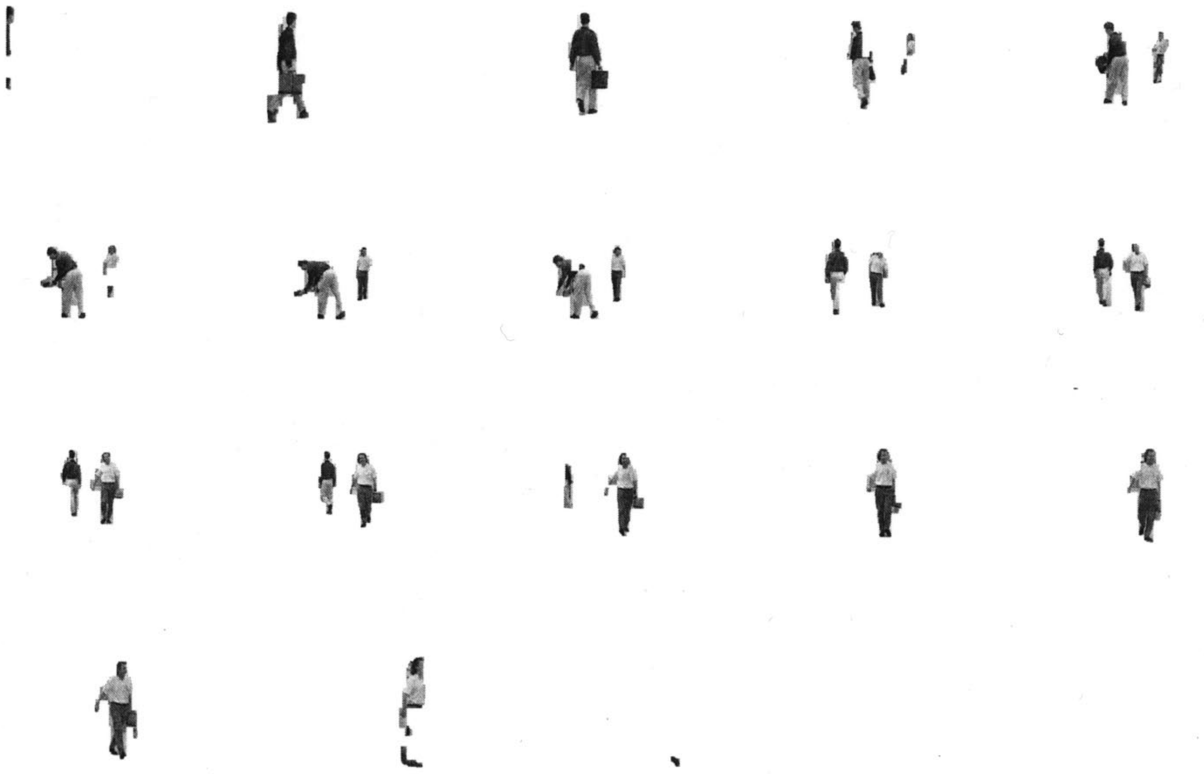


Fig. 13. Extracted key frames generated by the integrated system. Numbers denote the frame numbers in raster scan order: $\mathbf{17}, 28, 51, \mathbf{80}, 91, 102, 113, 126, 162, 197, 208, 219, 245, \mathbf{256}, 269, 280, \mathbf{306}, \mathbf{317}$. Numbers in bold denote key frames by event. Note that some regions smaller than a predefined size (300 pixels) are ignored in the KFE stage, but are shown in the images.

Closing helps to achieve spatial continuity, while opening suppresses small false regions. The abstraction results show little difference from Fig. 12, capturing important events and shape changes, although some object boundaries are not as smooth as those of the ground truths.

In addition to the above MPEG4 test sequences, an indoor sequence captured in the office was also tested. The threshold values for the distance measure were the same as those used in the previous experiments. A person with a bag enters the scene and bends forward to put the bag down, and then exits the scene. Unlike the "Hall Monitor" sequence, this sequence was captured by a low-end PC camera and contains a dark shadow under the moving person's legs. It contains relatively large moving objects in a scene; thus, the bag is clearly detected even after the

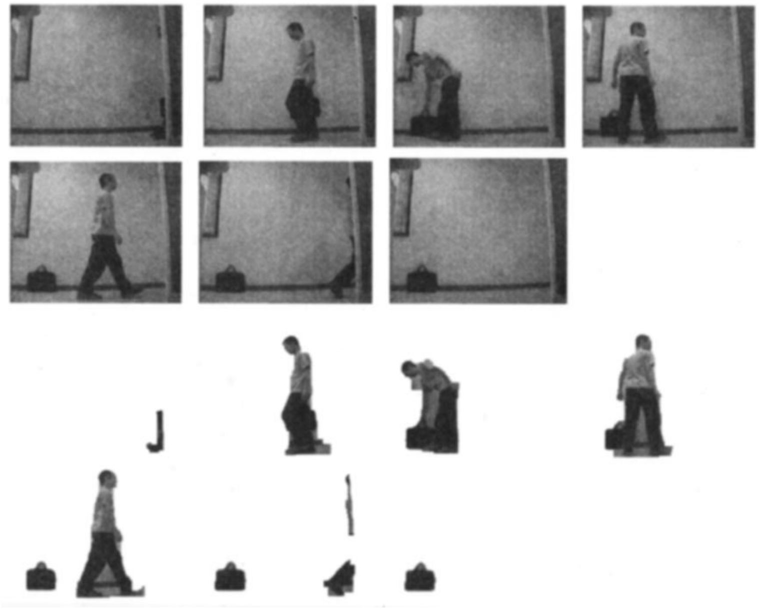


Fig. 14. Extracted key frames from indoor scene. Note that some regions smaller than a predefined size (300 pixels) are ignored in the KFE stage.

person has left the scene. The abstract in Fig. 14 shows significant instances, such as walking, bending, turning and splitting into two objects, as well as entering and leaving.

VI. Conclusion

We have shown an integrated scheme for object-based video abstraction. The contributions and characteristics of the proposed scheme are summarized as the following.

- *Efficiency*: Easy to implement and fast in computation.
- *Effectiveness:* Able to capture the salient contents based on observations made on objects.
- *Online processing:* Easy to implement online, since it depends only on the last selected key frame and the current frame.
- *Open framework:* In this paper, we used a moment-based descriptor for shape description. A combination with any useful features is possible. Some feasible low-level features to describe an object are color, texture, shape, spatial relationship, and motion. A performance study for each feature should be conducted in order to find crucial features.

In this paper, we have tested two shape descriptors to adopt a suitable feature vector for our object-based KFE algorithm: one is Hu's seven moments, belonging to region-based descriptors, and the other is FCs, belonging to contour-based descriptors. The former is found to be suitable for the proposed abstraction scheme, whereas the latter is expected to be more useful for a shape-based image retrieval system, as shown in previous literature [15].

The proposed the KFE scheme, combined with an efficient VOS system, generates desirable summarization of the video sequences containing one or more VOs. It can be easily applied to an intelligent video security system, since appearance/disappearance of objects, as well as their significant changes in shape, can be automatically detected.

Acknowledgment

The authors appreciate the help from Dr. H. Zhang and the anonymous reviewers of this paper for their valuable and constructive comments.

References

[1] T. Sikora, "The MPEG-4 video standard verification model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 19–31, Feb. 1997.

[2] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recog.*, vol. 30, no. 4, pp. 643–658, 1997.

[3] A.M Ferman, B. Gunsel, and A. M. Tekalp, "Object-based indexing of MPEG-4 compressed video," in *Proc. SPIE-3024*, San Jose, CA, Feb. 1997, pp. 953–963.

[4] B. Erol and F. Kossentini, "Automatic key video object plane selection using the shape information in the MPEG-4 compressed domain," *IEEE Trans. Multimedia*, vol. 2, pp. 129–138, June 2000.

[5] C. Kim and J.-N. Hwang, "An integrated scheme for object-based video abstraction," in *Proc. ACM Int. Multimedia Conf.*, Los Angeles, CA, Oct. 2000, pp. 303–311.

[6] T. Aach, A. Kaup, and R. Mester, "Statistical model-based change detection in moving video," *Signal Processing*, vol. 31, no. 2, pp. 165–180, Mar. 1993.

[7] R. Mech and M. Wollborn, "A noise robust method for segmentation of moving objects in video sequences," in *Proc. ICASSP'97*, vol. 4, Apr. 1997, pp. 2657–2660.

[8] A. Neri, S. Colonnese, G. Russo, and P. Talone, "Automatic moving object and background separation," *Signal Processing*, vol. 66, pp. 219–232, 1998.

[9] C. Gu and M-C Lee, "Semantic segmentation and tracking of semantic video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 572–584, Sept. 1998.

[10] C. Kim and J.-N. Hwang, "Fast and robust moving object segmentation in video sequences," in *Proc. IEEE Int. Conf. Image Processing (ICIP'99)*, vol. 2, Kobe, Japan, Oct. 1999, pp. 131–134.

[11] C. Kim and J.-N. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, pp. 122–129, Feb. 2002.

[12] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inform. Theory*, vol. IT-8, no. 2, pp. 179–182, Feb. 1962.

[13] A. Khotanzad and Y.H. Hong, "Invariant image recognition by zerike moments," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 489–498, May 1990.

[14] W. Kim and Y. Kim, “A region-based shape descriptor using zernike moments,” *Signal Processing: Image Commun.*, vol. 16, pp. 95–102, 2000.

[15] M. Bober, “MPEG-7 Visual Shape Descriptors,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 716–719, June 2001.

[16] S. Loncaric, “A survey of shape analysis techniques,” *Pattern Recog.*, vol. 31, no. 8, pp. 983–1001, 1998.

[17] W. E. Grimson, *From Images to Surfaces*. Cambridge, MA: MIT Press, 1981, pp. 3–5.

[18] J. F. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, pp. 679–698, Nov. 1986.

[19] H. J. Zhang, C. Y. Low, and S. W. Smoliar, “Video parsing and browsing using compressed data,” *Multimedia Tools Applic.*, vol. 1, pp. 89–111, 1995.

[20] H. J. Zhang, J. H. Wu, C. Y. Low, and S. W. Smoliar, “A video parsing, indexing and retrieval system,” in *Proc. ACM Int. Conf. Multimedia*, 1995, pp. 359–360.

[21] L. Shapiro and G. Stockman, *Computer Vision*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[22] R. C. Gonzalez and P. Wintz, *Digital Image Processing*, 2nd ed. New York: Addison-Wesley, 1987, pp. 173–174.

[23] E. Durucan and T. Ebrahimi, “Change detection and background extraction by linear algebra,” in *Proc. IEEE*, vol. 89, Oct. 2001, pp. 1368–1381.

[24] H. J. Zhang, J. Y. A. Wang, and Y. Altunbasak, “Content-based video retrieval and compression: A unified solution,” in *Proc. Int. Conf. Image Processing*, vol. 1, Oct. 1997, pp. 13–16.

Changick Kim (M’01) was born in Seoul, Korea, in 1967. He received the B.S. degree in electrical engineering from Yonsei University, Seoul, Korea, in 1989, the M.S. degree in electrical engineering from Pohang University of Science and Technology, Pohang, Korea, in 1991, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2000.

Since December 2000, he has been with the Epson Palo Alto Laboratory, Palo Alto, CA, as a Senior Member of Technical Staff. His research interests include object-based video analysis, media security/management, and video coding/transmission over wireless channels.

Jenq-Neng Hwang (F’01) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1981 and 1983, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in December 1988.

After two years of obligatory military service, in 1985 he joined the Signal and Image Processing Institute, Department of Electrical Engineering, University of Southern California, as a Research Assistant. He was also a Visiting Student at Princeton University, Princeton, NJ, from 1987 to 1989. In the summer of 1989, he joined the Department of Electrical Engineering, University of Washington, Seattle, where he is currently a Professor. He has published more than 180 journal and conference papers and book chapters in the areas of image/video signal processing, computational neural networks, multimedia system integration, and networking. He is the co-author of the book *Handbook of Neural Networks for Signal Processing* (Boca Raton, FL: CRC Press, 2001).

Dr. Hwang served as Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE TRANSACTIONS ON NEURAL NETWORKS, and is currently an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He was a Guest Editor for the IEEE TRANSACTIONS ON MULTIMEDIA (Special Issue on Multimedia over IP) in March/June 2001. He is also on the Editorial Board of the *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology.* He served as the Secretary of the Neural Systems and Applications Committee of the IEEE Circuits and Systems Society from 1989 to 1991, was a member of the IEEE Signal Processing (SP) Society’s Technical Committee on Design and Implementation of Signal Processing Systems, and was a Founding Member of the Multimedia Signal Processing Technical Committee of the IEEE SP Society. He served as the Chairman of the Technical Committee of the IEEE Neural Networks for Signal Processing of the SP Society from 1996 to 1998, and was the Society’s representative to the IEEE Neural Network Council from 1997 to 2000. He was the Conference Program Chair of the 1994 IEEE Workshop on Neural Networks for Signal Processing held in Ermioni, Greece, in 1994, the General Co-Chair of the International Symposium on Artificial Neural Networks held in Hsinchu, Taiwan, R.O.C., in December 1995, Chair of the Tutorial Committee for the IEEE International Conference on Neural Networks (ICNN’96) held in Washington, DC, in June 1996, and the Program Co-Chair of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) held in Seattle, WA, in 1998. He received the 1995 IEEE Signal Processing Society’s Annual Best Paper Award (with Shyh-Rong Lay and Alan Lippman) in the area of Neural Networks for Signal Processing.