An Automatic Segmentation Algorithm for Moving Objects in Video Sequences under Multi-Constraints

Wu Si Zhang Yong-dong Lin Shou-xun
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.
Email: {sw, zhyd, sxlin}@ict.ac.cn

Abstract

A new algorithm under multi-constraints for automatic segmentation of video moving objects is proposed. First, the temporal segmentation separates the initial areas including moving objects accurately from the background by continuous frame difference. Then, the spatial segmentation segments the initial areas into spatially consistent regions by Watershed algorithm based on a color gradient. Finally, regions are classified as foreground/ background by maximizing the a posterior probability (MAP) of the MRF with spatial, temporal and adjacent constraints. Experimental results demonstrate that the algorithm is not sensitive to objects' irregular movement and illumination, and it can extract moving video objects accurately.

1. Introduction

Automatic moving video objects segmentation, the fundament of the content-based video application, has been intensively studied in the past. The existing automatic moving video objects segmentation techniques can be classified into three approaches, namely, 1) spatial segmentation, 2) temporal segmentation, 3) hybrid segmentation.

1) Spatial segmentation [1], [2] can obtain precise objects' boundary using the spatial information (color, texture, gradient, etc.). However, the segmentation results are not always intact objects due to only the spatial information used.

2) Temporal segmentation [3], [4] can detect changes between frames rapidly using temporal (motion) information (e.g. the frame difference). However, the segmentation results don't track the objects' boun-dary precisely due to only the temporal information used.

3) Hybrid segmentation [5], [6] should be the optimal approach currently to extract the video objects by taking spatio-temporal information into account. The major steps of these approaches can be summarized as follows. First, spatial segmentation technique, such as region growing, watershed algorithm etc, is applied to segment the whole image into spatially consistent regions. Then, motion information of each region is calculated by motion estimation. Finally, regions with similar motion are merged together to form the final objects. Hybrid segmentation can extract object accurately. However, the computation complexity is very high because the procession is performed on the whole image. Besides, the region classification is unreliable because the motion estimation is sensitive to the objects' irregular movement (e.g. rapid or non-rigid movement) and obvious illumination.

An automatic hybrid segmentation algorithm under multi-constraints is proposed. To reduce the computation complexity, spatial segmentation and region classification are only performed on the initial areas which are approximate objects' areas separated from the background by temporal segmentation using continuous frame difference (CFD). To overcome the drawback of region classification based on motion estimation, regions are classified as foreground/background by estimating the MAP of the MRF with spatial, temporal and adjacent constraints, according to the similarity between the region and current background obtained by CFD, the region and the segmentation results of previous frame, the region and its neighbors.

The proposed algorithm is divided into three major steps. Firstly, temporal segmentation separates initial areas from current background by CFD. Secondly, only initial objects' areas are segmented into spatially consistent regions by watershed algorithm [7] based on a color gradient. Finally, regions are classified by estimating the MAP of the MRF with spatial, temporal and adjacent constraints.

2. Temporal Segmentation

Temporal segmentation separates initial areas from current background, which premises the reduction of computation complexity in the

0-7803-8603-5/04/$20.00 ©2004 IEEE

following procession. It is implemented using CFD with two preliminary steps, namely, global motion estimation and compensation, change detection.

1) Global Motion Estimation

To estimate the global motion, the camera motion is modeled by the six-parameter affine motion model:

$$\begin{cases} x' = ax + by + e \\ y' = cx + dy + f \end{cases} \quad (1)$$

the global motion is estimated iteratively using Gauss-Newton (GN) algorithm. The initial parameters of global motion estimation are obtained by least square method. To reduce the computation time, the GN algorithm is applied within a three-level multi-resolution pyramid, which is generated using [1/4, 1/2, 1/4] filter. For global motion compensation the bilinear interpolation is used.

2) Change Detection

Let I_t and $I_{t'}$ denote two consecutive frames, $d_{t,t'}$ is the frame difference:

$$d_{t,t'}(p) = W \times I_t(p) - W \times I_{t'}(p') \quad (2)$$

The change detection mask $D_{t,t'}$ is defined as:

$$D_{t,t'}(p) = \begin{cases} 1 & if d_{t,t'}(p) > T \\ 0 & else \end{cases} \quad (3)$$

where W is a smooth filter (e.g. 3×3 Gaussian filter), T is the threshold depending on the camera noise, which can be selected within [5, 10] according to the specific occasion of video application.

Connected components analysis [9] is applied to eliminate those smaller noise regions in $D_{t,t'}$.

3) Continuous Frame Difference (CFD)

There are always some occluded or disoccluded background areas left in $D_{t,t'}$. To obtain more accurate initial areas, the CFD is calculated. Let $D_{t,t-1}$, $D_{t,t+1}$ denote the change detection masks of current frame I_t and previous frame I_{t-1}, next frame I_{t+1}, the CFD mask D_t is defined as:

$$D_t = D_{t,t-1} \cap D_{t,t+1} \quad (4)$$

By calculating the CFD, more accurate initial areas IF_t are separated from current background IB_t, and there are few background areas left in the IF_t.

3. Spatial Segmentation

Watershed algorithm based on the gradient image [7] is applied to segment the initial areas IF_t due to its robustness and affectivity. To improve the accuracy of segmentation, the IF_t's gradient in the *YCbCr* color space is calculated.

Let $G_Y^{'}$, $G_{Cb}^{'}$ and $G_{Cr}^{'}$ denote the normalized gradient images of the three color components (*Y*, *Cb* and *Cr*), which are calculated using Canny's gradient approximation [9], and then normalized into [0, 255]. The color gradient G_{col} is calculated as:

$$G_{col}(p) = \begin{cases} \max\left\{\omega_Y \cdot G_Y^{'}(p), \omega_{Cb} \cdot G_{Cb}^{'}(p), \omega_{Cr} \cdot G_{Cr}^{'}(p)\right\} & if D_t(p) = 1 \\ 0 & else \end{cases} \quad (5)$$

where $\omega_Y, \omega_{Cb}, \omega_{Cr}$ are the weight coefficients of the gradient magnitudes of *Y*, *Cb* and *Cr*. In experiments, we have used ω_Y =0.5, $\omega_{Cb} = \omega_{Cr}$ =0.25.

Watershed is implemented using fast immersion simulation [7] on the initial areas IF_t. To overcome the watershed's inherent drawback, namely, over segmentation, a spatio-temporal merging scheme [5] is adopted to merge small regions.

For the convenience of narration in the rest of the paper, let $R_t = \{R_t^1, \cdots, R_t^K\}$ denote the set of regions after spatial segmentation. N_i is the size of R_t^i, $Nor(R_t^i)$ is the set of neighbors of R_t^i. $IF_t = \bigcup_{i=1}^{K} R_t^i$ is the initial areas obtained by CFD, $IB_t = I_t - IF_t$ is the current background. $IO_{t-1} = \bigcup_{L(R_{t-1}^i)=F} R_{t-1}^i$ is the objects segmented from previous frame, where $L(R_{t-1}^i)$ is the classification of R_{t-1}^i ($L(R_{t-1}^i) \in \{F, B\}$, F denotes foreground, B denotes background).

4. Region Classification

MRF model is a very prominent stochastic model applied comprehensively in image processing and computer vision. For region classification, MRF is defined using spatial, temporal and adjacent constraints based on the similarity between the region and current background, the region and the segmented results of previous frame, the region and its neighbors. Then, regions are classified by estimating the MAP of the MRF.

According to the Hammersley-Clifford theorem [8] and Bayes' rule, the MAP of MRF can be estimated by minimizing the posterior energy [5], [8]. So we define the posterior energy $U_p(X \mid O)$ of MRF as:

$$U_p(X|O) = \sum_{i=1}^{K} \alpha \cdot V_i^S(X,O) + \beta \cdot V_i^T(X,O) + \sum_{(i,j)\in E} \gamma \cdot V_{ij}^R(X,O) \quad (6)$$

where $V_i^S(X,O)$, $V_i^T(X,O)$, $V_{ij}^R(X,O)$ are the energy functions corresponding to spatial, temporal and adjacent constraints, α, β, γ are the associated weight coefficients, $E = \{(i,j) \mid R_t^j \in Nor(R_t^i)\}$ is the set of all adjacent relationship.

1) spatial constraint energy $V_i^S(X,O)$

$$V_i^S(X,O) = \begin{cases} f(SD(R_t^i), T_s, SD_{max}, SD_{min}) & X_i = B \\ 1 - f(SD(R_t^i), T_s, SD_{max}, SD_{min}) & X_i = F \end{cases} \quad (7)$$

where $SD(R_t^i)$ is the spatial similarity of R_t^i and IB_t:

$$SD(R_t^i) = \min_v \frac{1}{N_b} \sum_{l=1}^{3} \omega_l \cdot \sum_{p \in R_t^i} | I_t^l(p) - I_t^l(p+v) |,$$

$$D_t(p+v) = 0, N_b > \tfrac{2}{3} N_i \quad (8)$$

$I_t^1(p), I_t^2(p), I_t^3(p)$ are the associated intensity functions of Y, Cb, Cr, $\omega_1, \omega_2, \omega_3$ are associated weight coefficients same as $\omega_Y, \omega_{Cb}, \omega_{Cr}$ in section 3. v is the matching vector within a $w \times w$ window. N_b is the number of background pixels matching with R_t^i. $f(d, T, d_h, d_l)$ is a scaling function which normalizes d to [0, 1]:

$$f(d,T,d_h,d_l) = \begin{cases} 0.5 \times (d - d_l)/(T - d_l) & if d < T \\ 0.5 + 0.5 \times (d - T)/(d_h - T) & else \end{cases} \quad (9)$$

$V_i^S(X,O)$ represents the likehood of the region R_t^i to be classified as foreground/background based on the magnitude of the spatial similarity $SD(R_t^i)$ and threshold T_s. In IF_t, only a little part of regions are non-object regions and very similar to nearby areas in IB_t, thus, the likehood of region to be classified as foreground/background can be estimated according to the spatial similarity. Doing this way, the algorithm can overcome the drawback of region classification based on motion estimation which is sensitive to irregular movement and illumination.

2) temporal constraint energy $V_i^T(X,O)$

$$V_i^T(X,O) = \begin{cases} f(TD(R_t^i), T_t, TD_{max}, TD_{min}) & X_i = F \\ 1 - f(TD(R_t^i), T_t, TD_{max}, TD_{min}) & X_i = B \end{cases} \quad (10)$$

where $TD(R_t^i)$ is the temporal similarity of R_t^i and IO_{t-1}:

$$TD(R_t^i) = \min_{R_{t-1}^j} \sum_{l=1}^{3} \omega_l \cdot | \underset{p \in R_t^i}{avg} I_t^l(p) - \underset{p \in R_{t-1}^j}{avg} I_{t-1}^l(p) |,$$

$$L(R_{t-1}^j) = F \quad (11)$$

$V_i^T(X,O)$ represents the likehood of the region R_t^i to be classified as foreground/background based on the magnitude of temporal similarity $TD(R_t^i)$ and threshold T_t. The more similar R_t^i and IO_{t-1} are, the more likely R_t^i is to be classified as foreground. By considering the temporal constraint, mistaken classifications of regions similar to the background can be prevented.

3) adjacent constraint energy $V_{ij}^R(X,O)$

$$V_{ij}^R(X,O) = \begin{cases} (RD(R_t^i, R_t^j) - RD_{min})/(RD_{max} - RD_{min}) & X_i = X_j \\ 1 - (RD(R_t^i, R_t^j) - RD_{min})/(RD_{max} - RD_{min}) & X_i \neq X_j \end{cases} \quad (12)$$

where $RD(R_t^i, R_t^j)$ is the adjacent similarity of R_t^i and its neighbor R_t^j:

$$RD(R_t^i, R_t^j) = \sum_{l=1}^{3} \omega_l \cdot | \underset{p \in R_t^i}{avg} I_t^l(p) - \underset{p \in R_t^j}{avg} I_t^l(p) |,$$

$$R_t^j \in Nor(R_t^i) \quad (13)$$

$V_{ij}^R(X,O)$ represents that the more similar R_t^i and its neighbor R_t^j are, the more likely R_t^i and R_t^j belong to the same class.

The constants α, β, γ determine the relative proportion of the three terms in the posterior energy. In experiments, we found that using $\alpha = 1.0$, $\beta = \gamma = 0.6$ can obtain satisfactory results. The thresholds T_s and T_t depend on the complexity of video's background. To choose accurate thresholds adaptively, we select the spatial similarity and temporal similarity as classification plane respectively, and choose the plane that maximize the distance between classes as the value of T_s and T_t according to the Fisher linear discrimination criterion. The minimization of posterior energy is performed using an iterative deterministic relaxation scheme known as HCF [8].

5. Experimental Results

Experiments have been carried out on a Pentium IV 2.4G PC, the frame size is 352*288, the processing speed per frame is about 1.5~2.0 s, the average processing time for each step is shown in Table I. If the objects segmentation is performed on the whole image, only the time spent on the classification will exceed 10 s. This shows that temporal segmentation before spatial segmentation and classification reduces the computation complexity largely.

Table I Average Processing Time for Each Step

Steps	Temporal segment	Spatial segment	Classify	Total
Time (ms)	450	482	798	1730

Segmentation results of a diving sequence are shown in Fig. 1. In this sequence, the illumination is obvious and the motion of the athlete is rapid and non-rigid, experiments results demonstrate that the proposed algorithm is not sensitive to objects' irregular movement and illumination.

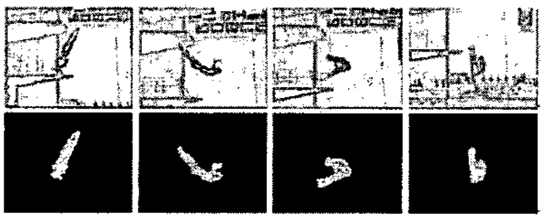
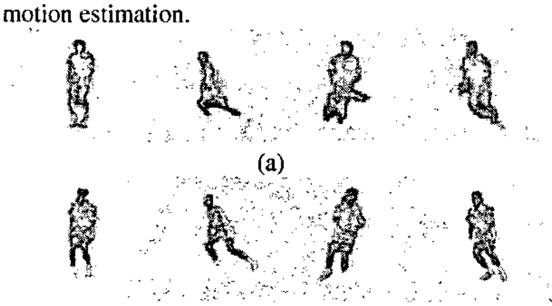


Fig. 1 Segmentation results of Diving sequence

The contrastive experimental result on the 53th、82th、125th and 142nd frame of Stepan sequence are giving in Fig. 2. (a) is the results of the algorithm proposed by Y. Tsaig [5] using region classification based on motion estimation, (b) is the results of our algorithm. According to the contrastive experiments, we can see that the region classification based on motion estimation results in wrong classification of some disoccluded background regions, however, the algorithm proposed in this paper can overcome the drawback of the region classification based on motion estimation.



(a)

(b)

Fig. 2 contrastive experiments

6. Conclusions

An automatic segmentation algorithm for moving objects in video sequences under multi-constrains is proposed. Compared with existing hybrid segmentation algorithms, the proposed algorithm has the following advantages. First, temporal segmentation before spatial segmentation and classification reduces the computation time of segmentation procession largely. Secondly, the region classification is implemented by estimating the MAP of the MRF using spatial, temporal and adjacent constraints. Doing this way, the proposed algorithm overcomes the drawback of region classification based on motion estimation, which is sensitive to objects' irregular movement and illumination. Experimental results validate the efficiency of the proposed algorithm.

7. References

[1] D. Wang, Unsupervised video segmentation based on watersheds and temporal tracking, *IEEE Trans. Circuits and Systems for Video Technology*, 1998, 8(5): 539-546.

[2] Jianping fan *et al*, Automatic image segmentation by integrating color-edge extraction and seeded region growing, *IEEE Trans. Image Processing*, 2001, 10(10): 1454-1466.

[3] Til Aach, Andre Kaup, "Statistical model-based change detection in moving video," *Signal Processing*, vol. 31, pp. 165-180, 1993.

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

[5] Y. Tsarg and A. Averbuch, "Automatic Segmentation of Moving Objects in Video Sequences: A Region Labeling Approach," *IEEE Trans. Circuits and Systems for Video Technology*, 2002, 12(7): 597-612.

[6] I. Patras, E. A. Hendriks and R. L. Lagendijk, "Video Segmentation by MAP Labeling of Watershed Segments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 326-332, 23(3), March 2001.

[7] L. Vincent and P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, *IEEE Trans. Pattern Analysis and Machine Intelligent*, 1991, 13: 583-598.

[8] P. B. Chou and C. M. Brown, "The theory and practice of Bayesian image labeling," *Int. J. Comput. Vis.*, vol. 4, pp. 185-210, 1990.

[9] M. Sonka, V. Hlavac and R. Boyle, *Image Processing, Analysis, and Machine Vision*, International Thomson Publishing, 1998.