

VIDEO OBJECT SEGMENTATION AND TRACKING USING PROBABILISTIC FUZZY C-MEANS

Jian Zhou, Xiao-Ping Zhang

Department of Electrical and Computer Engineering, Ryerson University
350 Victoria Street, Toronto, Ontario, Canada, M5B 2K3
E-mail: {jzhou, xzhang}@ee.ryerson.ca

ABSTRACT

Automatic video object segmentation and tracking is a challenging problem. In this paper, we introduce a new systematic method for fully automatic object segmentation and tracking using probabilistic fuzzy c-means and Gibbs Random Fields. The spatial segmentation is based on probabilistic fuzzy c-means clustering and Gibbs sampling. The obtained segmented mask is then refined by taking into account of motion information. Motion vectors are calculated using block matching method based on phase correlation. The motion features and their spatial relationships are used to associate the segmented regions to form video objects. Temporal tracking is achieved by projecting the blocks in current frame to the next frame. The motion-compensated prediction is carried out directly over membership matrix which is used as the initialization of probabilistic fuzzy c-means clustering for the next frame. Experimental results show that the proposed method can automatically extract and track the video object in cluttered background. The major advantages of the proposed method are its ability to deal with deformable objects and being fully automatic.

1. INTRODUCTION

Analyzing spatio-temporal patterns is a fundamental research in digital video. One important characteristic of video is its temporal dimension. Traditional video coding standards, such as MPEG-1/MPEG-2, exploit the similarities between neighboring frames and reduce the temporal redundancy by using block-based motion estimation methods. However, a human viewer does not view the video as a collection of rectangular blocks. Recently, partitioning video sequences into semantic video objects has been an active research area. The MPEG-4 [1] video standard introduces a framework for video object based coding. A video object may have arbitrary shape and may exist for an arbitrary length of time. The concept of video object not only allows more flexible options for video coding, it also supports high-level interpretation and manipulation of video contents.

Applications to object-based video representation include video surveillance for security, video editing, animation, video conference, content-based video indexing and retrieval.

Automatic video object segmentation and tracking is difficult in that most sub-problems such as spatial segmentation, motion segmentation, occlusion, video object formation, appearance/disappearance of video objects and tracking of deformable objects are all non-trivial. Thus, a simplified formulation is often used among existing techniques. For example, the background is assumed to be static, or the system is semi-automatic such that the video object boundaries are already coarsely initialized by users. Many segmentation and tracking techniques have been proposed in literatures. Classical methods are mainly based on motion estimation and motion segmentation. In [2], image sequence is decomposed into layers by estimating and clustering affine parameters. Borshukov, *et al.* [3] improved this method by replacing adaptive K-means with a merging algorithm and implementing the block-based affine modeling in a multistage. In [4], a multi-resolution iterative refinement algorithm based on Kalman filtering was proposed. More recently, many researchers [5][6] built their trackers on particle filtering framework since, in theory, particle filters can deal with non-linear and non-Gaussian estimations. Other methods based on Mean-shift algorithm [7], spatio-temporal information [8], edge maps [9] are also developed to segment and track video objects.

Due to limitations of motion estimation, methods based on motion segmentation may not give accurate object boundaries. Active contours (i.e., snakes) have been widely used to track non-rigid objects. However, most motion-based techniques generally require user initialization and need additional models to process occlusion and de-occlusion. Spatio-temporal segmentation and tracking techniques consider both spatial and temporal information. Such techniques typically have a spatial segmentation step and a merging step based on motion features.

The new method presented in this paper can be categorized into spatio-temporal in a sense we utilized the

spatial features and temporal information in different stages. The new method aims at extracting and tracking deformable video objects and is fully automatic. The segmentation algorithm is based on probabilistic fuzzy c-means clustering with integration of Gibbs random fields that is employed to compute the local conditional probability as neighborhood constraints. During image segmentation, spatially connected pixels tend to belong to the same segment. However, this constraint is usually not well utilized in classical c-means or fuzzy c-means clustering techniques. In the new method, we bring Gibbs Random Fields into probabilistic fuzzy c-means framework to compute the local conditional probabilities as spatial neighborhood constraints. For motion segmentation, the block matching method using phase correlation is used to compute the temporal features. The segmented regions are analyzed and labeled to form video objects. Motion-compensated predictions are also applied to track and estimate the interested regions for the next frame. Experimental results show that the proposed method can automatically extract and track the video object in cluttered background.

2. VIDEO OBJECT SEGMENTATION AND TRACKING

The presented new video segmentation and tracking method includes the following steps: (i) Spatial segmentation; (ii) Motion segmentation; (iii) Data association; (iv) Temporal tracking. Motion segmentation and spatial segmentation are processed in different steps. Their results are analyzed and combined in data association step to define and label the video objects. The temporal tracking is introduced as the motion-compensated predictions of regions. Details are described in the following subsections.

2.1. New Video Object Spatial Segmentation Method Based on Probabilistic Fuzzy C-means and Gibbs Sampling

As an important visual cue, the color features from the perceptually uniform CIE (Commission Internationale d'Eclairage) L*u*v color space are extracted in pixel domain. The L*u*v color space can linearize the perceptibility of color difference. Thus the difference measured in Euclidean distance is consistent with the perceptual color difference viewed by human. For each pixel, a three-dimensional color feature vector is computed. Denote $\mathbf{y}_i = [y_i^{(L)} \ y_i^{(u)} \ y_i^{(v)}]^T$ as the color feature vector for the i -th pixel, where $y_i^{(L)}$, $y_i^{(u)}$, and $y_i^{(v)}$ are L, U, and V components of pixel i in L*u*v

color space, respectively. The pixels in each frame are quantized into N (the number of clusters) colors according to the rule proposed in [10]. During spatial segmentation process, only color features are used. The spatial color segmentation process is based on probabilistic fuzzy c-means framework and Gibbs sampling.

2.1.1. Probabilistic Fuzzy C-means Clustering

Fuzzy c-means clustering techniques are generalized in [11]. In standard fuzzy c-means clustering, denote \mathbf{y}_k by the color feature vector for the k -th pixel, and given the image of N pixels, i.e., $Y = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N] \subset \mathfrak{R}^n$, the algorithm aims at finding a fuzzy partition \mathbf{U} of the N elements based on the following objective function [11]:

$$J_{FCM}(\mathbf{U}, \mathbf{V}) = \sum_{k=1}^N \sum_{i=1}^c u_{ik}^m \cdot (d_{ik})^2, \quad (1)$$

where

$$d_{ik}^2 = \|\mathbf{y}_k - \mathbf{v}_i\|_{\mathbf{A}}^2 = (\mathbf{y}_k - \mathbf{v}_i)^T \cdot \mathbf{A} \cdot (\mathbf{y}_k - \mathbf{v}_i), \quad (2)$$

in which c is the number of classes, $m \in [1, \infty)$ is the weighting exponent which controls the amount of fuzziness, u_{ik} is the degree of membership of \mathbf{y}_k to the class i , the three dimensional column vector \mathbf{v}_i represent the center of cluster i , $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_c]$ is the matrix of cluster centers, \mathbf{A} is a positive-definite weight matrix, and d_{ik} is the distance measure between sample \mathbf{y}_k and cluster center \mathbf{v}_i . The partition matrix \mathbf{U} is also called membership matrix.

A probabilistic fuzzy c-means clustering is introduced in [12]. The fusion of probabilistic and fuzzy information can be represented as:

$$u_{ik}^* = u_{ik} \cdot p_{ik}, \quad (3)$$

where $k = 1, 2, \dots, N$, and p_{ik} is the probability of data point k belonging to cluster i . Based on the above modification, the objective function becomes:

$$J_{FCM}(\mathbf{U}, \mathbf{V}) = \sum_{k=1}^N \sum_{i=1}^c (u_{ik}^*)^m \cdot (d_{ik})^2. \quad (4)$$

The cluster centers and the membership matrix can be iteratively updated as:

$$\mathbf{v}_i = \frac{\sum_{k=1}^N (u_{ik}^*)^m \cdot \mathbf{y}_k}{\sum_{k=1}^N (u_{ik}^*)^m}, \quad (5)$$

$$u_{ik} = \frac{\sum_{i=1}^c u_{ik} \cdot p_{ik}}{\sum_{s=1}^c (d_{ik}/d_{sk})^{2/(m-1)}}. \quad (6)$$

In [12], a method based on indicator and *ordinary kriging* is proposed to calculate p_{ik} . Different from [12],

we propose to bring Gibbs sampling into fuzzy c-means framework. Details are presented in the following section.

2.1.2. Integrating Probabilistic Fuzzy C-means and Gibbs Sampling

Standard fuzzy c-means has been a popular technique for image segmentation [13]. However, the relationship between pixels in spatial domain is not well utilized. Adding the spatial constraints directly in the objective function could be one possible solution. The other direction is to relax the constraints $\sum_{i=1}^c u_{ik} = 1$ and include spatial probabilistic information. The probabilistic fuzzy c-means framework introduces another way to associate the probabilities with membership matrix. That makes it possible to incorporate additional constraints and prior knowledge into the learning process.

Markov random fields (MRF) is the two-dimensional extension of Markov models. An image in spatial domain can be modeled by Markov random fields since it captures the context-dependent relationship of pixels within a neighborhood area. According to [14], an equivalent relationship is built upon between Markov random fields and Gibbs distributions. Thus the computationally more tractable Gibbs random fields can be used to compute the conditional probabilities.

To spatially segment the video frames, we proposed a new method which integrate the Gibbs Random Fields into probabilistic fuzzy c-means framework. The Gibbs sampler is used to compute the local conditional probabilities as local neighborhood constraints. Those probabilities are directly associated with membership matrix and updated at each iteration. In Gibbs sampling, the local conditional probability in spatial domain is of the form [15]:

$$p(z(\mathbf{x}_i) | z(\mathbf{x}_j), \forall \mathbf{x}_j \neq \mathbf{x}_i) = Q_{\mathbf{x}_i}^{-1} \cdot \exp\left\{-\frac{1}{T} \sum_{C|\mathbf{x}_i \in C} V_C(z(\mathbf{x}) | \mathbf{x} \in C)\right\}, \quad (7)$$

where

$$Q_{\mathbf{x}_i} = \sum_{z(\mathbf{x}_i) \in \Gamma} \exp\left(-\frac{1}{T} \cdot \sum_{C|\mathbf{x}_i \in C} V_C(z(\mathbf{x}) | \mathbf{x} \in C)\right), \quad (8)$$

in which $z(\mathbf{x}) \in \Gamma = \{0, 1, \dots, L-1\}$ is a discrete-valued random field evaluated at location \mathbf{x} , C is a *clique* which consists of a single pixel or a set of pixels, $Q_{\mathbf{x}_i}$ is a normalizing constant such that probabilities sum up to 1, T is a parameter and also known as *temperature*, $V_C(\cdot)$ are functions of the states of the pixels in the *cliques* set. The exponent function during the implementation is chosen as:

$$-\frac{1}{T} \cdot z_{ij} (\alpha_1 + \alpha_2 (z_{i-1,j} + z_{i+1,j}) + \alpha_2 (z_{i,j-1} + z_{i,j+1})), \quad (9)$$

where $v_{i,j}$ is the class label at location (i, j) , and the α_i 's are constants that depends on the local configuration of on the cliques [15].

The new segmentation algorithm based on Gibbs sampler and fuzzy c-means are summarized as follows:

1. Set values for the number of clusters C , the weighing exponent m , the termination criterion \mathcal{E} , and the maximum iteration steps. The number of clusters is determined by the relative smoothness of the whole image according to the guideline in [10].
2. Initialize the membership matrix \mathbf{U} .
3. Remove the fuzzyness in the membership matrix by choosing the maximum element at each column, construct the image in pixel domain, and compute the local conditional probability p_{ik} using (7) which is based on Gibbs sampler.
4. Evaluate the current cluster centers according to (5).
5. Update the membership matrix according to (6).
6. Compare the current membership matrix and the one obtained in previous loop, if $\|\mathbf{U}^{(t+1)} - \mathbf{U}^{(t)}\| < \mathcal{E}$ or maximum iteration steps are reached, then stop; otherwise, return to step 3.
7. Remove the fuzzyness in the membership matrix by choosing the maximum element in each column. The result is the segmentation mask.

Several parameters during implementation are chosen as follows: the termination condition $\mathcal{E} = 1e-5$; the temperature is chosen as 1.

Since the proposed spatial segmentation mainly depends on the colors, the learning process allows regions with any arbitrary shape to be detected as long as the interested neighboring regions are distinguishable by their colors, and thus makes it possible to segment and track de-formable objects.

After the spatial segmentation mask is obtained, each region is given a unique label. A median filter is then applied to fill isolated small holes.

2.2. Motion Segmentation

In motion segmentation, the block matching method based on phase correlation [15] is used to compute the motion feature vectors. Note that as initial segmentations, motion segmentation and spatial segmentation are separately processed. To extract motion feature during motion segmentation, each video frame is partitioned into 16-by-16 non-overlapping blocks. Fourier transform is used to calculate the spectrum. The motion vectors can be found by locating the peaks in the phase-correlation function, since a translational shift in spatial domain results in a phase change in spectrum domain. For each block, a two dimensional motion vector $v = [v_x, v_y]$, is obtained, which represent the translational shift along the horizontal and

vertical directions. Block matching using phase correlation has some desirable properties. First, it is relatively insensitive to illumination changes since the shifts in the mean value do not affect Fourier phase. Second, it is computationally efficient, compared to other motion estimation techniques such as pixel-level optical flow estimation methods. Note that this method models the motions as two-dimensional shift between two image blocks. Therefore, complex motions such as rotational motions cannot be captured.

In addition, the texture feature is utilized during motion segmentation. The reliability of motion features depends on the variations within a block. For example, if there is no texture within an area, i.e. the color in that area is almost uniform, good matches can always be found even there are no motions in the blocks. Those large motion vectors should be considered as noises. To this end, we introduced a criterion which evaluates the texture within a block to validate or reject the motion features. During implementation, we choose the variance of the block to evaluate the amount of texture within a block. If the matched block contains little texture, the motion vector for this block is rejected (motion vectors are assigned zero). Otherwise, it is accepted.

After blocks with motions are identified, a post-filtering based on the image dilation and erosion is applied to absorb nearby neighboring blocks. Then the output is the motion segmentation mask.

Note that static background is not assumed. Otherwise, a simple technique such as the image difference can be used to directly identify the region of interest (ROI). In our work, motions are estimated over all the blocks. As long as the background motion is relatively smaller than that of ROI, the motion segmentation described above may still be able to identify the interested regions.

2.3. Data Association

The motion vectors obtained by block matching method are assigned to each region. Regions which show consistency in motions and are spatially connected are identified and grouped together to form semantic objects. This data association step is important since it bridges the low-level features and high-level semantics. There are two tasks in this step. The first task is to build semantic video objects from low-level segmented regions. The second task is to label the video objects and keep track of their labels over the time.

The video objects formation is achieved by combing the spatial and motion segmentation results. The motion vector for the block is assigned to each pixel within the block, and thus the motion feature is extended from block level to pixel level. Then, in each spatial segmented region, the summation of absolute values of motion

vectors is calculated. A predefined threshold is used to select those candidate regions that show certain amount of motions. Other regions are removed and assumed as static regions. Then, a dilation operator is applied on motion segmentation masks. The output, combined with the candidate regions, is used to calculate the final output. The overlapped areas are identified as semantic video objects.

After video objects are detected, they are put in correspondence over time. We implemented a solution that uses positions (centroids of video objects) and normalized histogram to compute the dissimilarity between video objects. In the video object pool in previous frame, the one which has the minimum distance with the current video object is associated with each other and labeled as the same object.

2.4. Temporal Tracking

We consider temporal tracking as the motion-compensated predictions of the interested regions.

Temporal information is directly applied in membership matrix instead of pixel domain. The reason is that the membership matrix itself already contains all the information about how the image is partitioned into regions since the spatial segmentation is essentially the removal of fuzzyness in the membership matrix. The advantage of performing tracking on membership matrix is that the matrix estimated by motion-compensated from current frame can be directly used as the initialization for the next frame. This strategy can reduce the learning time since the initial membership matrix for the next frame is already a close approximation of true membership matrix which will be learnt. The similar idea has been proposed in [16]. However, the motion features used in [16] is pixel-level optical flow motion estimation which is computationally demanding.

The motion vector for a block can be considered as the average motion of all pixels in that block. Therefore, we use the block motion vector to approximate the pixel motions. Denote $\mathbf{v}(x_1, x_2) = [v_x \ v_y]$ by the motion feature vector of the pixel at location (x_1, x_2) and denote \widehat{S} by the estimated spatial segmentation result for the next frame. \widehat{S} can be obtained by:

$$\widehat{S}^{(n+1)}(x_1, x_2) = S^{(n)}(x_1 - v_x, x_2 - v_y). \quad (10)$$

Then, the initialization of membership matrix for the next frame is given by:

$$u_{ik} = \begin{cases} 1 & \text{if } \widehat{S}^{(n+1)} = i \\ 0 & \text{otherwise} \end{cases}. \quad (11)$$

3. EXPERIMENTAL RESULTS

The image sequence “Hall Monitor” is used to test the proposed video object segmentation and tracking method. The test data contains 298 frames and each frame is of size 353-by-240. The spatial and motion segmentation results are represented as masks which are superimposed to the original images for display purpose. Figure 1 and Figure 2 show the motion segmentation results for frame 31. It can be seen that motion feature alone can only produce very coarse boundaries. Figure 3 and Figure 4 show the spatial segmentation results based on probabilistic fuzzy c-means and Gibbs random fields. An example of spatial segmentation is shown in Figure 4. As it can be seen that some background areas are mistakenly segmented into potential region of interests since their colors are very similar. Such errors can be eliminated during data association step. Data association combines both spatial segmentation results and motion segmentation results. We first identify the spatial regions which show a certain amount of motions. An example is shown in Figure 5. Then, motion segmentation masks are combined with the identified region to create the final segmented results. The temporal tracking through motion-compensated prediction of membership matrix is verified in the experiments. Figure 7 illustrates the tracking results for frame 23, 32, 36, and 42. As it can be seen that the tracking performs very well on the major target but the boundaries are not very accurate. The error is mainly caused by the estimation noise from motion segmentation.

4. CONCLUSIONS

In this paper, we present a new fully automatic video segmentation and tracking method that combines probabilistic fuzzy c-means and Gibbs random fields. Color, motion and texture features are utilized together. In the spatial segmentation process, Gibbs sampling is integrated into probabilistic fuzzy c-means framework to compute the local conditional probabilities as spatial constraints. Motion segmentation is based on block matching method using phase correlation. In data association, motion segmentation masks and spatial segmentation masks are combined together to create video objects. The temporal tracking is performed for the motion-compensated prediction of membership matrix. The proposed method brings the probabilistic fuzzy c-means clustering into video object extraction and tracking, and integrated Gibbs random fields into the framework. The experimental results show that the proposed method can detect and track de-formable objects and being fully automatic. We note that complex motions such as rotational motions cannot be captured due to the limitation of the motion model. Future work will include such situations and the occlusion problem under this framework.

5. REFERENCES

- [1] I.E.G. Richardson, *H.264 and MPEG-4 Video Compression: Video Coding for Next-Generation Multimedia*, Wiley, 2003.
- [2] J.Y.A. Wang, E.H. Adelson, “Representing moving images with layers,” *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625-638, 1994.
- [3] G.D. Borshukov, G. Bozdagi, Y. Altunbasak, and A.M. Tekalp, “Motion segmentation by multistage affine classifications,” *IEEE Transactions on Image Processing*, vol. 6, no. 11, pp. 1591-1594, 1997.
- [4] F.G. Meyer and P. Bouthemy, “Region-based tracking using affine motion models in long image sequences,” *CVGIP: Image Understanding*, vol. 60, no. 2, pp. 119-140, 1994.
- [5] Y. Wu and T.S. Huang, “A co-inference approach to robust visual tracking,” *ICCV’01*, vol. 2, pp. 26-33, 2001.
- [6] D. Serby, E.K. Meier, and L. Van Gool, “Probabilistic object tracking using multiple features,” *ICPR’04*, vol. 2, pp. 184-187, 2004.
- [7] B. Han, D. Comanicui, Y. Zhu, and L. Davis, “Incremental density approximation and kernel-based Bayesian filtering for object tracking,” *CVPR’04*, vol. 1, pp. 638-644, 2004.
- [8] F. Moscheni, S. Bhattacharjee, and M. Kunt, “Spatiotemporal segmentation based on region merging,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 897-915, 1998.
- [9] C. Kim and J.N. Hwang, “Fast and automatic video object segmentation and tracking for content-based applications,” *IEEE Transactions on Circuits and Systems For Video Technology*, vol. 12, no. 2, pp. 122-129, 2002.
- [10] Y. Deng, C. Kenney, M.S. Moore, and B.S. Manjunath, “Peer Group Filtering and Perceptual Color Image Quantization,” *Proc. of IEEE Intl. Symposium on Circuits and Systems*, vol. 4, pp.21-24, 1999.
- [11] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
- [12] T.D. Pham, “Image Segmentation using Probabilistic Fuzzy C-Means Clustering,” *ICIP’01*, vol. 21(3), pp. 722-725, 2001.
- [13] Z. Chi, H. Yan, T. Pham, *Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition*, World Scientific, Singapore, 1996.
- [14] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6(6), pp. 721-741, 1984.
- [15] A.M. Tekalp, *Digital video processing*, Prentice-Hall, Upper Saddle River, NJ, USA, 1995.
- [16] R. Castagno, “Video Segmentation Based on Multiple Features for Interactive and Automatic Multimedia Applications,” *Ph.D. Thesis*, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1998.

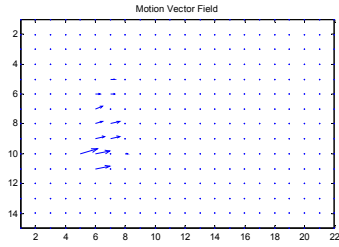


Figure 1: Motion vector field computed by block matching method using phase correlation.

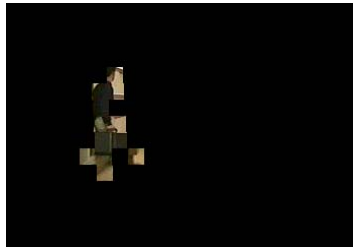


Figure 2: Motion Segmentation result.

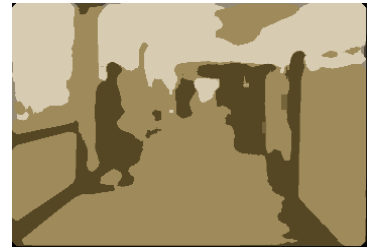


Figure 3: Regions obtained from spatial segmentation.



Figure 4: Spatial segmentation result.



Figure 5: Spatial region that contains motions.



Figure 6: Final results after data association.



(a)



(b)



(c)



(d)

Figure 7: Tracking results at (a) frame 23 (b) frame 32 (c) frame 36 (d) frame 42.