International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.8 (2015), pp.231-238
http://dx.doi.org/10.14257/ijsip.2015.8.8.25

Dynamic Scene Segmentation through Object Hypotheses Ranking

Yinhui Zhang, Zifen He* and Xing Wu

Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming, China, 650500
**zyhhzf1998@163.com*

Abstract

A novel framework for highly dynamic scene segmentation through foreground hypothesis is developed here. This framework enables robust foreground segmentation by ranking object hypothesis over spatial space to achieve consistent object candidates and binary segmentation of a video sequence. Inside object candidates derived from spatial features in each frame are first estimated. This is followed by ranking the object candidates over a specific hypothesis space so as to yield consistent and dense object proposals. An efficient higher-order graph-cut method is adapted to optimize a Markov Random Field (MRF) model, which is instantiated by the estimated foreground hypothesis with highest score. We demonstrate the performance of our approach through experimental evaluation on a typical dynamic scene benchmark from Freiburg-Berkeley Motion Segmentation Dataset. Compared with a state-of-the-art algorithm, our method achieves improved and robust segmentation performance when dealing with highly dynamic image sequences. The segmentation accuracy of the proposed method improved by 10.19% and 92.66% pixels are correctly classified.

Keywords: *Dynamic scene segmentation, Object hypothesis, Markov random field, Object ranking*

1. Introduction

Segmenting salient foreground regions from video sequences is the basis for many visual tasks. Imagine such a scenario that an autonomous robot is wondering in an unknown environment, the location information of foreground targets which can be inference in an unsupervised way will certainly facilitate to a large extent of subsequent recognition, tracking as well as visual servo manipulation of the targets. However, the unconstrained natural scene contains many dynamics as camera motion, dynamic backgrounds as well as object occlusions that cause the video segmentation task very challenging. Despite of much effort in this direction, designing an algorithm that is robust under a wide variety of dynamic scenes encountered in complex natural environments remains an open problem.

The bottle neck technique in dynamic scene segmentation lies in foreground object hypothesis stage. At present, there are mainly two category methods to deal with foreground hypothesis. One approach takes advantage of such object appearance cues as color and texture. Typically, this approach consists of bottom-up and top-down segmentation paradigms. Bottom-up [2] paradigm exploits appearance similarity of neighborhood pixels of images, which are then merged into hyper pixels or image patches. Actually, bottom-up segmentation based on appearance cues always leads to oversegmentation of dynamic scenes and cannot segment foreground objects from background scenes.

Global features such as object shape are combined in top-down segmentation paradigm. However, the segmentation model has to be trained using these features in the

ISSN: 2005-4254 IJSIP

first place, which hampered the autonomous segmentation process. A new trend is combining bottom-up segmentation with top-down detection [3] by using a deformable part-based object detection framework. An alternative approach is to use the motion cues produced by state-of-the-art variational dense optical flow [4] as additional features for video segmentation.

In this paper, we take advantage of standard spatial-temporal video segmentation approaches by computing a hypothesis of foreground seeds by inside region mapping and produce a binary segmentation by graph-cut using the mapping. Our goal is to make use of these foreground regions to obtain consistent foreground proposals over a specific feature space so as to yield more accurate segmentation in highly dynamic scene. Unlike the aforementioned traditional approaches, our method is robust to dynamic segmentation by ranking consistent foreground candidates over temporal space, which is demonstrated using extensive experimental results.

The work in this paper has been divided into six sections. Inside object candidates that are generated by making effective use of multiscale boundary information in each frame are estimated in Section 2. This is followed by ranking the object candidates over a specific hypothesis space so as to yield consistent and dense object proposals in Section 3. In Section 4, an efficient higher-order graph-cut method is adapted to optimize a Markov Random Field (MRF) model, which is instantiated by the proposed foreground hypothesis with highest ranking score. Section 5 demonstrates the performance of our approach through experimental evaluation on a typical dynamic scene benchmark from Freiburg-Berkeley Motion Segmentation Dataset and then draw a conclusion in Section 6.

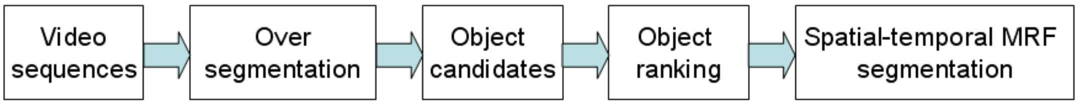

Figure 1. Block Diagram of Our Video Segmentation Algorithm

2. Object Candidates Generation

Due to the highly dynamic and complexity nature of natural scenes, a set of object candidates are generated in the first place in order to form a hypotheses pool. Instead of deformable part model based object detection approaches, which produce object candidates that are represented in terms of image windows, we choose to perform segment based object hypothesis through low-level appearance features computed on superpixels. To this end, we first compute superpixels of original images through SLIC [5] algorithm. In this paper, the size of each superpixel region is fixed at 10 pixels and the regularizer is selected as 0.2 during superpixel estimation. Then the superpixels are used as foreground and background seeds to generate a pool of object hypotheses. This leads to a large and diverse pool of segments at higher-order level. Given a set of superpixel seeds, the popular constrained parametric min-cut (CPMC) [6] method is used to hypothesize a series of foreground proposals at a variety of seeds and unary potentials of energy function. The main difference between ours method with CPMC is that, instead of hypothesize seeds on regular grids at pixel level, we perform seeds selection at high-level patch level, which is shown to be more robust than the standard method.

Moreover, a specific energy function is formulated instead a set of parametric functions. In precise, we compute a binary segmentation by minimizing the energy:

$$E(X)=\sum_{u\in I}\text{[illegible]}+\sum_{(u,v)\in V}\text{[illegible]} \qquad （1）$$

where u and v are indices of superpixels. U and V represents unary and pairwise potentials of the energy function E. For each superpixel u, we regard it lies in forground region according to the normalized probability distribution:

$$U(x_u) = -\log(\exp(-\min \| I_v - I_u \|)) \tag{2}$$

where I_u denotes the mean color in RGB color space of the superpixel u. This assignment is reasonable as neighboring superpixels with similar colors are tend to by the same segments. In this sense, similar colored superpixels are encoded with higher unary potentials, which will consume more energy if segment them into different regions.

Pairwise potentials are used to encode smoothness between superpixels. Since the global probability boundary (gPb) algorithm [7] has relative good performance in detecting salient contours in whole images, thus the standard CPMC algorithm employs gPb features to capture pairwise potentials. However, the classic global probability boundary extraction method is very time and memory consuming which hamper the foreground hypothesis process. To this end, a newly published Multiscale Combinatorial Grouping (MCG) [8] boundary detection approach is used in this paper, in which the standard spectral clustering algorithm is speed up by making effective use of multiscale information. For these reasons, in this paper, for each neighboring superpixels u and v, the pariwise term is formulated as

$$V(x_u, x_v) = -\log(\exp(-\| C_u - C_v \|)) \tag{3}$$

Note that the range of pairwise potentials are normalized in the zone of $[0 \quad 1]$. Intuitively, neighboring superpixels with similar boundary mappings are assigned with higher potentials.

3. Object Hypotheses Ranking

The objective of ranking is to assign hypotheses that exhibit object-like regularities with higher ranking score. As in CPMC method, several kinds of features are computed to capture *objectness* of object proposals that hypothesized in Section 2. The feature consists of graph partition properties, region properties such as area, perimeter, bounding box location, major and minor axis lengths of the ellipse as well as Gestalt properties such as curvilinear continuity and convexity. Gestalt properties are computed in terms of χ^2 distance, which is based on color histogram measures between two half discs of a circular image patch.

$$\chi^2(h,g) = \sum_i \frac{(h_i - g_i)^2}{h_i + g_i} \tag{4}$$

where h and g represents color histograms of two half discs. The radius of each disc is fixed at 6 pixels throughout the paper.

In fact, straightforward ranking of object proposals in terms of the objectness features would result that similar segments tend to be ranked in adjacent positions. To alleviate this problem, maximum marginal relevance criteria is employed to diversify the proposals in the ranking stack. This criteria consists of a score term and a diversity term. If high precision is desired, then a higher weight should be given to the predicted score term, whereas if recall is more important, then a higher weight should be given to diversity term [6]. In this paper, we reasonably emphasize the diversity term and associate this term with a weight of 0.6. In addition, traditional CPMC algorithm takes advantage of a regressor to learn the ranking of object hypotheses through ground truth segmentations of PASCAL VOC training. In this paper, however, we aim to perform unsupervised video segmentation. Thus the ground truth of the video sequence is not provided in the ranking stage. To overcome this difficulty, we employ the trained random forest regressor model through PASCAL VOC 2010 training data sets. In our method, the random forest regressor model consists of 50 trees.

4. Spatial-temporal MRF Segmentation

The goal of this section is to exert spatial and temporal constraints on the image sequences so as to yield a binary segmentation of videos. To this end, the popular spatial-temporal Markov random field (MRF) is used to model the correlation in spatial domain an along temporal axis. In particular, the MRF model consists of two unary terms and two pairwise terms. The unary terms exert appearance and position constraints on foreground object hypotheses. Following the method in [1], we formulate MRF model as

$$\begin{aligned} E(X) &= \sum_{t,u} A_u^t(x_u^t) + \alpha_1 \sum_{t,u} L_u^t(x_u^t) \\ &+ \alpha_2 \sum_{u \overset{s}{\longleftrightarrow} v} V_{uv}^t(x_u^t, x_v^t) + \alpha_3 \sum_{u \overset{t}{\longleftrightarrow} v} V_{uv}^t(x_u^t, x_v^{t}) \end{aligned} \tag{5}$$

where A denotes appearance potential that encourage superpixels with similar color values in RGB color space to cluster in the same category.

The L exert location constraints on foreground hypothesis to encode that foreground objects appear at near positions in consecutive frames. The symbol $u \overset{s}{\longleftrightarrow} v$ denotes the node u and v are connected in spatial domain. Similarly, $u \overset{t}{\longleftrightarrow} v$ denotes the node u and v are connected in time domain. The computation of spatial connectivity is straightforward as long as two superpixels share a same pixel at their borders. Whereas temporal connections established by using optical flow vectors at each node, which is derived by a variational dense optical flow method such as in [9]. The potential of temporal pairwise term is instantiated by superpixels overlap between two neiboring superpixels that are connected by a optical flow vector. The parameters in the energy model α_1, α_2 and α_3 are fixed as 1.2, 1.8 and 2.0 by experimental cross-validation method.

5. Experiments

To demonstrate the effect of the proposed method, we use the Freiburg-Berkeley Motion Segmentation Dataset [10], which contains of 59 video sequences. Since typical challenges appear in all sets, thus the *tennis* data set is employed to test the algorithm. There are 100 frames in this data set and 6 frames are annotated. The original image sequences and annotated ground truth images is illustrated in Figure 1. From the sequence we can find that the foreground is moving and the cluttered background is dynamic. Most importantly, the camera is moving randomly at the same time of foreground moves, which cause the segmentation task very difficult.

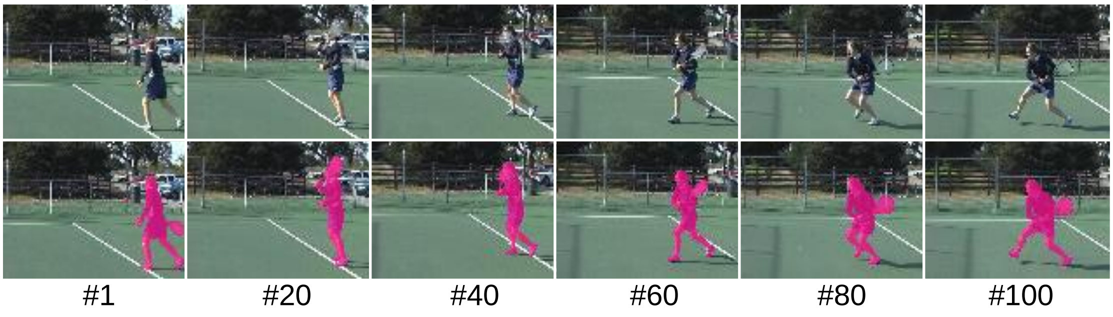

Figure 2. Original Video Sequences and Ground Truth

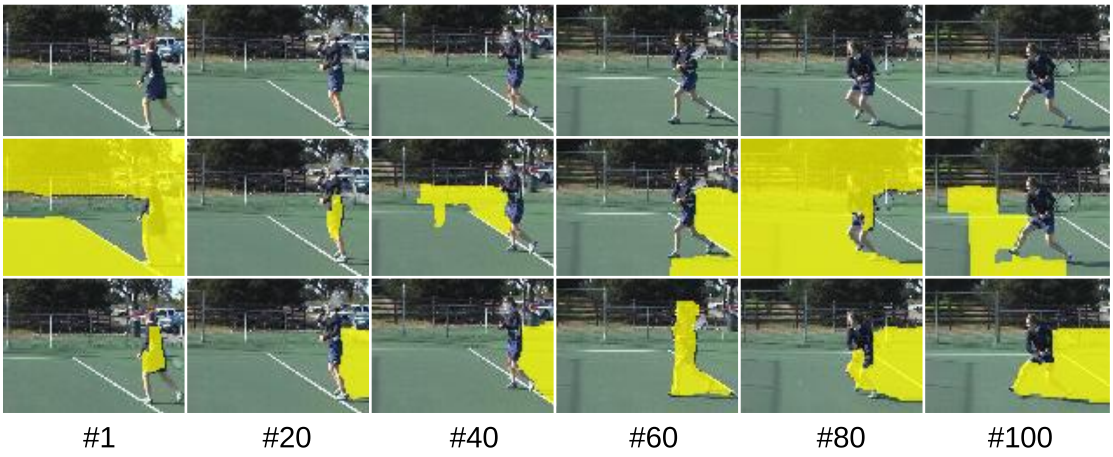

Figure 3. The First Row: Object Hypothesis with Highest Ranking Score; The Second Row: Segmentation Results after MRF energy Minimization

The second row of Figure 2 shows the object hypothesis result with the highest score using the proposed ranking method. From the hypothesis result we find that frame #1 and frame #40 are relatively good result of object candidates. Frame #40 and frame #100 have false positive object regions around the foreground object. The object region hypotheses in frame #1 and frame #80 estimate the whole playground as the foreground candidate. The unary and pairwise potentials of the spatial-temporal MRF is initialized using the object hypotheses with the highest score shown in the second row. The MRF model is minimized using graph-cut algorithm, which is very efficient since the MRF is based on superpixel patches of the original image sequences. In addition, since the problem is binary valued segmentation, thus the pairwise term is guaranteed to be submodular functions such that the MRF model can be minimized exactly using the graph-cut algorithm [11]. The segmentation result is shown in the third row of Figure 2. From the segmentation result we can find that, qualitatively, most of the foreground objects are segmented correctly using the unsupervised method.

Table 1. Quantitative Segmentation Results of *tennis* Data Set

Mean	AUC	AP	CR	CRp (%)
FOS [1]	0.0441	0.0442	13942	84.09
Our method	0.5707	0.5657	14301	92.66
Std	Std AUC	Std AP	Std CR	Std CRp
FOS [1]	±0. 0057	±0. 0057	±143	±0. 97
Our method	±0.0605	±0.0605	±64	±0.44

To assessment the segmentation result quantitatively, we use four segmentation criteria, which include Area Under Curve (AUC), Average Precision (AP), Correct Rate (CR) and CR in percent (CRp). The AUC and AP is derived precision recall curve by using vl-feat toolbox [12] over each frame. Then the AUC and AP are averaged over six frames of #1, #20, #40, #60, #80 and #100. The CR is computed using XOR operation of ground truth and segmented images and then minus by total number of pixels in each frame. The CRp is derived using CR and divided by the total number of pixels in each frame. The third row of Table 1 shows the quantitative segmentation result using the proposed algorithm and their respective standard division is shown in the last row. Compared with the result of a state-of-the-art Fast Object Segmentation (FOS) method [1], the result demonstrates that our method achieves robust and improved segmentation accuracy of the highly dynamic image sequences, the CRp improved by 10.19% and 92.66% pixels are correctly classified.

6. Conclusion

In this paper, we propose a foreground hypothesis framework for object segmentation in presence of highly dynamic scene. This framework enables foreground segmentation by ranking object hypothesis over spatial space to achieve a robust binary segmentation of a video sequence in an unsupervised way. Object candidates derived from spatial features in each frame are used to rank the object candidates over a specific hypothesis space, so as to yield object proposals with varying ranking scores. Based on the hypothesis with the highest ranking score, a Markov Random Field (MRF) model is use to estimate foreground hypothesis with highest score. We demonstrate the performance of our approach through experimental evaluation on a difficult dynamic scene segmentation benchmark from Freiburg-Berkeley Motion Segmentation Dataset, and show that our method achieves robust segmentation performance when dealing with highly dynamic image sequences.

Acknowledgements

This work was supported by the National Science Foundation of China (NSFC) under Grants 61461022 and 61302173. Yinhui Zhang was supported by the China Scholarship Council (File No. 201208535035) while studying as a visiting scholar at University of Miami.

References

[1] A. Papazoglou and V. Ferrari, “Fast object segmentation in unconstrained video”, IEEE International Conference on Computer Vision (ICCV), (**2013)**, pp. 1777-1784.

[2] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation”, International Journal of Computer Vision, vol. 59, no. 2, (**2004**).

[3] S. Fidler, R. Mottaghi, A. Yuille and R. Urtasun, “Bottom-up segmentation for top-down detection”, IEEE Conference on Computer Vision and Pattern Recognition, (**2013**), pp. 3294 – 3301.

[4] T. Brox and J. Malik, “Large displacement optical flow: Descriptor matching in variational motion estimation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33 no. 3, (**2011**) pp. 500-513.

[5] A. Radhakrishna, S. Appu, S. Kevin, L. Aurelien, F. Pascal and S. Susstrunk, “Slic superpixels”, Technical Report 149300 EPFL, (**2010**).

[6] J. Carreira and C. Sminchisescu, “CPMC: Automatic object segmentation using constrained parametric min-cuts”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 7, (**2012**), pp. 1312-1328.

[7] P. Arbelaez, M. Maire, C. Fowlkes and J. Malik, “Contour detection and hierarchical image segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, no. 5, (**2011**) pp. 898-916.

[8] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques and J. Malik, “Multiscale combinatorial grouping”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (**2014**), pp. 328-335.

[9] T. Brox and J. Malik, “Large displacement optical flow: descriptor matching in variational motion estimation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no.3, (**2011**), pp. 500-513.

[10] http://lmb.informatik.uni-freiburg.de/resources/datasets/moseg.en.html

[11] V. Kolmogorov and R. Zabih, “What energy functions can be minimized via graph cuts?”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 2, (**2004**), pp. 147- 159.

[12] D. Hoiem, Y. Chodpathumwan and Q. Dai, “Diagnosing error in object detectors”, ECCV, (**2012**).

Authors

Yinhui Zhang, he received the Ph.D. degree in the direction of image segmentation from Kunming University of Science and Technology, Kunming, China, in 2010. He is currently an Associate Professor of the Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology. His research interests include image processing and computer vision.

Zifen He, she received the Ph.D. degree in the direction of digital image halftoning from Kunming University of Science and Technology, Kunming, China, in 2013. She is currently an Associate Professor of the Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology. Her main areas of research are image processing and computer vision.

Xing Wu, he received the Ph.D. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2005. He is currently a Professor and Dean of the Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology. His research focuses on modern signal processing technology and its application in fault diagnosis. He published more than 20 journal and conference papers in the areas of remote monitoring and fault diagnosis and holds three national patents.