Journal of Southeast University (English Edition) Vol. 23, No. 4, pp. 500 – 506 Dec. 2007 ISSN 1003—7985

Motion connectivity-based initial video object extraction

Wang Yujian[1,2] Wu Zhenyang[1]

([1] School of Information Science and Engineering, Southeast University, Nanjing 210096, China)

([2] CBG/MMPD, Alcatel-Lucent Shanghai Bell Co. Ltd., Shanghai 201206, China)

Abstract: In order to obtain the initial video objects from the video sequences, an improved initial video object extraction algorithm based on motion connectivity is proposed. Moving objects in video sequences are highly connected and structured, which makes motion connectivity an advanced feature for segmentation. Accordingly, after sharp noise elimination, the cumulated difference image, which exhibits the coherent motion of the moving object, is adaptively thresholded. Then the maximal connected region is labeled, post-processed and output as the final segmenting mask. Hence the initial video object is effectively extracted. Comparative experimental results show that the proposed algorithm extracts the initial video object automatically, promptly and properly, thereby achieving satisfactory subjective and objective performance.

Key words: video object extraction; motion connectivity; adaptive threshold; cumulated difference image

With the increasing popularity of multimedia applications and content-based interactivity, new video describing and coding schemes are necessary. The standard MPEG-4, enabling content-based functionalities, introduces the concept of video object planes (VOPs). Each frame of the input sequence is composed of arbitrarily shaped image regions such that each VOP describes one semantically meaningful object or video content of interest. Real-time object-based video applications demand automatic extraction of semantic video objects (VOs).

An intrinsic problem of VO extraction is that objects of interest may not be homogeneous with respect to low-level features such as color, intensity, texture, edge, or optical flow[1-2]. Thus, conventional algorithms for video object segmentation may fail to obtain meaningful partitions. Due to the inherent difficulty of defining semantic video objects[2], automatic extraction of video objects, especially the initial video objects, is still quite a challenging problem.

Moving objects are often characterized by a coherent motion that is distinct from that of the background. The connectivity of the motion, semantically meaningful in a sense, is an advanced feature for segmenting video sequences into VOs. Detecting regions of change in images of the same scene is of widespread interest due to a large number of applications in diverse disciplines[3]. Conventional change detection-based video object segmentation algorithms[1,4-5] are conducted on the change mask comprised by the set of pixels that are "significantly different" between the last image of the sequence and the immediately previous image. These methods make no use of the motion connectivity so that rather complicated post-processing is necessary to obtain satisfactory segmentation results.

Contrarily, the cumulated difference image (CDI) of a certain number of successive frames exhibits the motion connectivity. Accordingly, an improved motion connectivity-based algorithm to automatically extract the initial video objects is proposed in this paper. Experimental results demonstrate that this algorithm is very effective and efficient. And the extracted initial video objects are fairly good for some real-time object-based video applications, which helps to alleviate the difficulty of semantic video object extraction.

1 Segmentation Algorithm

1.1 Motion connectivity as cue for segmentation

Foreground moving objects are distinguished from the background by their different coherent motions. Motion information can complement other features that are commonly adopted for segmentation, such as color, intensity, or edges. This makes motion a very useful feature incorporated into video object segmentation algorithms.

Moving objects are highly connected and structured[1]. That is, the coherent motion of an object is the gathering of all the motions of different components, and hence is assigned to a structured moving object.

Received 2007-02-28.

Foundation item: The National Natural Science Foundation of China (No. 60672094).

Biographies: Wang Yujian (1980—), male, doctor, yujian.wang@alcatel-sbell.com.cn; Wu Zhenyang (1949—), male, professor, zhenyang@seu.edu.cn.

Thereby, the motion connectivity of an object is an advanced feature indicating semantic meanings and is useful for segmentation.

In a video sequence with a still or a global-motion-compensated background, the difference image of successive two frames demonstrates the motion of the foreground object. But due to the motion locality and instantaneity, plus the influences of noises, the change detection mask (CDM) of immediately successive two frames only covers topical and localized moving parts, as illustrated in Fig. 1(b). Change detectors also mark occlusion areas as changed, while the object itself is unchanged unless it contains sufficient texture. This makes exact boundary localization very difficult, so that an additional mechanism is necessary to fill the holes inside the object. On the contrary, the cumulated difference image of a number of successive frames accumulates all the motions during a certain period of time, and thereby sufficiently exhibits the coherent motion of the object. In the mean time, in comparison with the connected and structured moving object, the stochastic noises are isolated and accumulate somewhat more slowly. Thus, in the cumulated difference image, the object motion is far more significant, as illustrated in Fig. 1(c).

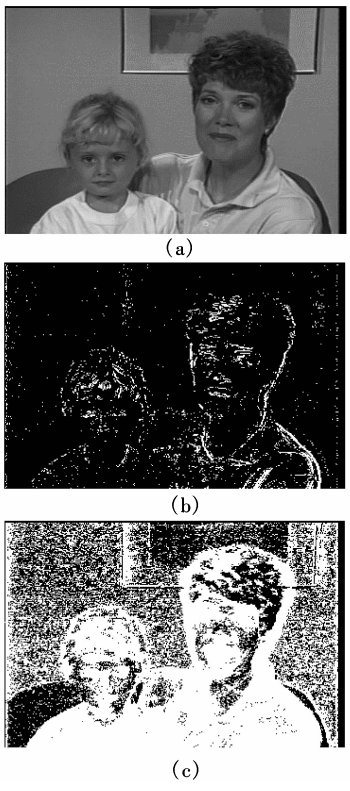

Fig. 1 Illustration of motion connectivity. (a) Initial frame of M & D sequence; (b) Difference image of the first two frames; (c) Cumulated difference image of 10 frames

1.2 Segmentation based on motion connectivity

Traditional change detection-based VO segmentation methods include the following procedures: threshold the difference image of immediately successive two frames, post-process the output and obtain a CDM, logically add an appropriate number of successive CDMs to obtain the final segmenting mask and extract the video object.

In comparison with the conventional methods, a novel initial video object extraction algorithm is presented here, which is subdivided into the following four steps: ① Calculate a certain number of immediate frame difference images (IFDIs) and accumulate them to obtain the CDI; ② Nonlinearly transform the CDI with an adaptive threshold; ③ Label and post-process the maximal connected region (MCR) and output it as the final segmenting mask; ④ Extract the initial video object. The flowchart of the proposed algorithm is shown in Fig. 2.

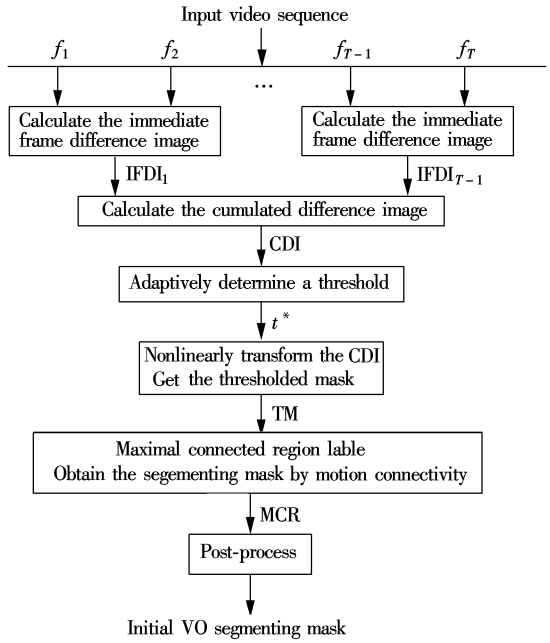

Fig. 2 Flowchart of the proposed algorithm

1.2.1 Calculate the cumulated difference image

Human eyes are more sensitive to luminance than to colors, so intensity is adopted here. We denote by $f_k(x, y)$ the intensity or luminance of pixels (x, y) in frame k, and $f_{k+1}(x, y)$ in frame $k+1$. Then the immediate frame difference is

$$d_k(x, y) = |f_k(x, y) - f_{k+1}(x, y)| \quad (1)$$

And the cumulated difference image is

$$\mathrm{cd}(x, y) = \sum_{k=1}^{T} d_k(x, y) \quad (2)$$

where T is the number of frames to be accumulated.

1.2.2 Adaptively threshold the cumulated difference image

Selecting an appropriate value to distinguish the foreground from the background under certain criteria is critical in the threshold algorithms. The classical threshold algorithm analyzes the histogram of features such as luminance, hue and saturation. The threshold value is determined by manual interactions, which restricts the algorithm from automatic functioning. Many adaptive threshold algorithms have been proposed, such as a threshold based on a genetic algorithm[6]. Unfortunately, due to the inevitable sharp noise in the CDI, these global threshold methods cannot afford satisfactory performance here.

In video sequences of natural scenes, slow changes in textures and stochastic changes in brightness caused by noises are regarded as Gaussian signals[7]. That is, the noise in the k-th difference image d_k follows a Gaussian distribution, assuming $N(0, \sigma_k^2)$. As the noises in the difference images are independent, the summed noise in the CDI also follows a Gaussian distribution $N(0, \sigma^2)$, where $\sigma^2 = \sum_{k=1}^{T} \sigma_k^2$. As the distribution of foreground object's motion information is rather different from that of Gaussian signals, the foreground moving object is very significant in the CDI. Encouragingly, because Gaussian distribution features in the fact that its higher statistics is zero, a higher order statistics (HOS) algorithm has the unique merit of extracting non-Gaussian signals from the Gaussian background[8]. Consequently, fourth order moment statistics in the CDI is applied here to remove the noise[4].

Intuitively, a moving foreground object often occurs on the central position of the frame image. Four blocks in the CDI, so-called still blocks, are specifically selected for background noise estimation, as shown in Fig. 3.

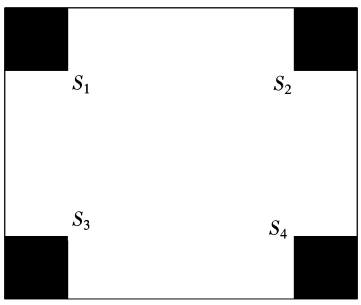

Fig. 3 Selection of still blocks

In order to obtain a threshold adherent to the real background noise statistics, the pixels in each still block are sorted according to their intensity. And the pixels with the highest and lowest intensities are eliminated to remove the blight of sharp noises. The variance of the remaining pixels in each still block is calculated. The median of the four values is regarded as the estimated variance of background noise, and is adopted for the following nonlinear threshold transformation. The estimation procedure is described as follows:

$$m_i = \frac{1}{N_i} \sum_{(x,y) \in S_i} \mathrm{cd}(x, y) \qquad i = 1, 2, 3, 4 \tag{3}$$

$$\sigma_i^2 = \frac{1}{N_i} \sum_{(x,y) \in S_i} [\mathrm{cd}(x, y) - m_i]^2 \tag{4}$$

$$\sigma^2 = \mathrm{median}\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\} \tag{5}$$

where S_i corresponds to the remaining pixels in the i-th still block, $\mathrm{cd}(x, y)$ denotes the intensity of pixel (x, y) in the CDI, m_i is the average intensity, and N_i is the number of the remaining pixels. In applications, the size of still blocks is typically set to 16×16.

Thus, the threshold $t^* = c(\sigma^2)^2$ is obtained, where c is an experiential factor corresponding to a certain type of video sequence.

Considering any pixel p in the CDI, the fourth order moment in its neighboring patch is calculated,

$$m_{\eta_p}^{(4)} = \frac{1}{M} \sum_{(x,y) \in \eta_p} [\mathrm{cd}(x, y) - \overline{m}_{\eta_p}]^4 \tag{6}$$

where

$$\overline{m}_{\eta_p} = \frac{1}{M} \sum_{(x,y) \in \eta_p} \mathrm{cd}(x, y) \tag{7}$$

η_p corresponds to the patch of p; $\overline{m}_{\eta_p}$ is the average intensity of this patch; M is the number of the pixels inside the patch. In applications, the patch size is usually set to 5×5 for CIF sequences and 3×3 for QCIF sequences.

Compare the moment $m_{\eta_p}^{(4)}$ with t^*, and the thresholded mask (TM) is obtained as follows:

$$\mathrm{TM}(p) = \begin{cases} 1 & \text{if } m_{\eta_p}^{(4)} \geqslant t^* \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

1.2.3 Maximal connected region labeling and post-processing

After the thresholding operation, there still exists noise greater than the threshold. Accordingly, there is discontinuity in the thresholded mask. One of the conventional methods dealing with this problem is morphological opening-closing filtering[1,4]. These methods can further reduce the noise and fill the "holes" and "cracks" inside the object. However, these methods do not work well for dramatically changing noise regions and big shadow regions.

Because of the motion connectivity of the fore-

ground object and the isolation of the noise, in the thresholded mask the area of the noise regions is normally much smaller than that of the semantic object. So in a video sequence with one object or multiple connected objects, the maximal connected region is where the objects are. Thus, maximal connected region labeling is adopted in our algorithm to extract the moving object. For cases of two or more separated objects, the number of the maximal connected regions to be labeled is accordingly increased. This method guarantees the integrality of the extracted object. And it is capable of removing the shadow regions that are not close bounding to the object, as well as the big noise regions that conventional methods cannot work well with.

In order to guarantee that the final mask has good continuity without dissociative small "holes", post-processing including morphological filtering and optional scan-filling[8] operation according to the spatial uniformity is applied. Morphological opening-closing by reconstruction filters with appropriate structural elements can remove the "holes" and "cracks" caused by noise and interior texture consistency. For cases of single objects with slight movements inside still backgrounds, optional scan-filling operation further helps to achieve a satisfactory segmentation mask.

After the post-processing, the final segmenting mask is obtained. And the initial foreground object is effectively extracted.

2 Experimental Results and Analysis

An experiment is implemented on the Claire sequence (CIF), which is one of the standard testing sequences for MPEG-4. The experimental results are shown in Fig. 4.

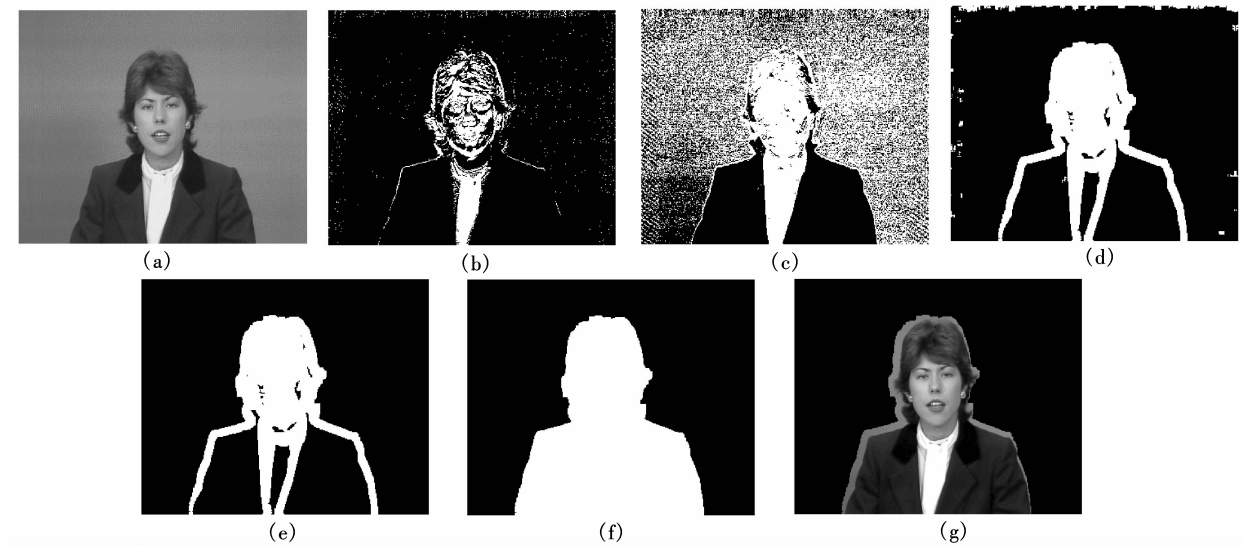

Fig. 4 Experimental results of the proposed algorithm on the Claire sequence. (a) Original initial frame; (b) IFDI of first two frames; (c) CDI of first five frames; (d) Thresholded mask; (e) Maximal connected region; (f) Final segmenting mask; (g) Extracted initial VO

Once the original initial frame (see Fig. 4(a)) is read, followed by sequential frames, the immediate frame difference images are calculated. And normally 5 IFDIs (the first one is shown in Fig. 4(b)) are added up to generate the cumulated difference image, as shown in Fig. 4(c). An adaptive threshold algorithm based on background noise estimation works to signify the foreground object, where the experiential factor c is typically set to 1 for sequences such as Claire. The output is the thresholded mask (see Fig. 4(d)). Then the maximal connected region (see Fig. 4(e)), labeled according to the motion connectivity, is post-processed with morphological filtering and scan-filling operations. The final initial VO segmenting mask and the extracted initial foreground object are respectively shown in Fig. 4(f) and Fig. 4(g).

In comparison with the background noise estimating method without sharp noise elimination in Ref. [4], in our experiment 10% of the pixels in the still blocks with the highest intensity and 10% of the pixels with the lowest intensity are excluded from the following noise estimation procedure. Because these pixels are typically influenced by the sharp noises and accordingly reduce the accuracy of the estimation from the actual statistics of noise. Take still block S_1 in Claire's CDI as an example. The data in this block are shown in Fig. 5.

Removing the 20% pixels of extrema, the estimated $\sigma_1^2 = 2.502\,4$. Without pixel elimination, the estimated $\sigma_1^2 = 5.316\,4$. With higher estimated noise vari-

4	4	14	16	9	6	15	9	15	17	14	15	6	14	19	10
7	7	13	12	10	10	20	8	14	14	12	13	17	19	14	12
9	8	9	14	5	9	18	13	19	12	14	14	18	15	18	18
14	9	11	9	9	12	19	10	15	10	17	13	16	12	15	16
14	16	7	17	11	9	14	8	12	13	17	11	13	12	9	15
8	13	11	15	8	10	14	14	18	11	9	7	7	15	14	18
14	10	17	14	9	11	19	9	12	13	16	13	14	16	13	13
19	9	16	18	8	8	13	14	17	10	20	12	11	20	16	11
10	10	14	17	13	10	11	7	17	9	17	8	21	16	13	21
12	16	12	18	13	7	14	9	17	10	15	11	13	14	15	17
10	10	11	18	13	13	15	13	10	17	15	14	17	15	15	13
12	15	10	17	15	15	11	14	9	16	16	13	18	11	11	14
10	15	13	12	12	14	13	10	12	13	12	13	14	14	9	9
13	15	7	12	7	12	18	13	13	13	13	15	13	12	17	14
13	16	5	11	8	14	16	18	18	14	10	11	11	13	15	15
15	15	11	13	13	11	16	13	14	17	10	14	11	14	8	15

Fig. 5 Still block S_1 of Claire's CDI

ance, the threshold to separate the foreground and the background increases. And more pixels of slow motion are classified to the background, as illustrated in Fig. 6. The post-processed segmenting mask is not unabridged and not good enough, and so is the extracted video object. In other words, in order to obtain a satisfactory segmentation result, a CDI of more incoming frames is necessary. This definitely leads to more time delay for the algorithm to work, which is not desired in some applications that are rigorous in time consumption.

In Ref. [9] the threshold value is the mean value of the intensity of the pixels after sharp noise elimination. This method takes no advantage of the spatial relationship of the neighboring pixels. The classification of the pixels is independent of one another, which results in the fact that a CDI of more incoming frames is *a priori* before obtaining an unabridged segmenting mask. Typically, the CDI of 20 frames is adopted in Ref. [9]. Similarly, as mentioned above, it is not acceptable for some real-time applications demanding low time delay. Furthermore, the CDI of more frames makes the extracted initial object have a much thicker boundary, as illustrated in Fig. 7. This makes the boundary refining algorithms take more time to converge the segmented boundary to the real one.

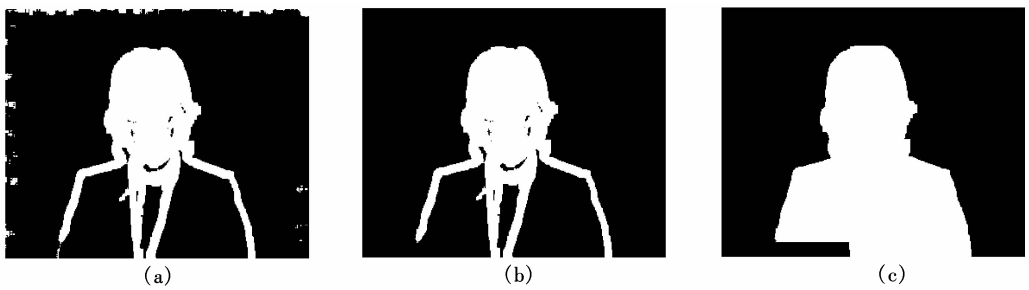

Fig. 6 Segmenting results without sharp noise elimination. (a) Thresholded mask; (b) Maximal connected region; (c) Final initial VO segmenting mask after the post-processing with scan-filling operation

Fig. 7 Comparative experimental results. (a) Extracted initial VO by the method in Ref. [9]; (b) Extracted initial VO by the proposed algorithm

Manual extraction of masks is taken as a reference (so-called ground-truth) for object performance evaluation. The space accuracy, one of the evaluation criteria for segmenting algorithms, is formulated as

$$S = d(A^{\text{est}}, A^{\text{ref}}) = 1 - \frac{\sum_{(x,y)} A^{\text{est}}(x,y) \oplus A^{\text{ref}}(x,y)}{\sum_{(x,y)} A^{\text{ref}}(x,y)} \tag{9}$$

where A^{est} and A^{ref} are the estimated and reference object masks, and $\oplus$ denotes the logical XOR operation.

For the Claire sequence, the space accuracy of the proposed algorithm is 88.9%, while that of the method in Ref. [9] is 85.6%.

More comparative experimental results are demonstrated in Fig. 8.

Figs. 8(a) to (d) are the experimental results on the M & D sequence, which is selected on purpose as a representative for cases of multiple connected foreground objects. Fig. 8(c) is the extracted initial VO by the method in Ref. [9]. Because the CDI of 5 frames (see Fig. 8(b)) is not sufficient for this method to extract a satisfactory unabridged initial VO, the CDI of 20 frames is applied. In comparison with this, the proposed algorithm achieves much better performance and the extracted object (see Fig. 8(d)) has a rather thinner boundary. Figs. 8(e) to (h) are the experimental results

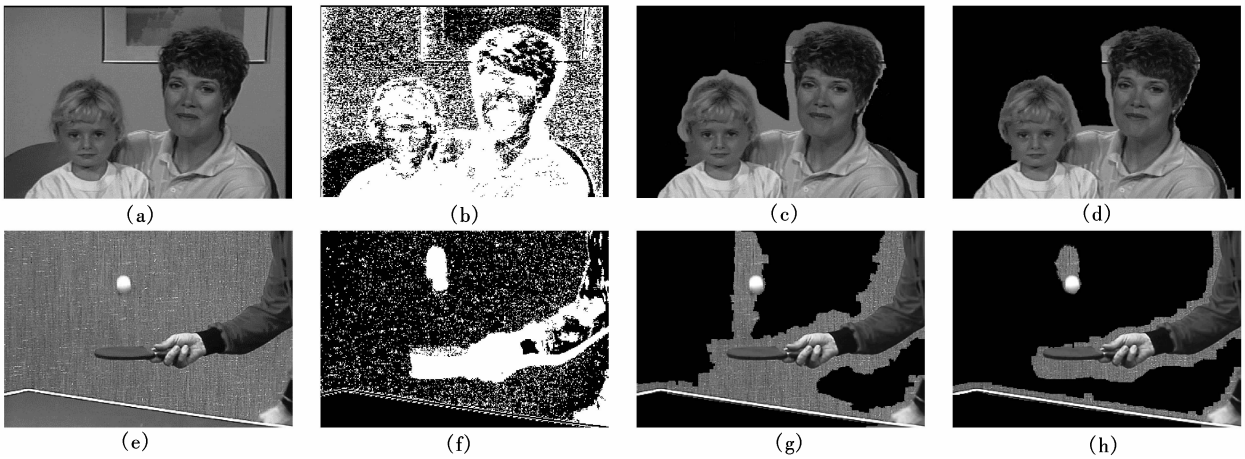

Fig. 8 Comparative experimental results. (a) Initial frame of the M & D sequence; (b) CDI of the first 5 frames; (c) Extracted initial VO by the method in Ref. [9] with the CDI of 20 frames; (d) Extracted initial VO by the proposed algorithm with the CDI of 5 frames; (e) Initial frame of the Tennis sequence; (f) CDI of the first 6 frames; (g) Extracted initial VO by the method in Ref. [9] with the CDI of 20 frames; (h) Extracted initial VO by the proposed algorithm with the CDI of 6 frames

on the Tennis sequence, which is selected as a representative of cases of separated foreground objects. Similarly, the proposed algorithm obtains rather better segmentation. Generally speaking, because of the improvements to the threshold calculation, the proposed algorithm works somewhat better than its predecessor.

From the above comparative experimental results, it is clear that the proposed algorithm can extract the initial video object efficiently and effectively, and it has fairly satisfactory subjective and objective performance.

3 Conclusion

An improved initial video object extraction algorithm is proposed in this paper. This algorithm is based on the inherent motion connectivity of moving objects. By adaptively thresholding the cumulated difference image with sharp noise elimination, the connected moving regions where the objects lie in are signified. After the maximal connected region labeling and post-processing, the final initial VO segmenting mask is obtained. And the unabridged foreground object is extracted automatically, promptly and properly.

This algorithm has fairly good performance for sequences of still backgrounds and foreground objects with slight movements, which appeals to applications such as video communication and video conferencing. For the cases of moving backgrounds, global inter-frame motion estimation and compensation is *a priori*.

Noticeably, as the cumulated difference image accumulates object motion for several timeslices, plus the effects of morphological filtering operations in post-processing, the segmented foreground object has a fairly thick boundary. This is acceptable for applications such as video surveillance and video retrieval. But for applications demanding high accuracy such as video compression and video synthesis, it is necessary to explore some methods to further refine the boundaries. Many applicable boundary refining algorithms have been proposed. An optional solution is proposed in Ref. [10]. Thus, the initial VO with a more accurate boundary is successively obtained, and then is available for a sequential video object tracking process.

References

[1] Meier T, Ngan K N. Automatic segmentation of moving objects for video object plane generation [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 1998, **8** (5): 525 − 538.

[2] Liang T M, Tang D X, Chen J. Research and implementation of a video object extraction system [J]. *Computer Engineering*, 2003, **29**(1): 182 − 193.

[3] Radke R J, Andra S, Al-Kofahi O, et al. Image change detection algorithms: a systematic survey [J]. *IEEE Transactions on Image Processing*, 2005, **14**(3): 294 − 307.

[4] Neri A, Colonnese S, Russo G, et al. Automatic moving object and background separation [J]. *Signal Processing*, 1998, **66**(2): 219 − 232.

[5] Mech R, Wollborn M. A noise robust method for segmentation of moving objects in video sequences [C]//*Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. Munich, Germany, 1997: 2657 − 2660.

[6] Zheng H, Pan L. The automatic selection of image threshold on the basis of genetic algorithms [J]. *Journal of Image and Graphics*, 1999, **4**(4): 327 − 330.

[7] Aach T, Kaup A, Mester R. Statistical model-based change detection in moving video [J]. *Signal Processing*, 1993, **31**

(2): 165 - 180.

[8] Kim C, Hwang J. Fast and automatic video object segmentation and tracking for content-based applications [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2002, **12**(2): 122 - 129.

[9] Liu M Y, Dai Q H, Liu X D, et al. Automatic extraction of initial moving object based on advanced feature and video analysis [C]//*Proceedings of Visual Communication and Image Processing*. Beijing, China, 2005: 160 - 168.

[10] Wang Y J, Gao J P, Wu Z Y. A novel video object spatial segmenting strategy based on morphological filtering [C]//*Proceedings of International Conference on Computational Intelligence and Security*. Guangzhou, China, 2006: 1677 - 1682.

基于运动连通性的初始视频对象提取

王煜坚[1,2] 吴镇扬[1]

([1] 东南大学信息科学与工程学院,南京 210096)

([2] 上海贝尔阿尔卡特朗讯股份有限公司网络融合集团多媒体事业部,上海 201206)

摘要:为了从视频序列中获取初始视频对象,提出了一种改进的基于运动连通性的初始视频对象提取算法.视频中的运动对象高度连通结构化,这就使得运动连通性是适用于视频对象分割的高级特征.据此首先对反映对象的一致性运动的累计帧差图进行尖锐噪声滤除,然后应用自适应阈值算法提取对象运动区域,接着根据运动连通性标记出最大连通区域,通过后处理得到视频对象的分割模版从而有效提取出初始视频对象.对比实验结果表明,该算法能自动、快速、准确地提取出初始视频对象,获得了理想的主客观分割效果.

关键词:视频对象提取;运动连通性;自适应阈值;累积帧差图

中图分类号:TN911.73