An Integrated Scheme for Object-based Video Abstraction

Changick Kim and Jenq-Neng Hwang
Information processing Laboratory
Dept. of Electrical Engineering, Box#352500
University of Washington, Seattle, WA 98195
{cikim, hwang}@ee.washington.edu

ABSTRACT

In this paper, we present a novel scheme for object-based key-frame extraction facilitated by an efficient video object segmentation system. Key-frames are the subset of still images which best represent the content of a video sequence in an abstracted manner. Thus, key-frame based video abstraction transforms an entire video clip to a small number of representative images. The challenge is that the extraction of key-frames needs to be automated and context dependent so that they maintain the important contents of the video while remove all redundancy. Among various semantic primitives of video, objects of interest along with their actions and generated events can play an important role in some applications such as object-based video surveillance system. Furthermore, on-line processing combined with fast and robust video object segmentation is crucial for real-time applications to report unwanted action or event as soon as it happens. Experimental results on the proposed scheme for object-based video abstraction are presented.

Keywords

Video abstraction, object-based key frame extraction, video object segmentation, MPEG-4/MPEG-7.

1. INTRODUCTION

The 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 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, which emphasizes to partition the video frames to semantically meaningful video objects and background, becomes an important issue for successful use of MPEG-4/MPEG-7. As an example in the MPEG-7, segmented results based on the frame-to-frame motion information or abrupt shape change can be utilized for a high-level (object-level) description .

Most of the traditional key frame extraction 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 [4]. 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 attempt to object-based key frame extraction has been reported in [7], where the ratio of the number of I-macroblocks (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 scheme requires an MPEG-4 encoder as well as pre-segmented VOPs and the accuracy of using the ratio of I-MBs is too low to be effective. In this paper, we extended previous work on VOP extraction [17] and propose a new object-based framework for video abstraction, where changes in contents are detected through observations made on the objects in the video sequence. Naturally, the efficient video object segmentation scheme is necessary for successful object-based key frame extraction.

This paper is organized as follows. In Section 2, an efficient video object segmentation algorithm is described as the first step of the object-based video abstraction system. The object-based key frame extraction algorithm using distance measure is introduced in Section 3, promising simulation results are also reported in this section. Conclusion is followed in Section 4.

2. VIDEO OBJECT SEGMENTATION

Many of video segmentation algorithms that specifically address VOP generation have been recently proposed due to the development of the new video coding standard, MPEG-4 [1]. The change detection for inter-frame difference is one of the most popular video segmentation schemes [5][6][9] because it is straightforward and enables automatic detection of new appearance. While the algorithms, which are based on inter-frame change detection, enable automatic detection of objects and allow larger non-rigid motion comparing to object tracking methods [10], the drawback is the noise (small false regions) created by decision error. Thus small holes removal using morphological operation and removal of false detection parts like uncovered background by motion information are usually incorporated [6][9][16]. Another drawback of change detection is that object boundaries are 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. Nevertheless, we believe that the spatial edge information should be incorporated in the

motion detection stage to simplify algorithm and generate better results. A desirable video object segmentation for real-time object-based applications should meet the following criteria.

- Segmented objects should conform to human perception, i.e., semantically meaningful objects should be segmented.
- Segmentation algorithm should be efficient and achieve fast speed.
- Initialization should be simple and easy for user to operate (human intervention should be minimized).

2.1. Extraction of Moving Edge (*ME*) Map

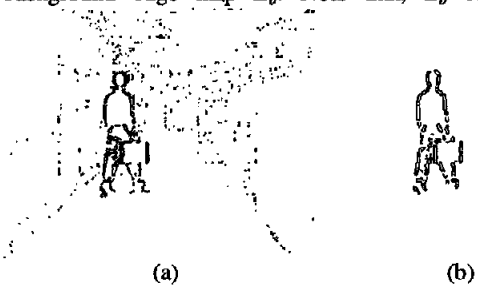
Our segmentation algorithm [17] starts with edge detection which is the first and most important stage of human visual processing as discovered by Marr and *et al.* [2]. While edge information plays a key role in extracting the physical change of the corresponding surface in a real scene, exploiting 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 one created in the successive frame. The difference of edges is defined as

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

where the edge maps $\Phi(f)$ are obtained by the Canny edge detector [8], which is accomplished by performing a gradient operation ∇ on the Gaussian convoluted image $G*f$, followed by applying the non-maximum 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 (see Eq. (2)) 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.

$$DE_n = \Phi(\left|f_{n-1}-f_n\right|) = \theta(\nabla G * \left|f_{n-1}-f_n\right|). \quad (2)$$

Fig. 2 shows the block diagram of our segmentation algorithm. After calculating edge map of difference of images using Canny edge detector (see Fig.1-(b)), we extract the moving edge ME_n of the current frame f_n based on the edge map DE_n of difference $\left|f_{n-1}-f_n\right|$, 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 in case of a still camera and is set at the first stage to increase the extraction performance. For video surveillance sequences, such as 'Hall Monitor', which contain no moving objects in the beginning of video clip, 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 moving edges of target objects (see Fig. 3). We define the edge model $E_n = \{e_1, \cdots, e_k\}$ as a set of all edge points detected by the Canny operator in the current frame n. Similarly, we denote $ME_n = \{m_1, \cdots, m_l\}$ the set of l moving edge points, where $l \le k$. The edge points in ME_n are not restricted to 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 moving edge model generated by edge change is given by selecting all edge pixels within a small distance T_{change} of DE_n, i.e.,



(a) (b)

Fig. 1: Edge maps resulted from Eq. (1) and (2)

$$ME_n^{change} = \left\{ e \in E_n \mid \min_{x \in DE_n} \|e-x\| \le T_{change} \right\}. \quad (3)$$

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

$$ME_n^{still} = \left\{ e \in E_n, e \notin E_b \mid \min_{x \in ME_{n-1}} \|e-x\| \le T_{still} \right\}. \quad (4)$$

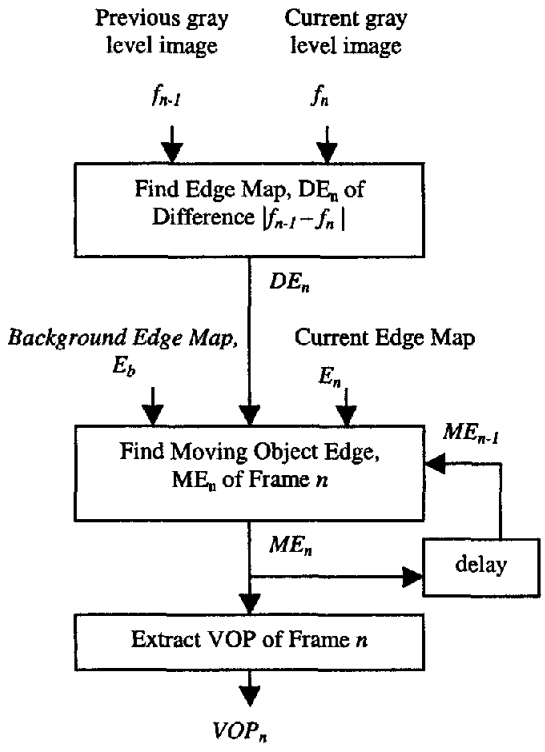


Figure 2: Block diagram of the segmentation

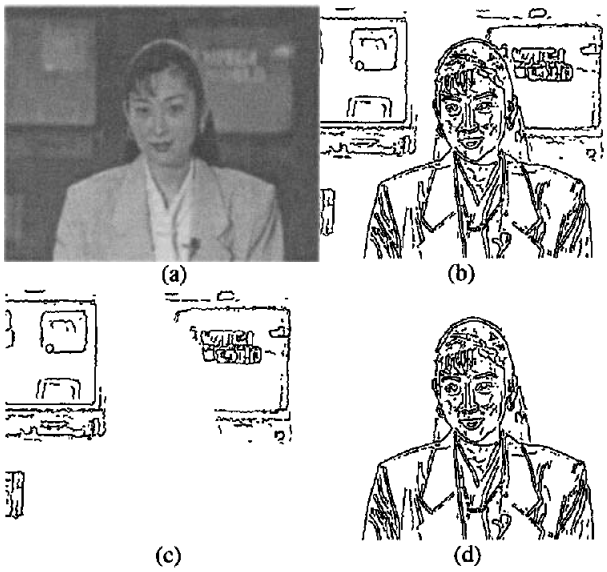


Fig.3: Finding initial edge maps: (a) The first frame, I_0 , (b) edge map of 1st frame, E_0, (c) Background edge map, E_b , (d) Moving edge map of first frame, $ME_0=E_0-E_b$

Finally, the moving edge model for current frame, f_n, is given by combining the two components.

$$ME_n = ME_n^{change} \cup ME_n^{still} . \quad (5)$$

For initial moving edge map ME_0, just a blank image is required in case of surveillance video such as 'Hall Monitor'. 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. Note that in case of little movement in the beginning frames, there is no way to detect moving edges, i.e., correct moving edges are not generated until we encounter the frame in which objects begin to move. Thus, to facilitate correct moving edge extraction, ME_0 is calculated by the equation, $ME_0 = E_0 - E_b$ (see Fig. 3-(d)). Note that such a manual initialization is not needed for the surveillance type of sequences, because moving objects will not appear in the very beginning of the sequences.

2.2 Extraction of VOP

With moving edge map ME_n, as shown in Figure 5-(a), detected from DE_n , we are ready to extract the VOPs (see Figure 4). The horizontal candidates are declared to be the region inside the first and last edge points in each row (see Fig.5-(b)) and the vertical candidates for each column. After finding both horizontal and vertical VOP candidates, the intersection regions (see Fig. 5-(c)) through logical AND operation are further processed by alternative use of morphological operations. For video sequences, such as 'Miss America', which contain only a partial moving object instead of the whole object, a rule is added to declare image boundary points as moving edge points if either of horizontal or vertical candidates touch image boundary points. This process for finding candidates is repeated to extract VOP.

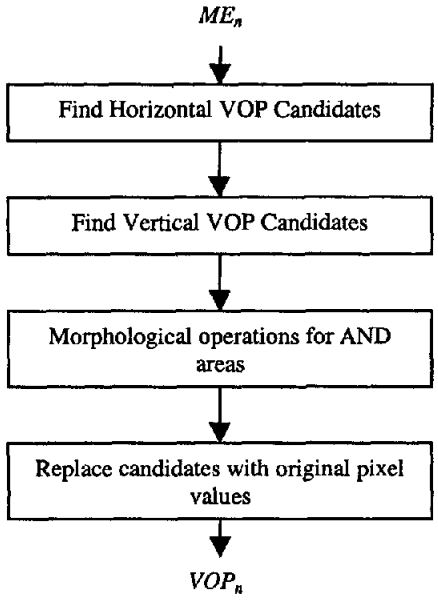


Figure 4: Block diagram of VOP extraction

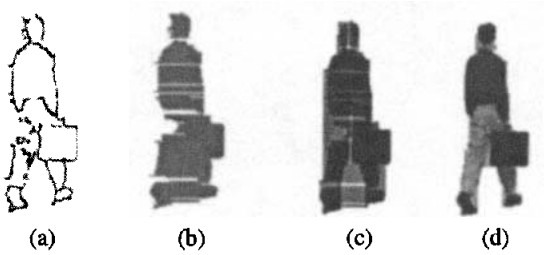


Figure 5: The VOP extraction process. (a) moving edge map, ME_{45}, (b) horizontal candidates, (c) logical AND (black areas) of horizontal and vertical candidates, (d) extracted VOP after morphological operations.

2.3 Simulation Results

In this section, the algorithmic performance of the proposed VOP extraction is illustrated through the subjective and objective evaluation .

2.3.1 Subjective Evaluation

The proposed algorithm was applied to 'Hall Monitor', which contains small moving objects and complex background in CIF format. It was also applied to 'Miss America' and 'Akiyo' sequences, which are typical head-and-shoulder type video in QCIF and CIF format respectively. Fig. 6 and Fig. 7 show several results, which show that the proposed algorithm is quite efficient on both surveillance and head-and-shoulder type sequences. Both of T_{change} and T_{still} values are 1. In order to eliminate small false regions, morphological opening can be applied. In our simulations, the morphological operations were only applied to 'Hall Monitor', which has floating noise in the background through the sequence. The entire resulting VOPs for 'Akiyo' sequence is accessible at our web site [12].

(a) (b)

(c) (d)

Fig.6: Extracted VOPs from 'Hall Monitor'
(a) fr.46 (b) fr.55 (c) fr.105 (d) fr.294

2.3.2 *Objective Evaluation of Performance*

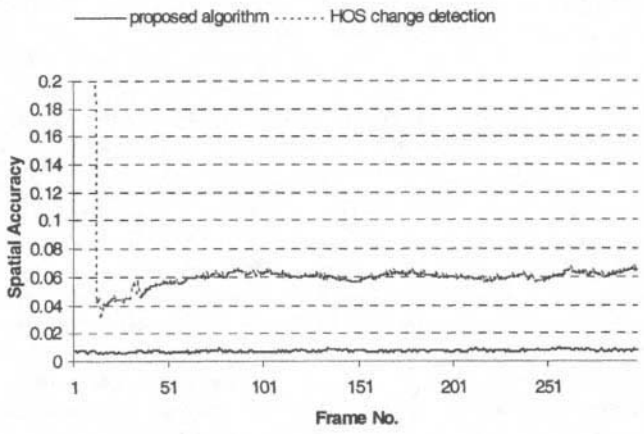
For objective evaluation of the proposed segmentation scheme, Wollborn and Mech [14] proposed a simple pixel-based quality measure. The spatial distortion of an estimated binary video object mask at frame t is defined as

$$d(O_t^{est}, O_t^{ref}) = \frac{\sum_{(x,y)} O_t^{est}(x,y) \oplus O_t^{ref}(x,y)}{\sum_{(x,y)} O_t^{ref}(x,y)}, \qquad (6)$$

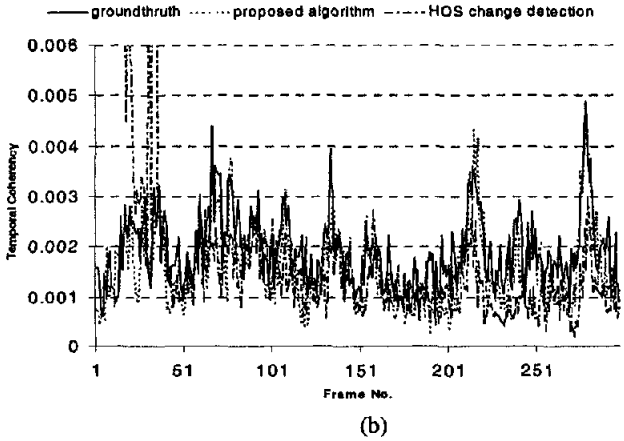
where O_t^{ref} and O_t^{est} are the reference and the estimated binary object masks at frame t, respectively, and $\oplus$ is the binary "XOR" operation. The temporal coherency $\eta(t)$ is defined by

(a) (b)

Fig. 7: Extracted VOPs from (a) Miss America (fr. 16) and (b) Akiyo (fr.148)



(a)



(b)

Figure 8 : The objective evaluation for objects from 'Akiyo' CIF sequence. (a) spatial accuracy and (b) temporal coherency

$$\eta(t) = d(O_t, O_{t-1}), \tag{7}$$

where O_t and O_{t-1} are binary object masks at frame t and t-1, respectively. Temporal coherency $\eta^{est}(t)$ of the estimated binary mask O^{est} should be compared to temporal coherency $\eta^{ref}(t)$ of the reference mask O^{ref}. Any significant deviation from the reference indicates a bad temporal coherency.

The segmentation performance of the proposed algorithm is evaluated by using both subjective and objective criteria described above. The corresponding results of the "Akiyo" sequence in CIF format are shown in Fig. 8. For the reference binary VO mask, the manually segmented mask from the original sequence is used. More specifically, edge map is taken from each frame and the background edges are manually deleted. Next, for the broken edges along the object boundary, manual linking is also performed. Finally, the inner area of an object is filled using 'paint bucket tool' in Adobe Photoshop software. In Fig. 8-(a), the dot line is obtained by using higher order statistics change detection described in [9] while the solid line is from our proposed algorithm. We see that the spatial accuracy of the proposed edge-based algorithm is much better than the inter-frame change detection scheme. The error is less than 1% in every frame. In Fig. 8-(b), the solid line denotes the reference mask while the dot line the proposed scheme. The temporal coherency curve from the proposed algorithm closely follows the curve from the reference mask. We see from these results that our moving-edge-based segmentation gives good spatial accuracy as well as temporal coherency since it provides accurate object boundary information.

3. OBJECT-BASED KEY-FRAME EXTRACTION

Object-based key-frames are extracted in a sequential manner through the sequence. Figure 9 illustrates the proposed integrated scheme for object-based key-frame extraction (KFE). It relies on two criteria to extract key-frames from a video shot.

3.1 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 numbers of regions. If the number reduces, this implies either disappearance of one or more objects, or overlap of two or more objects. If the number increases, it also implies either new appearance of one or more objects, or separation of two or more overlapped objects. In either cases, we declare the current frame as a new key-frame assuming an important event occurs. For this decision, connected components labeling [11] is first conducted to label separate regions. Specifically, we adopt a row-by-row labeling algorithm, which makes two passes over the image: the first pass to record equivalence and assign temporary labels; and 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. These labeled regions are also used in the key-frame extraction by action change which will be explained in the next subsection.

3.2 Key-Frames by Action (or Shape) Change

In case the number of labeled regions in the last selected key-frame and current frame are 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 L frames in an n-dimensional feature space belonging to two different classes +1 (=key-frames) or –1 (=non-key-frames)

$$\Delta = \left\{ (\mathbf{x}_k, y_k) \,\middle|\, k \in \{1, \cdots, L\}, \mathbf{x}_k \in \Re^n, y_k \in \{+1, -1\} \right\} \tag{8}$$

where L denotes the number of sequential frames in which the number of labeled regions is consistent.

A binary classifier that maps the points from their feature space to their label space can be used.

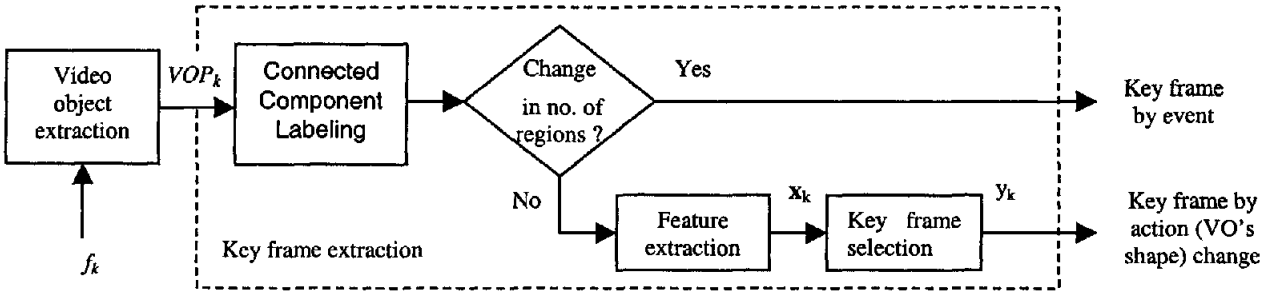


Fig.9: Block diagram for Integrated system for object-based key frame extraction

$$f : \Re^n \rightarrow \{+1,-1\}. \qquad (9)$$
$$\mathrm{x}_k \mapsto y_k$$

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 $R_i = \{r_{i,m}, m = 1, \cdots M\}$ and $R_j = \{r_{j,m}, m = 1, \cdots M\}$, then the distance between these two frames can be defined as

$$D(F_i, F_j) = \max[d(r_{i,1}, r_{j,match_{ij}(1)}), d(r_{i,2}, r_{j,match_{ij}(2)}), \cdots, d(r_{i,M}, r_{j,match_{ij}(M)})], \qquad (10)$$

where $match_{ij}(m)$ denotes the closest spatially labeled region in F_j for the *m-th* labeled region in F_i. We take city block distance [3] between center points of two regions to measure spatial closeness. We define the distance (weighted Euclidean distance) between two regions as

$$d(r_{i,m}, r_{j,n}) = \sum_p w_p \left[x_{i,m}(p) - x_{j,n}(p)\right]^2, \qquad (11)$$

where w_p is the weighting constant and $\mathbf{x}_{i,m}$ is a shape feature vector. The 7-dimensional feature vector $\mathbf{x}_{i,m}$ of the m-th labeled region in frame F_i is generated using seven Hu moments [15][3], which are known as reasonable shape descriptors in a translation- and scale-invariant manner. In order to reduce the range of values, the ***log*** of 7 moments are used.

If a distance $D(F_{last}, F_k)$ between the last selected key-frame and the current frame is greater than predefined threshould value, we regard the existence of quite different action or shape change from the last key frame. Details of our object-based key frame selection method is given as followed.

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 each other, the frame F_k is declared as a key-frame, otherwise the distance $D(F_{last}, F_k)$ 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, that is

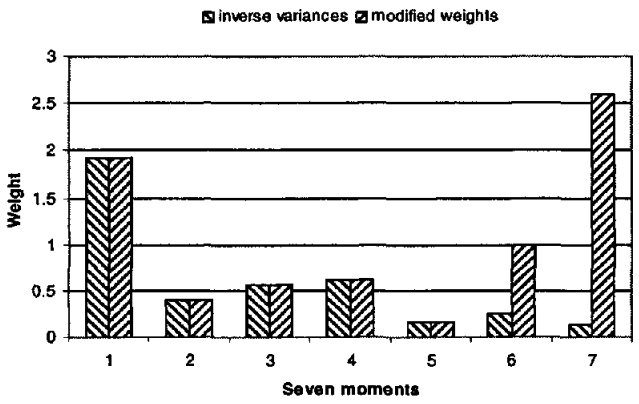


Fig. 10. Weights used for Eq. (11) in this experiments

*Step*1: $\forall k \in [1, N],\ y_k = -1$

*Step*2: $y_1 = +1, last = 1, k = 1$

*Step*3: $k = k + 1$,

if $n_k \neq n_{k-1}$ and $k - last > T_f \Rightarrow y_k = +1, last = k$

otherwise

if $D(F_{last}, F_k) > T_d$ and $k - last > T_f$

$\Rightarrow y_k = +1, last = k$

*Step*4: Iterate step3 until $k = N$.

Here, N is the number of frames within a shot and n_k is the number of labeled regions in F_k.

4. Experimental Results

In this section, we present experimental results on the MPEG-4 test sequence 'hall monitor'. For VOP extraction, we used morphological closing with structuring element (S.E.) 9×19 followed by 5×11 size of opening. The latter opening with structuring element of 5×11 suppresses small false regions. In feature extraction stage, 7 Hu moments are calculated from the ground-truth object masks manually generated for the first VO in 'hall monitor'. Weight vector **w** (see the first columns in Fig.10) is first selected by taking inverse of variances for each moment. We found the 6^{th} and 7^{th} moments are more influential to discriminate shapes from others. Thus we set weight vector with more weight on 6^{th} and 7^{th} moments (see the second column in Fig.10). Two threshold values used are 3.0 and 10 for T_d and T_f, respectively. In our labeling stage, the regions smaller than the size of 100 pixels are regarded as noise and ignored.

In order to evaluate the performance of the proposed key-frame extraction scheme, the ground-truth object masks are used, which are generated by the method explained in Subsection 2.3.2. The experimental results are shown in Figure 11, which reports important events in the sequence, such as birth (appearance) and death (disappearance) of two objects (see (a), (d), (l) and (q) in Fig. 11), and distinguishable action changes, such as different walking or turning scenes (see (b), (c), (e), (h), (i), (j), (k), (m), (n), (o) and (p) in Fig. 11), and bending scenes to put/take something (see (f) and (g) in Fig.11). Fig. 12 shows the experimental results applied to our integrated system. Note that the proposed automatic segmentation system has been used for this experiment. It shows little difference from Fig. 11, capturing important events and shape changes.

5. CONCLUSION

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

- Efficiency: Easy to implement and fast to compute.
- Effectiveness: Able to capture the salient contents based on observations made on objects.
- On-line processing: Easy to be implemented on-line since it only depends on the last selected key frame and the current frame. The novel selection scheme for weight vector in

Equation 11 is needed. Firstly, the effect of each moment should be studied carefully.

- Open framework: In this paper, we used Hu moments for similarity measure. A combination with any useful features is possible. The feasible low level features to describe an object are color, texture, shape, spatial relationship, motion, etc. Performance study for each feature will be conducted to find some crucial features.

In this paper, the terminology of 'event' and 'action' has been used in a broad sense. Eventually, our scheme should be able to discriminate semantic aspects of different roles, actions, and events for better abstractions of visual signs.

6. REFERENCES

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

[2] William E. Grimson, *From Images to Surfaces*, The MIT press, pp3-5, 1981.

[3] Rafael C. Gonzalez, Paul Wintz, *Digital Image Processing*, Addison Wesley, 2nd edition, pp173-174, 1987.

[4] H.J.Zhang et al. "An Integrated system for Content-based Video Retrieval and Browing", Pattern recognition, vol. 30, No.4, pp.643-658, 1997.

[5] Til Aach, Andre Kaup, Rudolf Mester, "Statistical model-based change detection in moving video", *Signal Processing*, Vol. 31, No. 2, pp. 165-180, March, 1993.

[6] R. Mech and M. Wollborn, "A noise robust method for segmentation of moving objects in video sequences", *ICASSP97*, Vol.4, pp.2657-2660, April 1997.

[7] A.M.Ferman *et al.*, "Object-Based Indexing of MPEG-4 Compressed Video", SPIE-3024, pp. 953-963, Feb. 1997, San Jose, CA.

[8] J.F. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6) pp.679-698, November 1986.

[9] A. Neri et al., "Automatic Moving Object and background Separation," *Signal Processing*, vol.66, pp219-232, 1998.

[10] C. Gu and M-C Lee, "Semantic Segmentation and Tracking of Semantic Video objects," *IEEE Trans. Circuits Syst. Video Technology*, vol. 8, pp.572-584, Sep. 1998.

[11] L. Shapiro, "*Computer Vision*," Prentice Hall, to be published.

[12] http://students.washington.edu/cikim/cidil/mos/mos3.html

[14] M.Wollborn and R. Mech, "Refined procedure for objective evaluation of video generation algorithms," Doc. ISO/IEC JTC1/SC29/WG11 M3448, March, 1998.

[15] M. Hu, "Visual pattern recognition by moment invariants", IRE Trans. Information Theory, IT-8(2), pp. 179-182, Feb. 1962.

[16] Ju Guo *et al.*, "Fast and accurate moving object extraction technique for MPEG-4 object-based video coding," SPIE, vol.3653, pp.1210-1221, January, 1999.

[17] Changick Kim and Jenq-Neng Hwang, "Fast and Robust Moving Object Segmentation in Video Sequences," *IEEE international conference on Image Processing* (ICIP'99), Kobe, Japan, Oct. 1999.

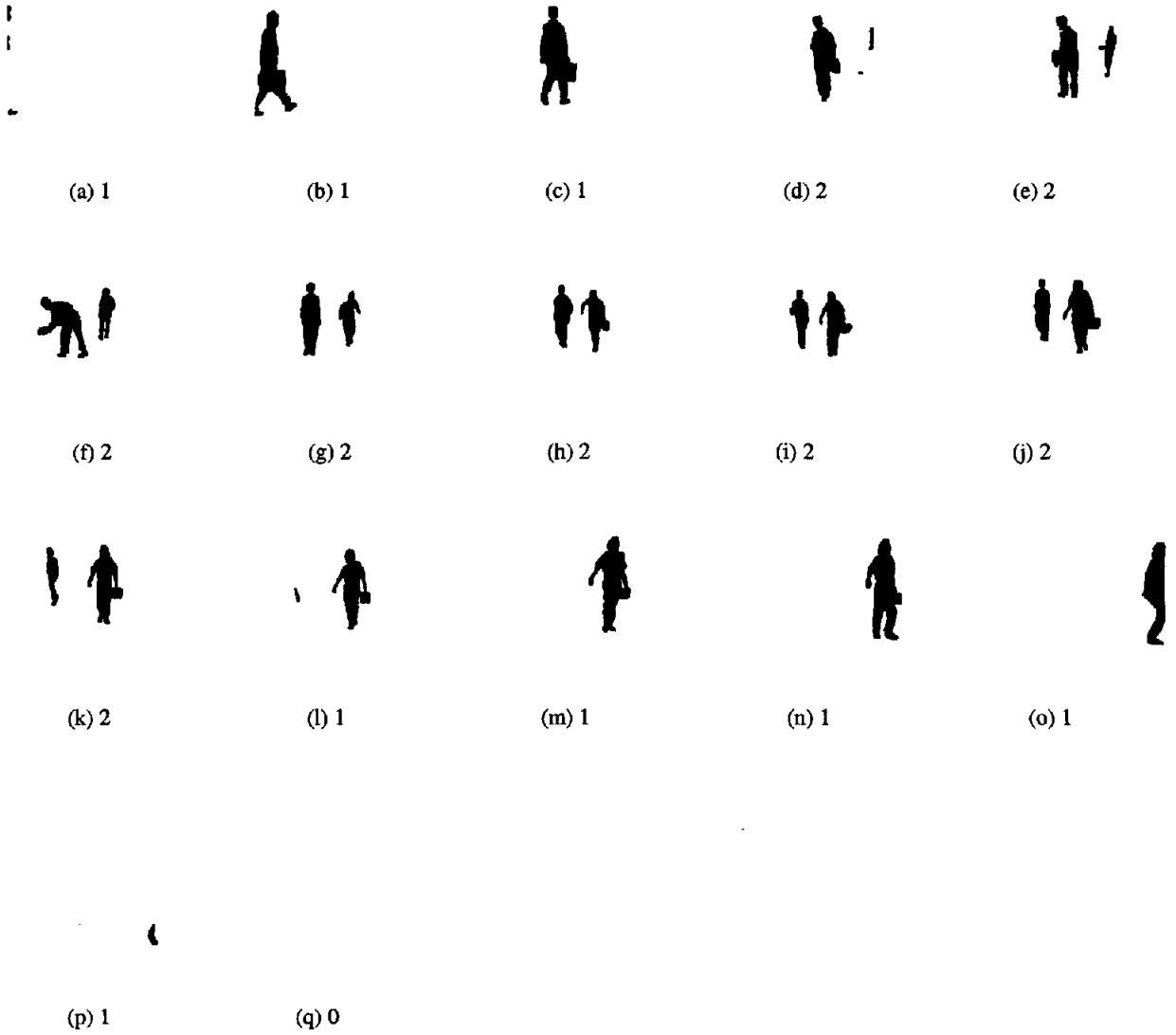


Fig.11: Extracted key frames from ground truths. Frame no. (a)15, (b)26, (c)48, (d)77, (e)88, (f)111, (g)151, (h)197, (i)210, (j)221, (k)239, (l)250, (m)274, (n)292, (o)307, (p)318, (q)329. Numbers denote the number of labeled regions in the frame. Note that some regions smaller than a predefined size are ignored in the KFE stage.

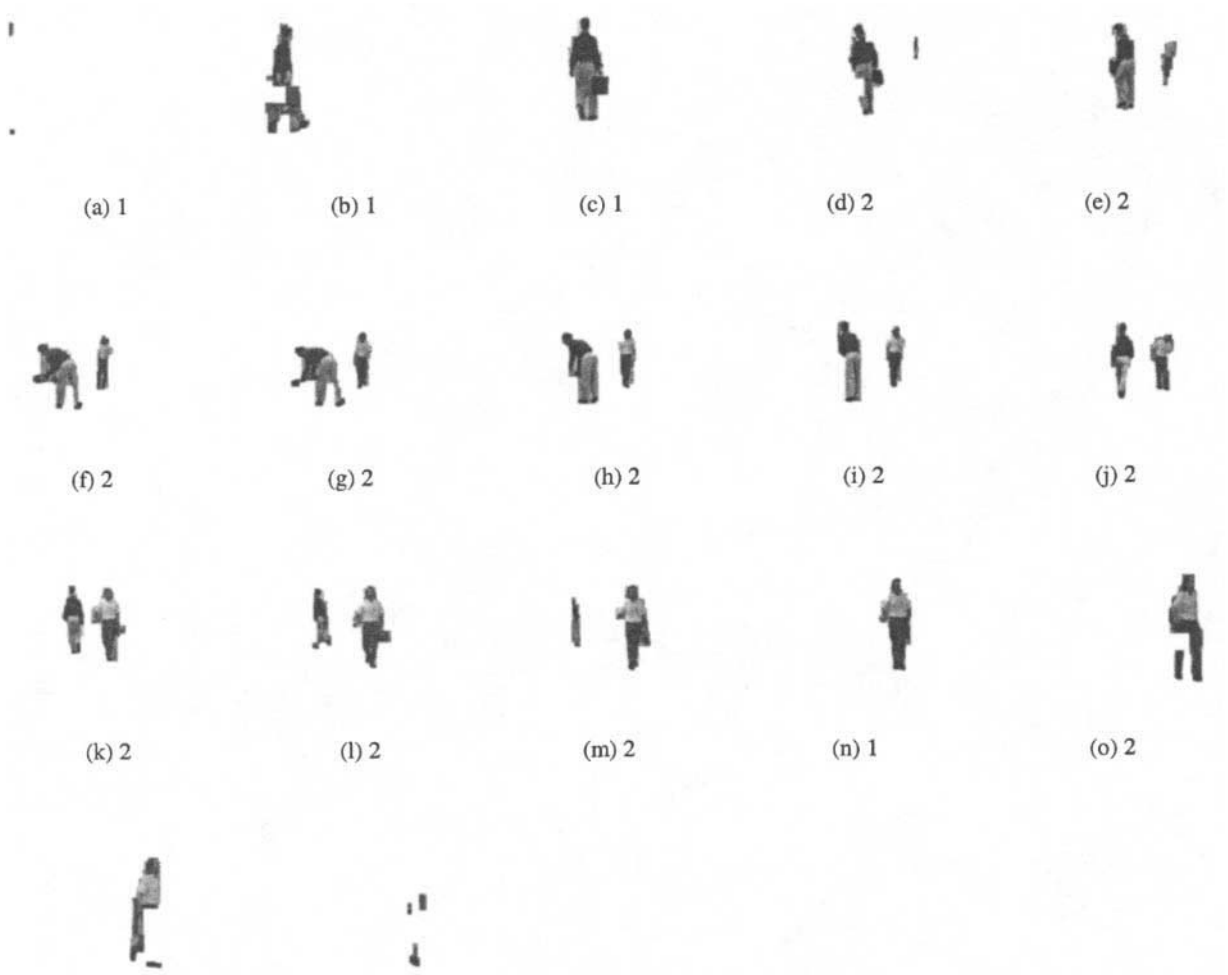


Fig.12: Extracted key frames generated by the integrated system. Frame no. (a)15, (b)26, (c)51, (d)76, (e)87, (f)106, (g)117, (h)132, (i)143, (j)164, (k)203, (l)235, (m)246, (n)257, (o)293, (p)304, (q)315, (r)326. Numbers denote the number of labeled regions in the frame. Note that some regions smaller than a predefined size are ignored in the KFE stage.