INTEGRATION OF MOTION AND IMAGE FEATURES FOR AUTOMATIC VIDEO OBJECT SEGMENTATION

Wei Wei and King N. Ngan, Fellow IEEE

Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

ABSTRACT

The paper presents an automatic video segmentation algorithm concerning multiple features. The *k*-median algorithm is employed to partition an image into a set of homogenous regions. The location of moving objects is determined by change detection and tracked by using region descriptors. Experiments have been carried out on several video sequences and results have shown the efficiency of this approach.

1. INTRODUCTION

Segmentation of moving objects in a video sequence has many potential applications in wide range of areas, including video surveillance, object detection and tracking, and object-based video compression [1].

For the reason that the moving object might rotate or change its shape, as it is moving through the video sequence, tracking is a fundamental step to follow semantic video objects in the scene and to update their 2D shape from frame to frame. At least two distinct approaches can be identified in tracking: (1) region-based tracking using motion clustering [2], and (2) contour tracking using curve evolution [3, 4]. While region-based tracking exploits full region information to reliably estimate global object motion even in case of partial occlusion, it is not very precise in extracting object edges. On the other hand, contour tracking used intensity gradient to guide curve evolution towards image edges, which means that object boundaries can be precisely estimated, however good initialization is required and thus this method can treat only slow object motion and simple background.

In this paper, a technique of automatic video segmentation based on the *k*-medians clustering method [5] and region tracking is described. In spatial segmentation, there are two stages. The first stage obtains initial regions using a single-feature clustering method. The second stage refines the regions using a combination of features according to a set of appropriate rules. Regions belonging to moving objects are tracked by using region descriptors from the current frame to the next frame. This allows the proposed method to track fast moving objects as well as to detect the disappearance of existing objects from the scene.

The remainder of this paper is organized as follows. Section 2 is concerned with the spatial segmentation. Section 3 describes the VOP extraction and tracking. Results and their interpretations are discussed in Section 4, while the Section 5 gives a summary of the work.

2. SPATIAL SEGMENTATION

The *k*-median clustering algorithm is used to partition an image into a set of disjoint homogeneous regions whose union is the entire image [5].

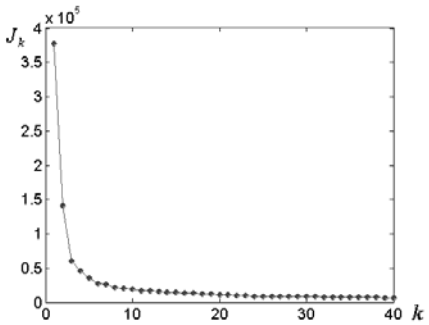
2.1 *k*-Medians Clustering Algorithm

The purpose of clustering is to partition a set of n feature vectors $C = \mathbf{u}_1,...,\mathbf{u}_n$ into k disjoint subsets, $C_1,...,C_k$. Each subset represents a cluster, with the feature vectors in the same cluster being more similar to each other than to the feature vectors in other clusters. J_k measures the total squared error incurred in representing the n samples $\mathbf{u}_1,...,\mathbf{u}_n$ by the k cluster centers, $\mathbf{m}_1,...,\mathbf{m}_k$. It is clear that J_k, must decrease monotonically as k increases, because the squared error can be decreased each time k is increased merely by transferring a single sample to new singleton cluster. If the n samples are really grouped into $\hat{k}$ compact, well-separated clusters, J_k should decrease rapidly until $k = \hat{k}$, and then decrease much more slowly thereafter until it reaches zero at $k = n$. Based on this property, the cluster number is estimated by analyzing the behavior of J_k for $k = 1, \ldots, k = k_{max}$ for the first frame [6].

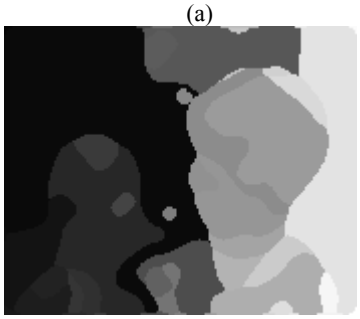
First, only luminance information is used to obtain an estimate of the cluster numbers in the first frame. Figure 1(a) shows the J_k versus k curve for the first frame of the *Mother & Daughter* sequence. Figure 1(b) depicts the corresponding segmentation field.

There are two problems to be solved before clustering the image. One is normalization of the different features.

0-7803-8554-3/04/$20.00 ©2004 IEEE.



(a)



(b)

Figure 1. *Mother & Daughter* sequence: (a) J_k versus k curve. (b) Segmentation fields

The other is that different weight should be assigned to different feature. The features that we propose to use in our segmentation scheme belong to five groups: luminance, chrominance, texture, position and motion. A common solution is to normalize with respect to the standard deviation over the entire image, which is adopted in [7]. A fixed weight is assigned to the position information. The weights assigned to luminance, chrominance and texture features are affected by the variances of the features, which are the medians of the corresponding features in the region $R_m^{(1)}$, $m = 1,...,M$, obtained by using the k-medians clustering algorithm based on the luminance information only in the first frame [5].

The exact k value will be obtained based on the multiple features, which is the same way as using one feature. The weight of motion information is computed in the similar fashion as the luminance, chrominance and texture information. The difference between them is that the weight of motion is computed adaptively by using the region information obtained in the previous frame, but the weights of other features are decided in the first frame, unless a scene change occurs. Figure 2 depicts the corresponding segmentation fields of *Mother & Daughter* sequence based on multiple features.

Figure 2. *Mother & Daughter* sequence: segmentation field based on multiple features.

2.2 Region Merging

The clustering algorithm is applied to partition the image into small regions that are homogeneous in terms of the multiple features. It results in over-segmentation, as shown in Figure 2. To solve the over-segmentation problem, the region merging method groups the sub-regions into larger regions based on a similarity criterion.

For every two neighboring regions R_i and R_j, we compute σ_i and σ_j, the standard deviations of R_i and R_j, respectively. We further calculate the standard deviation of the union of R_i and R_j as defined below:

$$\mathbf{m}_{i,j} = \frac{1}{(n_i + n_j)} \sum_{u \in (R_i \cup R_j)} \mathbf{u} \qquad (1)$$

$$\boldsymbol{\sigma}^2_{i,j} = \frac{1}{(n_i + n_j)} \sum_{u \in (R_i \cup R_j)} (\mathbf{u} - \mathbf{m}_{i,j})^2 \qquad (2)$$

where n_i and n_j are the number of pixels in region R_i and R_j. The decision rule of region merging is then given by

$$\sigma_{(i,j)l} \le \min(\sigma_{il}, \sigma_{jl}) \qquad (3)$$

for all $l = 1,...,k$. That is, the variance $\sigma_{(i,j)l}$ of the combined regions is smaller than the minimum of the two regions R_i and R_j, these two regions will be merged. Figure 3 shows the region merging results of Figure 2.

3. VOP EXTRACTION

The information derived from the change detection and spatial segmentation processes will be utilized to extract

Figure 3. Frame 1 of the *Mother & Daughter* sequence: Region merging result.

the VOPs. The VOPs are extracted from the video sequence by first detecting the moving regions, and then tracking those regions throughout the sequence.

3.1 Detection of Moving Regions

If majority of a region $R_c^{(t)}$ in current frame *t* is indicated as changed in the CDM, the region is declared as a moving region and marked in the VOP. A decision rule for the detection of moving region is defined as

$$T_c = \frac{N_{R_c^{(t)} \cap CDM}}{N_{R_c^{(t)}}} \tag{4}$$

where $N_{R_c^{(t)}}$ is the number of pixels in region $R_c^{(t)}$. If the value of T_c is greater than or equal to a given threshold, the region R_c^t is considered belonging to a moving object; otherwise it is background.

3.2 Tracking of Moving Regions

Through a video sequence, the topologies of the homogeneous region vary over time. So the corresponding regions at different time instances have to be linked together. This temporal linkage is achieved through tracking, which can be done by projecting the region descriptors from the current frame to the next frame. Projecting a region descriptor instead of the entire region is a simple and effective strategy. The simplicity comes from the fact that instead of projecting the entire region into the next frame, only the region descriptor needs to be processed. In addition, region descriptor projection is effective, since it can cope with deformation and complex motion. A region descriptor $\boldsymbol{\varphi}_r^{(t-1)}$ is computed for every moving region $R_r^{(t-1)}$ in frame *t*-1, which is given by

$$\boldsymbol{\varphi}_r^{(t-1)} = \underset{x \in R_r^{(t-1)}}{median}(\mathbf{u}) \tag{5}$$

Note that the region descriptor is a vector where each element of the vector is the median of the corresponding feature in the region $R_r^{(t-1)}$.

In order to track regions that may have temporarily ceased motion, the region descriptors $\boldsymbol{\varphi}_r^{(t-1)}$ of the regions belonging to moving objects in frame *t*-1 are compared with the region descriptors $\boldsymbol{\varphi}_c^t$ of regions in frame *t*. The region, which has the minimum distance with the region descriptor $\boldsymbol{\varphi}_r^{(t-1)}$, should be marked in the VOP as moving.

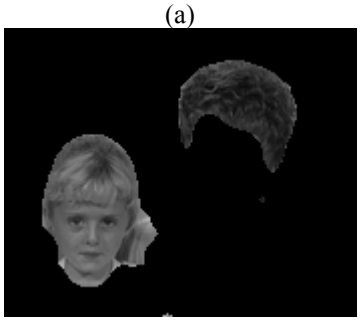
$$\min_C d(\boldsymbol{\varphi}_r^{(t-1)}, \boldsymbol{\varphi}_c^t) = \min_C \sum_{l=1}^{f} w_l \left| u_l^r - u_l^c \right| \tag{6}$$

where u_l^r (respectively, u_l^c) is the median of the feature u_l in region $R_r^{(t-1)}$ (respectively, R_c^t) and w_l is the weight assigned to the corresponding feature.

3.3 Hole Removal

If we only track the regions belonging to moving objects in the previous frame *t*-1, some small regions obtained by

(a)



(b)

Figure 4. Frame 87 of the *Mother and Daughter* sequence. (a) CDM (Change Detection Mask). (b) Moving regions with $T_c = 0.6$.

Figure 5. VOPs Extraction by using region descriptor

Figure 6. VOPs Extraction with hole removal

clustering method in the current frame *t*, which do not get any projected region descriptor from the previous frame, will be lost, as shown in Figure 5. These regions will be matched with the mask of the previous frame to decide if it should be marked in the VOP as moving. The result of VOP extraction with hole removal is depicted in Figure 6.

4. SIMULATION RESULTS

This algorithm described here has been applied to several different video sequences in QCIF format. Due to the limitation of space, the results of only two sequences will be shown. These sequences are *Mother* & *Daughter* and *Children*.

In both the *Children* and *Mother & Daughter* sequences, there are multiple objects required to be extracted. In the *Mother & Daughter* sequence, the head of the mother has a relatively large motion, while her body exhibits little motion. The motion of the daughter is little throughout the sequence. Our algorithm is still capable of determining the locations of the moving objects reasonably well, as demonstrated in Figures 6.

In order to demonstrate the accuracy of the proposed algorithm, the result of the *Children* sequence is compared with the segmentation results obtained manually. The object criterion named Validity Percentage can be described as in the following equation:

$$\text{validity percentage} = 1 - \frac{S_c^t \oplus S_m^t}{S_m^t} \qquad (7)$$

where S_m^t is the mask of moving object in the frame *t* obtained manually, S_c^t is the mask of segmentation results and $\oplus$ is logic XOR operation.

Figure 8 shows the validity percentage curve for *Children* sequence. The x-axes of the figure describe the number of frame and the y-axes are the corresponding validity percentage of results.

5. CONCLUSION

This paper proposed automatic VOPs generation method that continuously separates the moving objects in image frames through time evolution. The features employed are

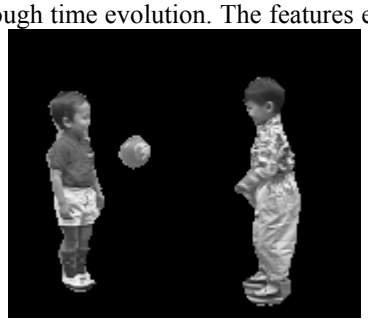


Figure 7. *Children* sequence: segmentation result

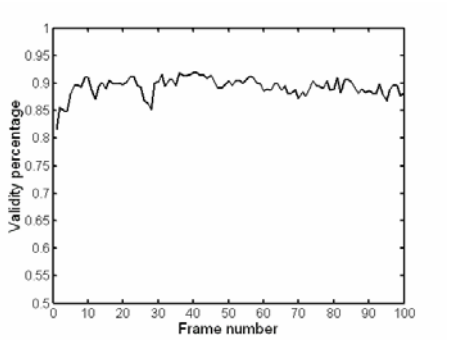


Figure 8. Validity percentage for *Children* sequence

luminance, chrominance, texture, position and motion, and their corresponding weights are determined by the characteristic of the sequence, which are determined by the variances in the first frame, except for the position and motion information. The weight of position is constant and the weight of motion is given adaptively by using the region information obtained in the previous frame. The location of moving objects depends on the change detection. The correspondence between instances of moving objects over frames is established by region tracking. Region descriptor used in this algorithm is an effective way, which is simple and can cope with deformation and complex motion.

Experimental results demonstrate that the proposed method is able to successfully extract moving objects from the sequence.

6. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 N2202 "Information Technology – Coding Of Audio-Visual Objects: Visual", Tokyo, March 1998.

[2] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal segmentation based on region merging, " IEEE Transactions on PAMI, vol. 20, No. 9, pp. 897-915, September 1998.

[3] N. Paragios, R. Deriche, "Unifying Boundary and Region-based information for Geodesic Active Tracking", Proc. CVPR, vol. 2, pp. 300-305, Fort Collins, Colorado, 1999

[4] T. Meier and K. N. Ngan, "Video segmentation for content-based coding," IEEE Transaction on Circuits and Systems for Video Technology, vol. 9, pp. 1190-1203, December 1999.

[5] W. Wei and K. N. Ngan, "Multiple feature clustering algorithm for automatic video object segmentation," ICASSP'04, Canada, May 2004.

[6] N. Nariman and K. N. Ngan, "Automatic Multi-cue VOP Extraction for MPEG-4," Picture Coding Symposium 2003, Saint Malo - France, April 2003.

[7] R. Castagno, T. Ebrahimi, and M. Kunt, "Video segmentation based on multiple features for interactive multimedia applications," IEEE Transaction on Circuits Systems for Video Technology, vol. 8, no. 5, pp. 562-571, September 1998.