Neurocomputing 172 (2016) 235–243

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Unsupervised pixel-level video foreground object segmentation via shortest path algorithm

Xiaochun Cao [a], Feng Wang [b,*], Bao Zhang [c], Huazhu Fu [d], Chao Li [e]

[a] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[b] School of Computer Software, Tianjin University, Tianjin 300072, China
[c] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[d] School of Computer Engineering, Nanyang Technological University, Singapore
[e] Shenzhen Key Laboratory of Data Vitalization, Research Institute of Beihang University in Shenzhen, Shenzhen 518057, China

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form
31 July 2014
Accepted 20 December 2014
Available online 9 May 2015

Keywords:
Video object segmentation
Shortest path solution

ABSTRACT

Unsupervised video object segmentation is to automatically segment the foreground object in the video without any prior knowledge. In this paper, we propose an object-level method to extract the foreground object in the video. We firstly generate all the object-like regions as the segmentation candidates. Then based on the corresponding map between the successive frames, the video segmentation problem is converted to corresponding graph model, which selects the most corresponding object region from each frame. The shortest path algorithm is explored to get a global optimum solution for this graph. To obtain a better result, we also introduce a global foreground model to restrict the selected candidates. Finally, we utilize the selected candidates to obtain a more precise pixel-level foreground object segmentation. Compared with the state-of-the-art object-level methods, our method does not only guarantee the continuity of segmentation result, but also works well even under the cases of fast motion and occlusion.

1. Introduction

Social media contains a large scale videos. How to extract the main contents for these videos automatically is a key problem for media analytics and learning. A common media issue is the video foreground segmentation. Video object segmentation methods generally include two categories: supervised segmentation and unsupervised segmentation. Supervised segmentation needs user interaction to initialize the key objects. In contrast, the unsupervised methods release the user input to automatically extract the foreground. There is relatively less attention focusing on the unsupervised method compared with the supervised one because it is difficult to define the foreground in video automatically. But the study of unsupervised video object segmentation is significant and it can be applied to many fields such as video analysis and understanding [1], video summarization and indexing [2,3], video retrieval [4], and activity understanding [5,6]. For instance, object-based video segmentation separates the meaningful object from the background. Video analysis and understanding can have a better understanding of foreground and background, as well as the semantic relationship of them. Moreover, for video retrieval, object segmentation can find the videos with related foreground, and remove unrelated videos which have the similar background.

Recently, there are many methods provided to predict foreground model from videos. For instance, visual saliency has been used to form foreground object model in the video [7–10]. Gu et al. [11] and Wang et al. [12] propose video scene segmentation approaches with content coherence and contextual dissimilarity. Brox and Malik [13] propose a method based on foreground object trajectory. A series of trajectories are firstly extracted from the video, then processed by the spectral clustering method. Since plenty of the trajectories are sparse, the calculated foreground also consists of sparse pixels set. Ochs and Brox [14] extend the method [13] to acquire a more dense foreground region. Brendel and Todorovic [15] argue that video object segmentation by tracking regions has many fundamental advantages over the approaches based on tracking points or jointly clustering of all pixels from all video frames. However, Lee et al. [16] point out that these methods lack an explicit notion of what a foreground object should look like in video data and the low-level grouping of pixels usually results in over-segmentation.

Since the above methods are based on the low-level visual features, which lack an explicit notion of what a foreground object should look like in video data, methods in [17–19,32,33] explore object-based segmentation in static image and achieve significant progress. Those methods generate multiple object hypotheses and rank hypotheses according to their scores. The model is learned for

* Corresponding author.
E-mail addresses: caoxiaochun@iie.ac.cn (X. Cao), wangf.tju@gmail.com (F. Wang), zhangbao@tju.edu.cn (B. Zhang), hzfu@ntu.edu.sg (H. Fu), licc@buaa.edu.cn (C. Li).

http://dx.doi.org/10.1016/j.neucom.2014.12.105
0925-2312/

a generic foreground object using several image features such as color, texture, and boundary, which is then object category independent.

In this paper, we propose an object-level method to segment the foreground object from the unlabelled. We firstly generate a set of the object-like regions by using method [17]. Then we construct the corresponding graph model based on the corresponding map using selected candidates between the successive frames. After that, the shortest path algorithm is explored to get a global optimum solution. To obtain a better result, we employ an interaction method that introduces a global model to restrict the selected candidates with the global object model. Finally, the selected candidates are utilized to obtain a more precise pixel-level foreground segmentation.

1.1. Related works

Lee et al. [16] introduce an unsupervised video object segmentation method based on object level. They firstly get a series of candidate regions [17]. Combined with the motion attribute, each proposal is scored. Then based on the color histogram of each proposal, different object clusters are acquired using spectral clustering. The cluster with the highest mean score is regarded as the foreground, which is used for video object segmentation. It is the common sense that the object going through the entire video always has more significance than the short-term appearing object, i.e. more likely to be the foreground object. However, there are three major problems when they utilize spectral clustering to form a foreground model. (1) The clustering is merely based on color histogram to measure the correspondence distance (correlation) between two proposals. Since no other features are taken into consideration to balance the correlation between two proposals in the successive frames, this might result in wrong classification when background color is similar to the foreground color. (2) They cannot guarantee that the proposals in their clustering result cover all the frames in the video, i.e. lack of continuity. (3) As they said, the cluster with highest score is selected to be the foreground object model. However, if the wrong foreground object is chosen (e.g. there is an interferent object with faster motion in the background, the interferent object could be selected to be the foreground object model), the final segmentation result would be wrong even after the post-processing steps. The comparison with Lee's method is shown in Fig. 1. We simulate a common scene happening all the time. It contains a walking people and a running one. In the first several frames, both people exist in the video. Since the running one is quicker to run out of the camera sight, the last few frames contain the walking person only. In contrast, our method works better than Lee's under such scene.

Ma and Latecki [20] attempt to address this video object segmentation problem by utilizing relationships between object proposals in adjacent frames. The object region candidates are selected simultaneously to construct a weighted region graph. This problem is modeled as finding a constrained Maximum Weight Cliques problem. However, this method also cannot guarantee that the proposals cover all the frames in the whole video. Moreover, the major problem is that the approach to solve maximum weight cliques is NP-hard. Only an approximate optimization solution is used to obtain the result, which may be not the global optimal solution. What is more, [16,20] have an additional limitations compared to the proposed method using object based segmentation approaches. The object proposal selection of a particular frame does not depend directly on adjacent frames in both approaches.

Zhang et al. [21] present a new approach to improve the methods [16,20], which uses a novel and efficient layered Directed Acyclic Graph (DAG) based approach to segment the primary object in videos, and the problem converts to find the highest weighted path in the DAG. The problem could be solved by dynamic programming in linear time. This approach also uses innovative mechanisms to compute the 'objectness' of a region and to compute similarity between object proposals across frames. However, the solving method of dynamic programming maybe causes curse of dimension, if the number of proposals and frames is large enough.

1.2. Our framework and contributions

In order to overcome the aforementioned drawbacks, we propose a new approach to segment the foreground object. Our method is to select object-like regions according to both static and dynamic cues, and then segment the main object of the unannotated video using

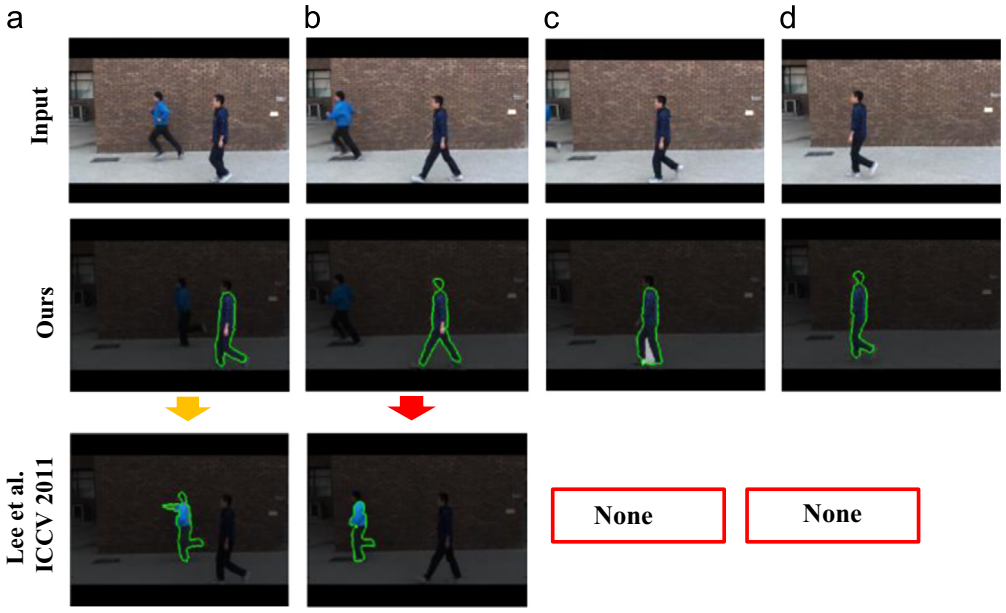

Fig. 1. Comparison with method [16]. The first row is input frames. Our result is shown in the second row, compared with the result of [16] shown in the last row. There is no corresponding segmentation result with respect to frames 9 and 13, marked by the red boxes. Note that the proposed method was able to find objects in all frames. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.) (a) Frame 1 (b) Frame 5 (c) Frame 9 (d) Frame 13.