A NOVEL VIDEO OBJECT TRACKING APPROACH BASED ON KERNEL DENSITY ESTIMATION AND MARKOV RANDOM FIELD

Zhi Liu[1,2], Liquan Shen[1], Zhongmin Han[1], Zhaoyang Zhang[1,2]

[1]School of Communication and Information Engineering, Shanghai University, China
[2]Key Lab of Advanced Display and System Application (Shanghai University), Ministry of Education, China
Tel.: +86-21-56331659-801, Fax: +86-21-56331194, Email: liuzhisjtu@163.com

ABSTRACT

In this paper, we propose a novel video object tracking approach based on kernel density estimation and Markov random field (MRF). The interested video objects are first segmented by the user, and a nonparametric model based on kernel density estimation is initialized for each video object and the remaining background, respectively. A temporal saliency map is also initialized for each object to memorize the temporal trajectory. Based on the probabilities evaluated on the non-parametric models, each pixel in the current frame is first classified into the corresponding video object or background using the maximum likelihood criterion. Starting from the initial classification result, a MRF model that combines spatial smoothness and temporal coherency is selectively exploited to generate more reliable video objects. The nonparametric model and the temporal saliency map for each video object are updated and propagated for the future tracking. Experimental results on several MPEG-4 test sequences demonstrate the good segmentation performance of our approach.

Index Terms— Video object segmentation, Video object tracking, Kernel density estimation, Markov random field

1. INTRODUCTION

Video object segmentation provides a content-based representation for the pixel-based or block-based source video, and it greatly benefits many multimedia applications including region of interest (ROI) coding, intelligent video surveillance, interactive video editing and manipulation, and specific object query and retrieval. There have been many approaches proposed for video object segmentation, both automatic and semi-automatic dependent on the target applications. Although researches on video object segmentation have progressed a lot in the recent years, automatic segmentation is still applicable to moving objects [1, 2] or some specific objects with a prior knowledge [3]. Semi-automatic segmentation is a more practical way to segment generic objects in a variety of video sequences [4-9]. A tracking based paradigm is commonly adopted in both automatic and semi-automatic video segmentation approaches, in which the video objects in the first frame can be automatically or interactively segmented by the end user, and the video objects of subsequent frames are generated by tracking of previous video objects. Therefore, the initial video object may be obtained by different ways, but the following video object tracking algorithm is interchangeable for both automatic and semi-automatic approaches.

The problem of video object tracking can be explicitly or implicitly transformed into a classification problem. Some approaches directly perform a binary classification on the pixel or region level [4-6]. The region classification scheme is proposed in [4], in which each segmented region in the current frame is backward projected into the previous frame, and then is classified into the video object or the background based on the overlapped area between the projected region and the previous object. The main drawback of region classification is that the extracted video objects are totally determined by the spatial region partition result of the current frame. The pixel classification is incorporated in [5, 6] to obtain a finer video object at the accuracy of pixel level, but the classification is still individually performed on each pixel based on the projection result. Although the above approaches are efficient, its main limitation is the lack of a clearly defined model for video object, and the spatial context from neighboring pixels or regions are totally ignored. Other approaches exploit some form of parametric models such as Gaussian mixture model (GMM) to compactly represent the whole video object [7, 8] or homogenous sub-object regions [9]. The model is propagated frame by frame and updated using the expectation maximization (EM) algorithm during the whole tracking process. However, the number of components in GMM needs to be predetermined, and the underlying assumption of Gaussian distribution is not always applicable to any sequence. The only use of chromatic features in GMM usually results in that other parts with a similar color distribution in the scene are likely to assign to the video object [7]. Although the addition of spatial features into GMM can be used to coarsely model the spatial extent of an object, the updating of model is not well adapted to the spatial extent over time [8, 9].

In this paper, we propose to use kernel density estimation [10] based nonparametric models to represent both video objects and background, which needs no assumption of underlying distribution compared with the commonly used parametric models mentioned above. The proposed approach can also be considered as a pixel classification approach, but the classification is not only dependent on the estimated nonparametric models, but also dependent on the classification result of the neighboring pixels by using Markov Random Field (MRF) [11]. Spatial context and temporal coherency modeled in MRF are exploited to ensure a more robust segmentation performance.

The rest of this paper is organized as follows. Section 2 describes the nonparametric modeling based on kernel density estimation for video objects and background. Section 3 details the proposed MRF classification approach for video object tracking. Experimental results are presented in Section 4, and conclusions are given in Section 5.

2. NONPARAMETRIC MODELING

Initial video objects can be first obtained for the following video object tracking in an automatic or interactive way. In this paper, the interested video objects are manually segmented in the first frame by the user, and an interactive segmentation tools proposed in our previous work [6] is employed to conveniently obtain the desired video objects. Multiple video objects may be segmented from the first frame and simultaneously tracked in the subsequent frames. An example of initial object segmentation is illustrated in Fig. 1. The first frame of the sequence *Table Tennis* is shown in Fig. 1(a), in which the interested objects and the background are marked by red and blue scribbles, respectively. The two extracted video objects, i.e., the arm holding the racket and the ball, are shown in Fig. 1(b). The obtained initial video objects and background are then represented using the nonparametric models based on kernel density estimation.

Generally, assume there are totally n video objects extracted in the first frame, all pixels in the inner region of each video object vo_k constitute a foreground pixel set denoted as $S_k (k = 1,2,...n)$, and the remaining pixels that do not belong to any video object constitute a background pixel set denoted as S_0. The three pixel sets for the two video objects and the background are shown in Fig. 1(c). Each pixel set is then used as samples to initialize a non-parametric model using kernel density estimation method for each video object or background, respectively. The feature representing the pixel sample is commonly denoted as a d-dimensional vector $\mathbf{x}_i \in \mathbb{R}^d$. Specifically, the color and position features are jointly used, $\mathbf{x}_i = [\mathbf{c}_i \ \mathbf{p}_i]^T (d = 5)$, in which the color feature $\mathbf{c}_i$ is denoted as (Y,U,V) due to the adopted YUV color space, and the position feature $\mathbf{p}_i$ is denoted as (x, y). Given the pixel sample set $S_k (k = 0,1,...n)$ for any video object or background, the probability that a candidate pixel sample $\mathbf{x}_c$ belongs to the corresponding video object or background is defined as

$$P(\mathbf{x}_c \mid S_k) = \frac{1}{|S_k|} \sum_{\mathbf{x}_i \in S_k} K_{\mathbf{H}_k}(\mathbf{x}_c - \mathbf{x}_i) \tag{1}$$

where $K_{\mathbf{H}_k}$ is a kernel function with the symmetric positive definite 5×5 bandwidth matrix $\mathbf{H}_k$. Specifically, d-variate Gaussian kernel is selected for its continuity, differentiability and locality properties, and the kernel function $K_{\mathbf{H}_k}$ is defined as

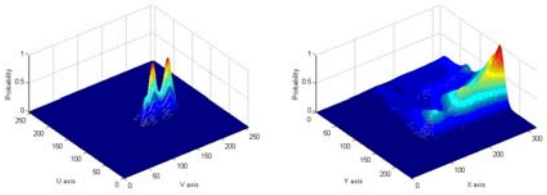
$$K_{\mathbf{H}_k}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\mathbf{H}_k|^{1/2}} \exp(-\frac{1}{2} \mathbf{x}^t \mathbf{H}_k^{-1} \mathbf{x}) \tag{2}$$

For a computationally efficient estimation of bandwidth matrix, it is reasonable to assume that the bandwidth for each component of the feature vector has no correlation with other components, and thus the probability density function in Eq. (2) can be further simplified as

$$P(\mathbf{x}_c \mid S_k) = \frac{1}{|S_k|} \sum_{\mathbf{x}_i \in S_k} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{k,j}^2}} \exp[-\frac{1}{2} \frac{(x_{c,j} - x_{i,j})^2}{\sigma_{k,j}^2}] \tag{3}$$

The bandwidth matrix is simplified as a diagonal matrix, i.e., $\mathbf{H}_k = diag(\sigma_{k,1}^2, \sigma_{k,2}^2, ..., \sigma_{k,d}^2)$, and the bandwidth for each feature component can be estimated independently. The binned kernel density estimator [12] is adopted due to its computational efficiency and the good approximation accuracy. Using the segmented arm object in Fig. 1(c) as an example, the marginal probability maps evaluated on the chrominance features (U,V) is shown in Fig. 1(d), and the marginal probability maps evaluated on the position features (x, y) is shown in Fig. 1(e), respectively.

(a) (b) (c)



(d) (e)

Fig. 1. Video object initialization and probability maps based on kernel density estimation.

Besides the above probability model for color and position features in Eq. (3), a temporal saliency map $TSM_k^t (k = 1,2,...,N)$ is also initialized for each video object

$$TSM_k^1(\mathbf{p}_i) = \begin{cases} 1, \mathbf{p}_i \in vo_k \\ 0, otherwise \end{cases} \tag{4}$$

where the superscript $t = 1$ denotes the first frame. The temporal saliency map is used to memorize the temporal trajectory of each video object till the current frame t, and its value $TSM_k^t(\mathbf{x}_i)$ can be considered as a prediction of possibility that the pixel $\mathbf{x}_i$ belongs to vo_k in the current frame t. The use and updating of TSM_k^t will be described in the following section.

3. MRF CLASSIFICATION

Based on the probability density functions defined for each video object with the form of Eq. (3), the most probable video object label k_i for each pixel $\mathbf{x}_i$ in the current frame is first determined by the following maximum likelihood criterion

$$k_i = \arg \max_{k=1,2,...,N} P(\mathbf{x}_i \mid S_k) \tag{5}$$

The likelihood that the pixel $\mathbf{x}_i$ belongs to the most probable video object rather than background is then defined as

$$lh(\mathbf{x}_i) = \ln \frac{P(\mathbf{x}_i \mid S_{k_i})}{P(\mathbf{x}_i \mid S_0)} \tag{6}$$

A label field $L = \{l_i \mid i \in \Lambda, l_i \in \Omega_i\}$ that represents the initial classification result is thus generated based on the above likelihood

$$l_i = \begin{cases} k_i, if \ lh(\mathbf{x}_i) > 0 \\ 0, otherwise \end{cases} \tag{7}$$

where the value of label l_i indicates whether the pixel $\mathbf{x}_i$ belongs to the most probable video object or the background. The set Λ contains the indices for pixels defined on the image lattice, and the label set $\Omega_i = \{k_i, 0\}$ for each pixel $\mathbf{x}_i$ is a limited set containing only two labels, which greatly reduces the computation load in the case of multiple objects.

The number of connected components n_c in the above initial label field is then determined by the connected component labeling algorithm. If $n_c = n$, it implies that each connected component may well represent each video object. In this case, the background is usually greatly different from video objects, and thus the initial classification is sufficient for obtaining a clean object segmentation result. However, n_c is usually greater than n for many sequences, it implies that some noisy regions and small gaps appear in the initial label field, which cannot represent an accurate video object segmentation result. For example, the 18th frame of the sequence *Table Tennis* is shown in Fig. 2(a), and the initial classification result shown in Fig. 2(b) contains small noisy regions and holes.

In the case of $n_c > n$, MRF model is exploited to refine the initial classification result to obtain more complete and compact video objects. Given the set of observations $O = \{o_i \mid i \in \Lambda\}$ where each observation is denoted as $o_i = \{\mathbf{x}_i, P(\mathbf{x}_i \mid S_0), P(\mathbf{x}_i \mid S_{k_i})\}$, the goal of MRF classification is to estimate an optimized configuration L_{opt}, which is obtained by minimizing the following energy function of the equivalent Gibbs distribution

$$U(L \mid O) = \sum_{i \in \Lambda} V_i^P(L,O) + \sum_{i \in \Lambda} V_i^T(L,O) + \sum_{(i,j) \in \mathbb{C}} V_{ij}^S(L,O) \qquad (8)$$

The first term in the energy function is a data-driven term, which represents the likelihood of the pixel $\mathbf{x}_i$ belonging to the most probable video object rather than the background

$$V_i^P(L,O) = \begin{cases} -w_1 \cdot lh(\mathbf{x}_i), l_i = k_i \\ w_1 \cdot lh(\mathbf{x}_i), l_i = 0 \end{cases} \qquad (9)$$

The second term is a temporal continuity term, which considers the trajectories of tracked objects in previous frames and thus enhances the coherency of objects through the whole sequence. The temporal saliency map TSM_k^t for the current frame is first updated as follows

$$\begin{aligned} TSM_k^t(\mathbf{p}_i) &= \max\{TSM_k^{t-1}(\mathbf{p}_i + \mathbf{mv}_i) + 1, 1\}, \mathbf{p}_i \in vo_k \\ TSM_k^t(\mathbf{p}_i) &= \max\{TSM_k^{t-1}(\mathbf{p}_i + \mathbf{mv}_i) - 1, 0\}, \mathbf{p}_i \notin vo_k \end{aligned} \qquad (10)$$

where $\mathbf{mv}_i$ is the estimated motion vector of the corresponding MB containing the pixel $\mathbf{x}_i$. It is observed from Eq. (10) that the previous video object tracking results are propagated into the current temporal saliency map using the estimated motion vectors. The temporal continuity term is then defined as

$$V_i^T(L,O) = \begin{cases} -w_2 \cdot TSM_{k_i}^t(\mathbf{p}_i), l_i = k_i \\ w_2 \cdot TSM_{k_i}^t(\mathbf{p}_i), l_i = 0 \end{cases} \qquad (11)$$

For the example frame in Fig. 2(a), the temporal saliency maps for the arm and ball are shown in Figs. 2(c) and (d), respectively. It can be seen that the trajectories of arm and ball are memorized in the two maps, in which brighter pixels indicate a higher saliency belonging to the corresponding video object.

The last term is a spatial smoothness term, which is defined on the clique set $\mathbb{C}$ containing all the two-site cliques $(\mathbf{p}_i, \mathbf{p}_j)$

$$V_{ij}^S(L,O) = \begin{cases} w_3 \dfrac{\mathbf{d}(\mathbf{c}_i, \mathbf{c}_j) - c_l}{c_h - c_l}, l_i = l_j \\ w_3 \dfrac{c_h - \mathbf{d}(\mathbf{c}_i, \mathbf{c}_j)}{c_h - c_l}, l_i \neq l_j \end{cases} \qquad (12)$$

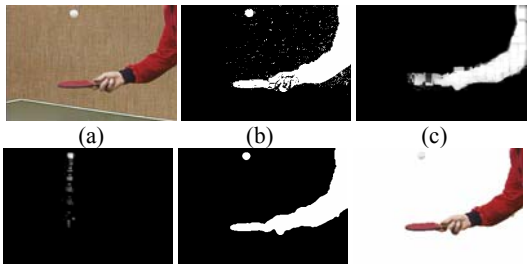


Fig. 2. Illustration of MRF classification process.

where $\mathbf{d}(\mathbf{c}_i, \mathbf{c}_j)$ is the Euclidean distance between the two color features $\mathbf{c}_i$ and $\mathbf{c}_j$ and is normalized into the range $[0,255]$, and c_h and c_l is the high and low limit for color dissimilarity. The adjacent pixels $\mathbf{p}_i$ and $\mathbf{p}_j$ are considered to exhibit the same color if the color distance $\mathbf{d}(\mathbf{c}_i, \mathbf{c}_j)$ is smaller than c_l, while they are considered to exhibit totally different color when $\mathbf{d}(\mathbf{c}_i, \mathbf{c}_j)$ is greater than c_h. Eq. (12) indicates that the adjacent pixels with similar color are likely to belong to the same object, while they are likely to belong to different objects when they exhibit different color. The two color limits c_h and c_l are set to 16 and 80, respectively.

The three weights w_1, w_2 and w_3 determine the relative contribution of the three energy terms. We found by the experiments that $w_1 = 0.25$, $w_2 = 0.1$ and $w_3 = 1$ lead to a satisfactory classification result. The minimization of the energy function $U(L \mid O)$ is performed using high confidence first (HCF) method. The resultant label field from the above MRF classification represents the video object mask as shown in Fig. 2(e), and the segmented two video objects are shown in Fig. 2(f).

The MRF classification result in the current frame is finally employed to update the nonparametric models for each video object and background. The pixel sample set generated for vo_k in the current frame t is denoted as $S_k^t = \{\mathbf{x}_i \mid l_i = k\}$, and the pixel sample set S_k used for estimating the object model is then updated as $S_k = \bigcup_{m=t-r+1}^{t} S_k^m$, which indicates that the tracked video objects in the most recent r frames are considered in the pixel sample set S_k. Accordingly, the corresponding non-parametric model based on kernel density estimation is updated using Eq. (3). The frame number r is set to 3 in our experiments, a moderate value that makes a good balance between the sensitivity and robustness of nonparametric models of video objects.

4. EXPERIMENTAL RESULTS

The proposed video object tracking algorithm is evaluated on several MPEG-4 test sequences. Experimental results on three of them are shown in Figs. 3, 4, and 5. For all sequences, video objects in the first frame are interactively segmented and the non-parametric model is initialized using kernel density estimation, like the example described in Section 2.

The first sequence *Bream* is with a uniform background, and the segmented objects shown in Fig. 3 is actually the initial

classification result based on the likelihood defined in Eq. (6), while the following MRF classification is skipped since there is only one connected component for the only one object in the initial label field. The second sequence *Table Tennis* is with a clutter background and a medium amount of motion. The interested two objects are the arm holding the racket and the ball. The third sequence *Foreman* is with a complex moving background that exhibits a low contrast with the interested object, the talking man. For these two sequences, the initial classification result cannot accurately represent the video objects, and the segmentation quality of video objects is gradually degraded in the whole sequence if the model is directly updated using the initial classification results. Therefore, MRF classification is exploited to obtain a refined object segmentation result. It can be seen from Figs. 3-5 that the video objects with good subjective visual quality are obtained during the whole tracking process.

The proposed approach is first subjectively compared with the GMM based object segmentation approach in [8] on the sequence *Foreman*. The extracted video object in the first frame is the same for both approaches, and some segmented objects from selective frames using both approaches are shown in Fig. 5. The number of Gaussian components in GMM is automatically determined by the criterion of Minimum Description Length (MDL). Comparing the segmentation results on the same frames (see the top and bottom rows), it is obvious that our segmented objects exhibit a better visual quality, while small isolated regions and holes in the bottom row are very annoying for human visual system. The segmentation performance of the proposed approach is further objectively evaluated using two measurements, that is, precision and recall. We manually segment the video object from the sequence *Foreman* as the ground truth, and the average precision and recall are 96.59% and 96.18% using our approach, while 96.54% and 95.15% using the GMM based segmentation approach [8]. The better segmentation quality of our approach is also demonstrated by the relatively higher values on both objective measurements.

5. CONCLUSION

We have presented a novel video object tracking approach based on kernel density estimation and Markov random field, which can be exploited in both automatic and semi-automatic segmentation. Each video object and background is represented by the kernel density estimation based nonparametric model, and initialized with a temporal saliency map, respectively. Using the maximum likelihood criterion, each pixel in the input frame is first classified into video object or background. The Markov random field that suitably models spatial smoothness and temporal coherency is selectively exploited to refine the classification result for more accurate video objects. The non-parametric models and temporal saliency maps are updated and propagated during the whole tracking process. Experimental results show that our approach can efficiently track video objects with good visual quality.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 60602012 and 60572127, by the Development Foundation of Shanghai Education Committee under grant No. 05AZ43, and by the Special Research Foundation of Shanghai Excellent Youth University Teacher Training (2006).

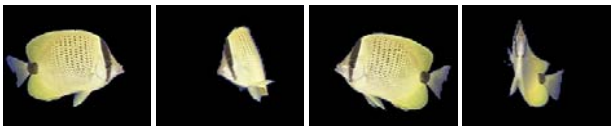


Fig. 3. Tracking results for *Bream* (Frames: 1, 120, 180, 220).

Fig. 4. Tracking results for *Table Tennis* (Frames: 5, 10, 20, 30).

Fig. 5. Tracking results for *Foreman* using our approach (top row) and GMM based segmentation approach [8] (bottom row). (Frames: 30, 60, 90, 150).

REFERENCES

[1] Y. Tsaig, and A. Averbuch, "Automatic segmentation of moving objects in video sequences: A region labeling approach," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 12, no. 7, pp. 597-612, 2002.

[2] H.F. Xu, A.A Younis, and M.R. Kabuka, "Automatic moving object extraction for content-based applications," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 14, no. 6, pp. 796-812, 2004.

[3] Z. Liu, J. Yang, and N.S. Peng, "An efficient face segmentation algorithm based on binary partition tree," *Signal Process. Image Commun.*, vol. 20, no. 4, pp. 295-314, 2005.

[4] C. Gu, and M.C. Lee, "Semantic video object tracking using region-based classification," *IEEE ICIP*, vol. 3, pp. 643–647, 1998.

[5] C. Kim, and J.N. Hwang, "Video object extraction for object-oriented applications," *Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 29, no. 1-2, pp. 7-21, 2001.

[6] Z. Liu, J. Yang, and N.S. Peng, "Semi-automatic video object segmentation using seeded region merging and bidirectional projection,"*Pattern Recogn. Lett.* vol. 26, no. 5, pp. 653-662, 2005.

[7] S. Marlow, and N.E. O'Connor, "Supervised object segmentation and tracking for MPEG-4 VOP generation," *IEEE ICPR*, vol. 1, pp. 1125-1128, 2000.

[8] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 384-396, 2004.

[9] D.J. Thirde, G.A. Jones, and J. Flack, "Spatio-temporal semantic object segmentation using probabilistic sub-object regions," *Proc. BMVC*, pp. 163-172, 2003.

[10] D.W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York, 1992.

[11] S.Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer, Berlin, 1995.

[12] P. Hall, and M. Wand, "On the accuracy of binned kernel estimators," *Journal of Multivariate Analysis*, vol. 56, no. 2, pp. 165-184, 1995.