Energy Minimization Based Semantic Video Object Extraction

김동현 최성환 김봉조 신형철 손광훈

연세대학교 전기전자공학과

khsohn@yonsei.ac.kr

Energy Minimization Based Semantic Video Object Extraction

Kim, Donghyun Choi, Sunghwan Kim, Bongjoe Shin, Hyungchul Sohn, Kwanghoon

School of Electrical and Electronic Engineering, Yonsei University

Abstract

In this paper, we propose a semi-automatic method for semantic video object extraction which extracts meaningful objects from an input sequence with one correctly segmented training image. Given one correctly segmented image acquired by the user's interaction in the first frame, the proposed method automatically segments and tracks the objects in the following frames. We formulate the semantic object extraction procedure as an energy minimization problem at the fragment level instead of pixel level. The proposed energy function consists of two terms: data term and smoothness term. The data term is computed by considering patch similarity, color, and motion information. Then, the smoothness term is introduced to enforce the spatial continuity. Finally, iterated conditional modes (ICM) optimization is used to minimize energy function in a globally optimal manner. The proposed semantic video object extraction method provides faithful results for various types of image sequences.

1. INTRODUCTION

A semantic video object is a meaningful part in image sequences. Since many content-based video applications such as video library [1], 2D to 3D video conversion [2], and MPEG-4 VOP [3] require object based functionality, the semantic video object extraction is very important issue.

In the literature, a large number of the semantic video object extraction methods have been proposed. These methods can be divided into automatic and semi-automatic methods. The automatic extraction method typically segments an image into homogeneous region by aggregating local cues such as color, texture, or motion. In [4], the authors proposed an automatic extraction method for video object segmentation which finds the best match of the binary model of object in subsequent frames using the Hausdorff distance. In [5], the authors proposed a region merging approach to identify a semantic object. Their method performs an over-segmentation in the current frame and then iteratively merges the regions based on spatiotemporal similarity in the following frames. Since a semantic video object contains multiple color, texture, and motion, [6] proposed multi-feature based method by combining color, texture, and motion to improve the extraction performance.

Although there are many automatic methods for semantic video object extraction, their results are not satisfactory. This is due to the fact that the semantic video object is too complicated to segment with local cues, such as color, texture, and motion. Therefore, a number of semi-automatic methods are proposed to incorporate user's interaction into the segmentation process for semantic video object extraction. In semi-automatic method, typically a user defines the semantic objects in the first frame, and then the semantic objects are tracked automatically in the following frames. In [7], the authors proposed semi-automatic extraction method in which intra-frame segmentation is performed by user's interaction, and then the inter-frame segmentation is performed by tracking object boundary. In [8], a video segmentation method based on multiple features was proposed. Their method incorporates user's interaction to define semantic object and represents it as multiple features. A constrained fuzzy C-mean algorithm is used to compute multiple feature vectors for each region. In [9], the authors segment semantic objects with user's interaction and a multivariate watershed algorithm is performed in the first frame. Then, the segmented region is tracked in the following frames. Since initial segmentation plays a important role in the system performance, [10] introduce an active contour model, VSnakes, to locate accurate object boundary.

In this paper, we propose a semi-automatic method for semantic video object extraction. The proposed method incorporates user's interaction and integrates multiple cues in the energy function. Given correctly segmented image acquired by user's interaction in the first frame, the semantic objects are extracted by minimization of the energy function in the following frames. This paper is organized as follows. Section 2 describes overview of our method. Section 3 introduces semi-automatic

semantic object extraction method and the corresponding experiment results are presented in Section 4. At the last, the paper concluded in section 5.

2. SYSTEM OVERVIEW

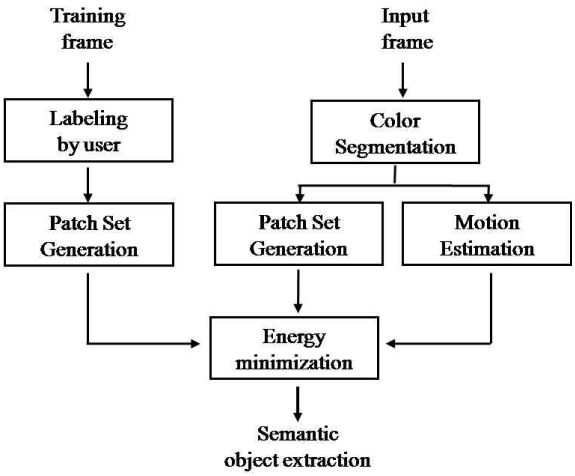


Fig 1. Block diagram of overall system

In the first frame the user labels semantic objects accurately using efficient segmentation tools such as lazy snapping [11] or spectral matting [12], and then patch sampling is performed over the labeled objects to represent semantic object as a set of patches. To automatically track the semantic objects in the following frames, we over-segment the input frames into small fragments using mean shift algorithm, and then patch sampling is also performed at each fragment. We construct an energy function which considers the possible labeling of each fragment using patch similarity, motion information, color, and local continuity. Finally, the semantic object extraction is formulated as an energy minimization problem and its solution is obtained by ICM [13]. The block diagram of the proposed method is shown in Fig. 1.

3. PROPOSED METHOD

3.1. Patch set generation

Since patch-based methods were shown to be extremely successful in object recognition [14], we represent the labeled region, semantic object, as a set of patches. Let I_{train} denotes the first frame and O_{train} is the labeling of I_{train} by k different labels. Here O_{train} is acquired by user's interaction with spectral matting. Given I_{train} and O_{train}, we randomly sample patches at each labeled region of the training frame to measure the similarity between the patch of input frame and the patch of the training frame. We use a fixed patch size of 7×7. After sampling, we have k-patch sets $\{S_i\}_{i=1}^{k}$, one for each label. Each set consists of a variable number of patches with respect to the size of labeled region.

Fig 2. Color segmentation result

3.2. Color segmentation

The computation of energy function which considers the possible labeling of each pixel is very intensive and sensitive to noise. In order to reduce system computation load and increase robustness, we define energy function at the fragment level, not pixel level. We divide input frame, I_{input}, into small homogeneous regions i.e. fragments according to some low level features. For the fragment segmentation, we use the mean shift algorithm [15]. Mean shift algorithm estimates density gradient of feature space and do not require multiple parameters which is important characteristics for robust color segmentation. As shown in Fig. 2, we over-segment the object to penalize two different objects share common fragment.

3.3. Motion estimation

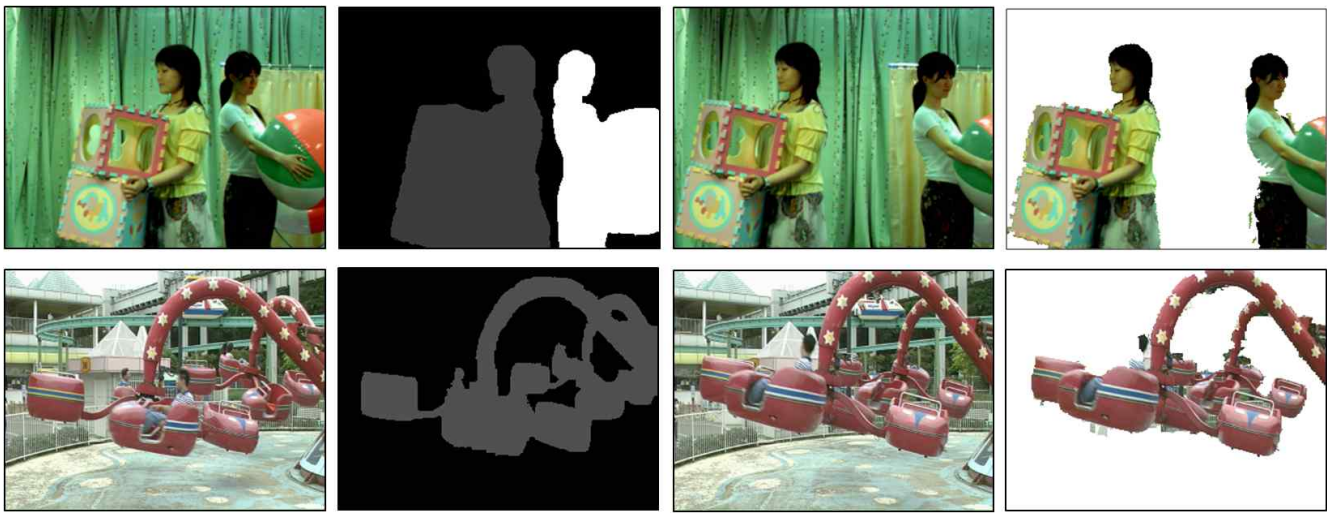
Generally, the motion of object in successive frames changes gradually over time. To take advantage of motion information, a bidirectional Kanade-Lucas- Tomasi (KLT) algorithm is performed at fragment level. We assume that the motion of each fragment is smooth. With this assumption, tracking a few features within each fragment enables to acquire accurate motion information. The feature points are selected from boundary of fragment and features are tracked by KLT feature tracker [16]. Bidirectional tracking is performed to increase accuracy of the feature tracking. When feature tracking fails or features are not extracted from fragment, interpolation is performed with neighbor fragment's color and distance information.

3.4. Energy minimization

Given a set of labels $\{L_i\}_{i=1}^{k}$ and a set of fragments F, the labeling problem is to assign a label $l_p = L$ to each of the fragment $p \in F$. An energy function is formulated as follows so that its minimum solution is an optimal label.

$$E(L) = \sum_{p \in F} E_{data}(l_p) + \lambda \sum_{q \in N} E_{smooth}(l_p, l_q) \qquad (1)$$

On the right hand side of eq. (1), the first term i.e. data term



(a) First frame (b) Training image (c) Input frame (d) Extraction result

Fig 3. Semantic object extraction results of "Akko&Kayo" and "Amusement park"

estimates the configuration l_p based on observation data at fragment p. We define $E_{data}(l_p)$ to be the cost of assigning label l to a fragment p as follows:

$$E_{data}(l_p) = \left[1 - \exp\left(-\frac{Dist[S(p),\ S_l]}{2\sigma^2}\right)\right] \times M\left(l_p, \Delta X^{p,t}, t, t-1\right) \quad (2)$$

, where $S(p)$ is a sampled patch set from fragment p, and S_l is one of the patch sets $\{S_i\}_{i=1}^{k}$ which label is l. $Dist[S(p),\ S_l]$ compares two patch sets by computing median value of the Euclidean distances between two patch sets, because the median value is robust to outlier. Parameter σ is related to the variation of patch similarity.

To enforce the continuity of the object motion along the time axis, $M\left(l_p, \Delta X^{p,t}, t, t-1\right)$ is constructed by following manner. As described in section 3.3, we estimate motion information by KLT method at fragment level. Let $\left\{X_i^{p,t}, \Delta X_i^{p,t}\right\}_{i=1}^{m}$ be extracted m-KLT features, and corresponding motion vectors of the p fragment in the t frame, respectively. $L\left(X_i^{p,t}, \Delta X_i^{p,t}\right)$ is the label of i^{th} KLT feature after translated by $\Delta X_i^{p,t}$ in t-1 frame. Finally, $M\left(l_p, \Delta X^{p,t}, t, t-1\right)$ is formulated by following equation.

$$M\left(l_p, \Delta X^{p,t}, t, t-1\right) = \frac{\sum_{i=1}^{m} T\left(L\left(X_i^{p,t}, \Delta X_i^{p,t}\right),\ l\right)}{m} \quad (3)$$

, where $T(\ )$ is an indicator function and written as:

$$T(a,b) = \begin{cases} 0, & \text{if} \quad a = b \\ 1, & \text{otherwise} \end{cases}$$

Data term measures not only the dissimilarity between the patches by l in the training image and the patches of fragment p but also the continuity of object label along the time axis.

The second term is the smoothness term which describes the label similarity of the adjacent fragments around fragment p. It is formulated as:

$$\sum_{q \in N} E_{smooth}\left(l_p, l_q\right) = \sum_{q \in N} T\left(l_p,\ l_q\right) \times adjac(p,q) \quad (4)$$

, where $T\left(l_p,\ l_q\right)$ is same manner as eq. (3), and q is the neighborhood fragment. Generally, neighborhood system often coincides with the set of regular grid of pixels such as 4-neighbourhood or 8-neighbourhood system. However, N is the set of variable number of fragments which share common boundary. The $adjac(p,\ q)$ represents the portion of shared common boundary and it can be computed as follows:

$$adjac(p,q) = \frac{\text{number of shared boundary pixels between } p \text{ and } q}{\text{number of boundary pixels of } p}$$

The parameter λ is used to control the relative importance of the data term versus the smoothness term.

Finally, the segmentation is formulated as a global minimum of the energy function and the minimization of this energy function is performed using iterative condition mode (ICM) for the sake of simplicity. The previously labeled result is used for initial solution of ICM.

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed method, we used two image sequences “Akko&Kayo” and “Amusement park”. The sequence, “Akko&Kayo″, contains smooth and almost rigid motion with two persons over a stationary background. Unlike “Akko&Kayo”, the “Amusement park” sequence is a combination of rapid, non-rigid motion over a stationary background. In all our experiment, λ is 0.1 for “Akko&Kayo” and 0.2 for “Amusement park”. We use fixed σ value of 25.

Fig. 3(a) shows the first frame and Fig. 3(b) is the training images labeled by user′s interaction. The black label denotes the background and other color labels denote semantic objects. Fig. 3(c) shows the input frame. The semantic object extraction results of frame of “Akko&Kayo” and “Amusement park” are illustrated in Fig. 3(d). As we can see in Fig. 3, the proposed method can obtain a good object extraction and handle the moving rigid and non-rigid object. However, errors may occur where the images are roughly segmented or untrained background occurred.

5. CONCLUSION

In this paper, we propose a semi-automatic method for semantic video object extraction where semantic video object extraction is formulated as an energy minimization problem. The proposed energy function takes into account local continuity, patch similarity, color, and motion information. Finally, minimization of the energy function leads to a robust semantic object extraction result. Experiment results demonstrate that our proposed method can effectively extract video object from various image sequences and handle both, rigid and non-rigid, moving objects. However, there are many problems with semantic video object extraction. For example, illumination variation, background change, error accumulation in successive frames, and occlusion and so on. These deficiencies are topic of further researches.

6. REFERENCES

[1] Y. Rui, T. Huang, and S. Chang, ″Digital Image/Video library and MPEG-7: Standardization and Research Issues,″ ICASSP′98, Seattle, May 1998.

[2] Chenglei Wu, Guihua Er, Xudong Xie, Tao Li, Xun Cao, and Qionghai Dai, ″A Novel Method for Semi-automatic 2D to 3D Video Conversion,″ 3DTV Conference: The True Vision Capture, Transmission and Display of 3D Video, 2008.

[3] ISO/IEC JTC1/SC29/WG11, Overview of the MPEG-4 Standard, MPEG98/N2323, Dublin, July 1998.

[4] T. Meier and K.N. Ngan, ″Automatic segmentation of movings for video object plane generation,″ IEEE Transactions on Circuits and Systems for Video Technology 8, pp. 525-538, 1998.

[5] F. Moscheni, S. Bhattacharjee, and M. Kunt, ″Spatio-temporal segmentation based on region merging,″ IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 9, pp. 897-915, 1998.

[6] J. Pan, S. Li, and Y. Zhang, ″Automatic extraction of moving objects using multiple features and multiple frames,″ IEEE Int. Symp. Circuits and Systems (ISCAS) 2000, Geneva, May 2000.

[7] Munchurl Kim, Jun Geun Jeon, Jinsuk Kwak, Myoung Ho Lee and Chieteuk Ahn, ″Moving Object Segmentation in Video Sequences By User Interaction and Automatic Object Tracking,″ Image and Vision Computing Journal, vol. 19, no. 5, pp. 245- 260, April 2001.

[8] R. Castagno, T. Ebrahimi, and M. Kunt, ″Video segmentation based on multiple features for interactive multimedia applications,″ IEEE Transactions on Circuits and Systems for Video Technology 8 (5) , pp. 562-571, 1998.

[9] C. Gu and M. C. Lee, ″Semantic video object segmentation and tracking using mathematical morphology and perspective motion model,″ International Conf. on Image Processing, pp. 514-517, 1977.

[10] S. Sun, D. R. Haynor, and Y. Kim, ″Semiautomatic video object segmentation using VSnakes,″ IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 1, pp. 75-82, 2003.

[11] Y. Li, J. Sun, C.K. Tang, and H.Y. Shum, ″Lazy snapping,″ Proceedings of ACM SIGGRAPH 2004, pp. 303-308, 2004.

[12] A. Levin, A. Rav-Acha, and D. Lischinski, ″Spectral Matting,″ IEEE Trans. Pattern Analysis and Machine Intelligence, Oct 2008.

[13] J. Besag, ″On the statistical analysis of dirty pictures,″ J. Royal Statist. Soc. B, vol. 48, no. 3, pp. 259-302, 1986.

[14] S. Ullman and E. Sali, ″Object Classification Using a Fragment-Based Representatio,″ BMVC 2000, Proc. Lecture Notes in CS 1811 Springer, pp. 73-87, 2000.

[15] Dorin Comaniciu and Peter Meer, ″Mean Shift: A Robust Approach Toward Feature Space Analysis,″ IEEE Transactions on Pattern Analysis and Machine Intelligence, v.24, no.5, pp. 603-619, May 2002.

[16] C. Tomasi and T. Kanade, ″Detection and Tracking of Point Features,″ Technical Report CMU-CS-pp. 91-132, 1991.