Object Segmentation from Long Video Sequences

Bing Luo, Hongliang Li, Tiecheng Song, Chao Huang

University of Electronic Science and Technology of China, Chengdu, China
mathild1987@163.com, hlli@uestc.edu.cn, tggwin@gmail.com, huangues@gmail.com

ABSTRACT

Most existing video segmentation methods are focused on extracting the primary objects in test video sequences. They assumed that only one object appeared through the whole video sequences, which is impractical in many applications. In this paper, we focus on the object segmentation from the long video sequences which consist of many different scenes, shot cuts and various motion patterns, etc. In order to solve this problem, we propose a framework to segment the objects in relative video shots, while discarding the irrelative video shots. A graph is constructed to model the video object detection and final segmentation is obtained by getting the superpixels in the detection boxes. We also introduce a new long video segmentation dataset which corresponds to the pixel-wise ground truth. The experiments demonstrate that our proposed method can deal with the object segmentation in long video sequence.

Categories and Subject Descriptors

I.4.6 [**Image Processing and Computer Vision**]: Segmentation—*Video segmentation*

General Terms

Algorithms

Keywords

Long video sequence, Segmentation, Graph

1. INTRODUCTION

Video segmentation extracts the object through the whole videos with relative frames, which is widely used in action recognition, video analysis. Compared with the object segmentation from images, video segmentation utilizes not only the information in single frame but also the relationship between adjacent frames. Many existing methods have been proposed to obtain the foreground in video by extending the static and dynamic information, such as shortest path video segmentation [18, 19], key-segments video segmentation [8], maximum weight clique based video segmentation [12], and incremental learning based video segmentation [9].

Although these methods have successfully segmented the objects in the popular test sequences, such as SegTrack Dataset [15] and GaTech Segmentation Dataset [6], they assumed that the primary objects appeared in each frames and no shot cuts were present in these sequences. While this assumption is strong and impractical in many real applications, such as movies, long videos and edited videos. Many complex situations may frequently appear in these videos, such as suddenly scene change, multiple objects appearing across in different shot cuts, uncertain motion pattern, etc. When all the situations arise in a long video sequence, how to extract the object in the relevant video shots is challenge.

In order to declare our new problem for long video sequences object segmentation, we introduce a new benchmark video segmentation dataset, LongVideoSeg dataset. The LongVideoSeg dataset contains one video which consists of 10000 frames from various shot cuts with multiple objects and different scenes. All the video frames are annotated with pixel-wise groundtruth for specific classes by human. In order to get the specific-class objects prior information, few bounding boxes corresponding to class labels are provided with some noise data.

In this paper, we proposes a new framework to solve the long video sequence segmentation. We cluster the given bounding boxes and utilize the correlation among them to discard noisy boxes. The long video sequence is divided into shot cuts with consistent scene and the relevant shot cuts are found with the help of clustering object priority. We model the video object detection process as an energy minimization problem which is optimized via dynamic programming. Finally, some subjective and objective experiments validate our method on our new dataset.

2. RELATED WORK

Most existing video segmentation methods pay much attention to extracting primary object in single video sequence. Several methods proposed to segment the primary object by discovering the most objectness segments with appearance and motion information. Lee and Kim [8] firstly utilized spectrum clustering to discover the segments with the high static and dynamic scores and generate foreground prior for spatial-temporal graph cuts. Ma [12] utilized the spatial and temporal mutex constrain to model objects discovery as constrained maximum weight cliques. Zhang [19] constructed

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806313.

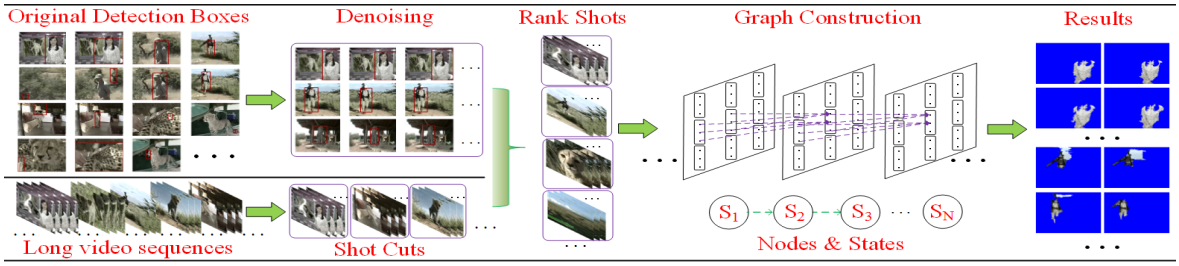

Figure 1: The flowchart of our proposed method.

directed acyclic graph to find the primary object regions based on appearance and motion similarity. These methods focus on single object appeared in each frames which are unsuitable for real unconstrained video sequences. Li [9] modeled the many figure-ground segments tracking as a incremental learning problem. Meanwhile, Banica [1] developed the multiple segments hypotheses as salient segment chain composition. These methods can deal with multiple objects in a successive video sequence.

Recently, in order to segment object in different video sequence, video cosegmentation is an alternative method to extract the common objects from two or multiple videos. Several methods [2, 14, 7] have been proposed to extract the common object from two or more videos, which assumed that the target object appeared in all videos. Liu [11] and Fu [5] proposed methods to solve the multiple foreground video cosegmentation. Their methods are constrained in the assumption that the common objects must appear in two or multiple videos. Zhang [20] proposed a maximum weight cliques based video cosegmentation to discard the assumption on the target objects appearance in all videos. Wang [17] presented an extremely weak supervision video cosegmentation to discover video object in relevant frames and cosegment in different videos.

3. PROPOSED METHOD

Given a long video sequence with a few specific-class bounding boxes, we present a new framework to solve the segmentation in a long video sequence. Fig. 1 overviews our proposed method. The framework consists of three steps: (1) We simultaneously cluster the given bounding boxes into subclasses and discard the noise boxes which are incorrect detection. (2) Given a long video sequence, we divide it into shot cuts and rank them by video-level class scores with the help of subclass bounding boxes to discover the relevant frames which the target object appears in them. (3) A graph is constructed to model the object detection process and the final segmentation is generated via these detection boxes.

3.1 Data Clustering and Denoising

Since the given specific-class bounding boxes are detected by DPM [3] and preserved upon a certain threshold, some noise data will be inevitably introduced into the class data. In this step, we avoid the noise data to contaminate the object prioriry. Firstly, we cluster the given bounding boxes via their height/width into K clusters. The noisy bounding boxes is independent of class information, so they are clustered into different subclasses which help us to prevent them into object trajectories. Secondly, for each subclass we calculate the correlative matrix in term of their corresponding

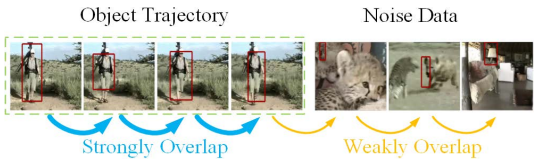

Figure 2: The object tracklet.

overlap rate. We observe that the more overlap the adjacent bounding boxes have, the more likely the two boxes are a same object. So an object trajectory corresponds to the higher overlap rate and successive bounding boxes, the noisy data have lower overlap rate due to their independence, as show in Fig. 2. Hence, we generate object tracklets as the subclasses positive samples and discard the noise data.

3.2 Generate and Identify Relevant Shot Cuts

Given a long video which is edited with many different shot cuts, we need to generate each single shot cut, i.e. identify the start and end points for a successive frame sequences. In general, a single shot cut is motion smoothness and object consistency, which is comfortable to handle with. We firstly identify the start and end point for the different shot cuts. Given a set of video frames $F = \{f_i, 1 \leq i \leq N\}$, where N is the frame number for the long video sequence. We denote $S = \{s_i, 1 \leq i \leq N-1\}$ as the inter-frame difference, i.e., $s_i = \frac{1}{\overline{N}} \sum_{x,y \in I} (f_{i+1} - f_i)$, where $\overline{N}$ is the pixel number in a frame, and x, y are the position in image I. We smooth s and add a threshold on it, and then the adjacent frames switched to different scene will yield a high response. The response points beyond the threshold are treated as the start and end points for the shot cuts.

After generating the shot cuts, we discovery the relevant shot cuts by matching the weighted histogram of shot cuts. We rank the matching distance and select the top N nearest shot cuts as the relevant shot cuts.

3.3 Video Object Detection

We treat the object detection as an optimization of model. A directed solution is to construct a graph such as Fig. 3. In this graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each node $v_i \in \mathcal{V}$ denotes a bounding box in corresponding frame and each edge $e_{ij} \in \mathcal{E}$ denotes the similarity between two bounding boxes in adjacent frames. In this model, the task is to obtain the state for each node, and the state space for each node $L = \{1 \leq l_i \leq \overline{N}, 1 \leq i \leq N\}$, is the bounding box position in current frame. Inspired by [4], our cost function is defined as:

$$E = \sum_{i \in \mathcal{V}} U_i(l_i) + \sum_{(i,j) \in \mathcal{E}} V_{i,j}(l_i, l_j) \qquad (1)$$

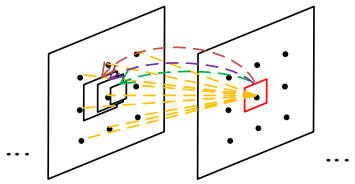

Figure 3: Connection of nodes in adjacent frames.

The energy function consists of unary terms $U_i(l_i)$ and pair-wise term $V_{i,j}(l_i, l_j)$, which represent the objectness scores of detected bounding box and smoothness between two bounding boxes in adjacent frames [10], respectively. $U_i(l_i)$ is calculated by $U_i(l_i) = \frac{B(l_i)}{Area(l_i)}$, where $Area(l_i)$ is the area of the l_i-th bounding boxes, and $B(l_i)$ is the score of the sliding windows detection. This unary term can preserve the tightness of the detection boxes. In order to obtain the objectness score, we train an appearance model for the target object via GMM in the color space, $O = \{\pi_o(k), \mu_o(k), \nu_o(k), o \in \{0,1\}, k \in \{1, ..., K\}\}$. Then frames in relevant shot cuts can be generated a pixel-wise foreground object likelihood. We use the integral image [16] to speed up the objectness scores calculation for unary terms. Hence, the unary term can be calculated in constant time. Then the pairwise terms are calculated the similarity of the normalized color histograms for adjacent nodes bounding boxes.

$$V_{i,j}(l_i, l_j) = 1 - exp\{-\chi^2(l_i, l_j)\} \quad (2)$$

Where $\chi^2(l_i, l_j) = \frac{1}{2}\sum_{b=1}^{N_d} \frac{(h_{l_i}(b) - h_{l_j}(b))^2}{h_{l_i}(b) + h_{l_j}(b)}$, h_{l_i} is the color histogram of bounding box l_i in frame i, and N_d is the number of histogram bins. Minimizing the cost function Eq. 1 will obtain the optimal state for each node, i.e., the optimal detection box for each frame.

In order to optimize Eq. 1, an optimization method via dynamic programming is summarized in Alg. 1. The algorithm finally outputs the optimal state for each node, i.e., the optimal center position Lab and scale S_{star} for detected bounding box. After obtaining the detected bounding boxes, we extract the superpixels in the detected bounding boxes as the final object segmentation.

4. EXPERIMENTS

In order to verify our proposed method, a new benchmark database, LongVideoSeg dataset, is introduced in this paper. The video sequence contains 10000 video frames with *people* and *cheetah* categories, which are annotated by pixel-wise groundtruth manually. Given by a few bounding boxes for above two categories with some noise detections, the task is to obtain the pixel-wise labels for the each object class.

We define the evaluation metrics in three situations:

(1) For the object class k, the groundtruth annotation and segmentation masks all exist in these frames. The evaluation score is calculated by intersection-over-union.

(2) The evaluation score is defined as zero if and only if one of groundtruth annotation and segmentation masks exists in the frames.

(3) The test frames don't contain the object and the segmentation masks also don't exist. We don't consider the evaluation into our scores.

Our evaluation metric is finally defined as follows:

Algorithm 1 Algorithm for Video Object Detection

Input:
Video shot cut $F = \{f_i, 1 \le i \le N\}$;
Foreground likelihood map, $M = \{m_i, 1 \le i \le N\}$;
Scale for sliding windows, $S = \{s_k, 1 \le s \le K\}$;
Output:
Detected bounding boxes $\{Lab(i), S_{star}(i), 1 \le i \le N\}$;

for $i = 1 : N$ **do**
  for $s = 1 : K$ **do**
    calculate $U_i(l_i, s)$ by integral image, where $l_i \in \{1, ..., |L|\}$ is the position for each node.
  end for
end for
for $i = 2 : N$ **do**
  for $l_i = 1 : |L|$ **do**
    calculate $w_{l_i,s}(l_{i-1}, s)$ by Eq. 2, where $l_{i-1} \in \mathcal{N}(l_i)$
    $l^*, s^* \longleftarrow argmax_{l_{i-1} \in N(l_i), s \in \{1,...,K\}}(B(l_{i-1}, s) + w_{l_i,s}(l_{i-1}, s))$
    $B(l_i, s) = min_{l_{i-1} \in N(l_i), s \in \{1,...,K\}}(B(l_{i-1}, s) + w_{l_i,s}(l_{i-1}, s) + U_i(l_i, s))$
    record the best child node for current position $L(l_i) = l^*$
    record the best state of child node $S(l_i, s) = s^*$ for position l^*
  end for
end for
trace back from $i = N$ frame
$l_*, s_* \longleftarrow argmax_{l_i \in L, s \in 1,...,K} B(l_i, s)$
$Lab(N) = l_*, S_{star}(N) = s_*$
for $i = N - 1 : 1$ **do**
  $Lab(i) = L(Lab(i + 1))$
  $S_{star}(i) = S(Lab(i + 1), S_{star}(i + 1))$
end for

$$Score = \begin{cases} \frac{M_i \bigcap G_i}{M_i \bigcup G_i} & \text{situation 1} \\ 0 & \text{situation 2} \\ not\ consider & \text{situation 3} \end{cases} \quad (3)$$

To our knowledge, there are no existing methods to perform multiple objects segmentation in long video sequences. We compare our proposed method with two baseline methods and FastVideoSeg [13]. In order to validate the first step, i.e., data clustering and denoising, Baseline1 is designed to directly model the foreground prior by GMM while step 2 and step 3 remain unchanged. This design evaluates the efficiency of the foreground map correctness by the noise data. The more correctly the foreground maps are generated, the higher accuracy our method gives. The second method Baseline2 is designed to randomly select the relevant shot cuts for step 2 and not to change other steps, which can evaluate the correctness of our step 2. Finally, [13] is used to compare with our step 3, i.e., detection and segmentation algorithm only on our detected relevant shot cuts.

Some subjective results are shown in Fig. 4 for our method and Baseline1. In this figure, our method generates more accurate foreground prior and leads to higher detection precision by eliminating the noise data than Baseline1. Specifically, the bounding box can detect the people more completely by our method. Cheetah detection by our method can locate the main object in these frames but Baseline1 loses the object due to the noise data. We also compare our

Figure 4: Comparison between our method and Baseline1. From top to bottom: original frames, foreground map by Baseline1, detection by Baseline1, our foreground map, our detection.

Table 1: IoU of our method and compared methods.

Class	Baseline1	Baseline2	[13]	Ours
people	38.62	9.27	20.38	**40.35**
cheetah	10.60	4.03	18.80	**24.97**

method with [13]. The subjective segmentation results are shown in Fig. 5. This figure shows that [13] only segments the face of the people and feet of the cheetah. The reason is that the performance of [13] seriously depends on optical flow, which is unreliable in many cases. In people shot cut, the person only twists her head. In cheetah shot cut, the optical flow calculation is inaccurate for the object. The above two situations will result in an error result by [13]. The results of our method don't utilize optical flow information and models the robust object detection as an energy minimization problem, which leads to the better performance.

Tab. 1 shows the results of our method and the compared methods. It can be seen that our method obtains 40.35% in people and 24.97% in cheetah, which outperforms other compared methods. From above experiments, the subjective and objective performance comparison demonstrates that our proposed method is effective.

5. CONCLUSIONS

In this paper, we propose a new framework to solve the long video sequence segmentation. We cluster the given bounding boxes and utilize the correlation among them to discard noisy boxes. The long video sequence is divided into shot cuts with consistent scene and the relevant shot cuts are found with the help of clustering object priority. We model the video object detection process as an energy minimization problem which is optimized via dynamic programming. Finally, a new database LongVideoSeg is introduced to evaluate our proposed method and some subjective and objective experiments validate our method.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Basic Research Program of China (973 Program 2015CB351804), National Natural Science Foundation of China (No. 61271289), and by The program for Science and Technology Innovative Research Team for Young Scholars in Sichuan Province, China (No. 2014TD0006).

7. REFERENCES

[1] D. Banica, A. Agape, A. Ion, and C. Sminchisescu. Video object segmentation by salient segment chain composition. In *ICCVW*, 2013.

[2] D.-J. Chen, H.-T. Chen, and L.-W. Chang. Video object cosegmentation. In *ACM MM*, 2012.

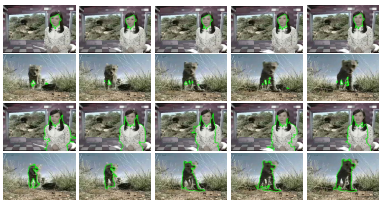

Figure 5: Segmentation results of FastVideoSeg [13] (top two rows) and our method (bottom two rows).

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2010.

[4] P. F. Felzenszwalb and R. Zabih. Dynamic programming and graph algorithms in computer vision. *IEEE TPAMI*, 2011.

[5] H. Fu, D. Xu, B. Zhang, and S. Lin. Object-based multiple foreground video co-segmentation. In *CVPR*, 2014.

[6] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010.

[7] J. Guo, Z. Li, L.-F. Cheong, and S. Zhou. Video co-segmentation for meaningful action extraction. In *ICCV*, 2013.

[8] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011.

[9] F. Li, T. Kim, A. Humayun, D. Tsai, and J. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013.

[10] H. Li, F. Meng, Q. Wu, and B. Luo. Unsupervised multiclass region cosegmentation via ensemble clustering and energy minimization. *IEEE TCSVT*, 2014.

[11] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu. Weakly supervised multiclass video segmentation. In *CVPR*, 2014.

[12] T. Ma and L. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012.

[13] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[14] J. C. Rubio, J. Serrat, and A. López. Video co-segmentation. In *ACCV*. 2013.

[15] D. Tsai, M. Flagg, and J. Rehg. Motion coherent tracking with multi-label mrf optimization. *BMVC*, 2010.

[16] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[17] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng. Video object discovery and co-segmentation with extremely weak supervision. In *ECCV*. 2014.

[18] B. Zhang, H. Zhao, and X. Cao. Video object segmentation with shortest path. In *ACM MM*, 2012.

[19] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013.

[20] D. Zhang, O. Javed, and M. Shah. Video object co-segmentation by regulated maximum weight cliques. In *ECCV*. 2014.