2016 IEEE 22nd International Conference on Parallel and Distributed Systems

Context-Aware Video Object Proposals

Wenjing Geng, Gangshan Wu
State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing, China
Email: jenngeng@gmail.com, gswu@nju.edu.cn

***Abstract*—Recent advances in object proposals have been achieved obvious performance to speed up sliding window based object detection or recognition. However, the spatial-temporal object proposal of multi-objects in video is still a challenging problem. Applying the existing image methods frame by frame will result in three defects. First, no guarantee to keep the consistent proposal results, i.e., it is hard to avoid omitting objects even in consecutive or similar sequences. Second, the latent information contained in time dimension would not be made best use of to improve the detection rate. Third, due to the motion blur caused by motion flow, the efficiency of object proposals relying on contour or edge features would be definitely degraded. In this paper, we propose an efficient method for video object proposals. By introducing image method into context-aware framework, we get the improved detection rate compared to the frame by frame usage, while keeping a controllable computing efficiency. Firstly, the bounding boxes produced by image proposals are used as the input. Then the candidate windows are scored with contextual information by generating motion-based mapping boxes. To evaluate the multi-object proposal results, we build a specific dataset. Experiments show that the proposed method can improve the detection rate of the original image method, and especially achieve better performance when proposing a small set of bounding boxes.**

***Keywords*-video multi-object proposals; motion based mapping; contextual re-scoring; multi-object detection dataset**

I. Introduction

Based on some visual cognitive and neuropsychological evidences, it is believed that human can quickly and accurately identify objects without recognizing them [1]. Due to the scientific acquaintance foundation, generating object proposals has become a promising and helpful pre-processing technique for fulfilling many computer vision tasks, such as object detection, recognition and classification. However, most state-of-the-arts are designed for image object proposals [2–7], for the common sense is that video object proposals can be solved by applying image methods frame by frame. Therefore, few solutions pay much attention to design specific pre-processing procedures for multi-object proposals in video. By thorough experiments, we found that it is not enough to directly apply image object proposals into video based on three facts.

- Omitting objects is inevitable even in consecutive or similar frames;
- Motion blur would degrade the edge or contour based proposal results;
- Temporal information preserved in neighboring frames should be utilized.

Fig. 1 (b) and (e) show the proposal results generated by [5] in five successive sequences. Red rectangles are the missing proposals compared to annotations. The inconsistency in time domain is obvious. Few methods specially serve for video multi-object proposals. Gilad et al. [8] aimed at finding the dominant objects in the scene and obtaining rough, yet consistent segmentations thereof. Due to the usage of multiple segmentations, it is inapplicable to serve as a pre-processing procedure. Van den Bergh et al. [9] proposed a novel method for the online extraction of video superpixels, contributing to delivering tubes of bounding boxes throughout extended time intervals. Though efficient in acquiring video superpixels, consistently tracking large amount of initial bounding boxes seems much too troublesome. Oneata et al. [10] explored the problem of generating video tube proposals for spatio-temporal action detection. This research is a branch of action detection in video, while our method devotes to proposing the category independent bounding boxes that probably contain objects no matter they are still or not. In brief, most related methods [11, 12] are explicitly defined to propose dominant objects or moving objects for video object detection. The complicated video content analysis determined that these methods may not suitable for pre-filtering. It is imperative to find a compromised way to tackle video multi-object proposals by leveraging advantages of image proposals and video intrinsic characters. Besides, the lack of comprehensive dataset with consecutively annotated ground truth in bounding boxes is another problem to be solved.

Generally, motivated by the challenges motioned above, we propose a context-aware video object proposal framework, aiming at minimizing the gap between image and video object proposals. First, candidate boxes are generated by image object proposals. Then these bounding boxes are scored by mapping them to respectively neighboring frames. The final score of each bounding box is summed up by weighted Gaussian. After ranking proposals based on these final scores, a series of bounding boxes are proposed. In order to quantitatively and objectively evaluate the proposed method, we build a new dataset with successive bounding box annotations for each frame in video. The explorations for these issues result in two main contributions as follows.

1521-9097/16 $31.00 © 2016 IEEE
DOI 10.1109/ICPADS.2016.158

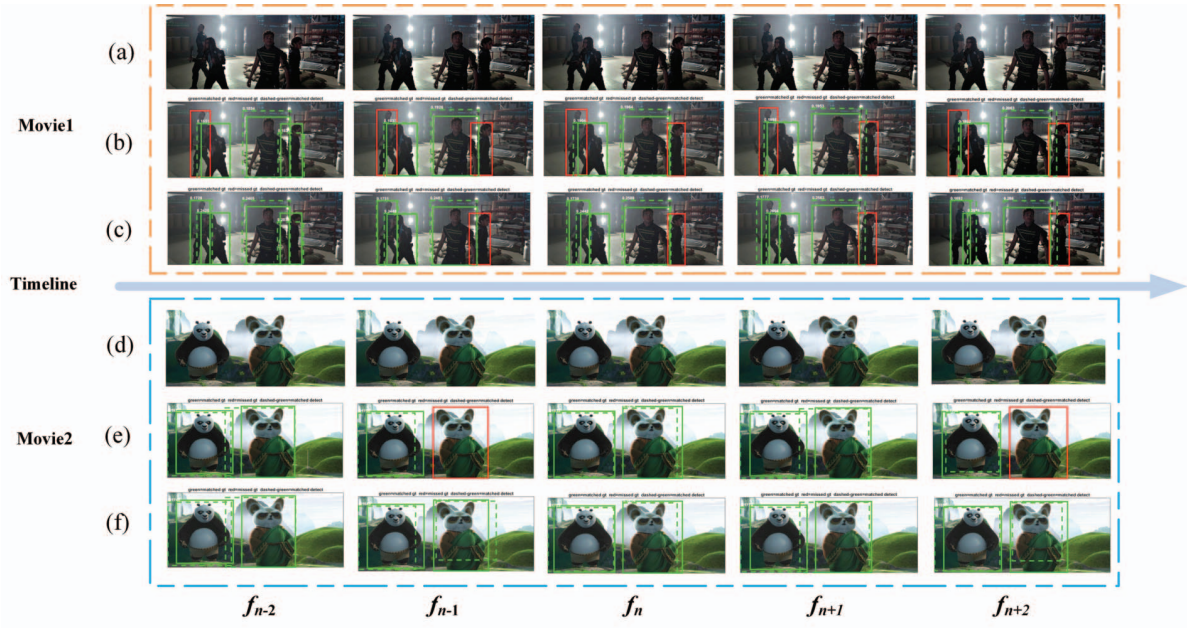


Figure 1. Results of object proposals generated by Edgebox [5] and ours. Solid red rectangles are the missing bounding boxes compared to the ground truth annotated by solid green boxes, and dashed green ones are the matched proposals. (a) and (d) are the successive video frames extracted from two movies, (b) and (e) are the object proposals generated by [5] frame by frame, and (c) and (f) are the proposal results generated by the proposed method.

- We propose a context-aware framework for building a bridge between image and video object proposals with minimum cost, which provides an efficient solution to improve the detection rate compared to the frame by frame usage;
- We build a specific dataset for video multi-object analysis with 25 different shots from 5 famous movies. The ground truth is annotated as the rectangle frame by frame, which can be used as an expert benchmark for multi-object detection in video.

II. Methodology

Fig. 2 depicts the overall procedures of the proposed framework, consisting of candidate bounding box generation, motion estimation, contextual-based re-scoring and bounding box re-ranking. Given a video, we aim at generating a serious of video object proposals by leveraging advantages of image object proposals and the basic feature in video. Our solution devotes to minimizing the additional computing cost as much as possible to make it suitable for a pre-filtering process and improving the detection rate compared to the frame by frame usage of image proposals.

A. Candidate bounding box generation

Our key idea is to utilize the redundant information preserved in neighboring video frames while maintain a controllable computing increase, for the solution should be a better alternative method for a video pre-filtering process. By comprehensively reviewed the related work about object

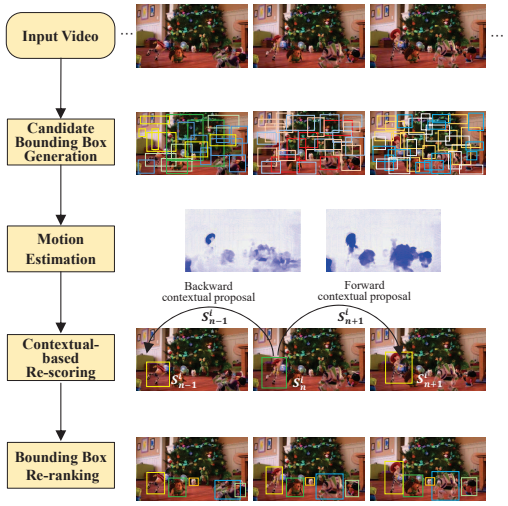
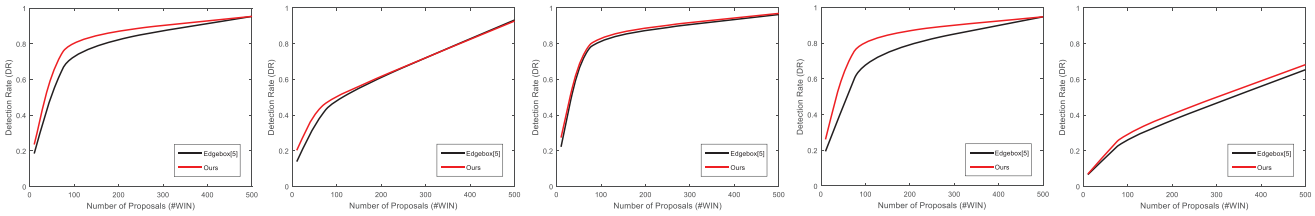


Figure 2. The framework of the proposed method.

proposals, objectness-based and merging-based methods can be viewed as two main trends. As merging similarity based methods aims at reaching accurate segments by requiring much more computations compared to the object-based solutions, they are not suitable for preprocessing. On the contrary, object-based methods can achieve high computing efficiency by designing some invariant features from generic objects [4–7]. Therefore, a set of candidate bounding boxes



(a) (b) (c) (d) (e)

Figure 3. The detection rate of the shots from different movies. (a)-(e) are separately from Mission Impossible, Monsters University, Kung Fu Panda, X-Men and Toy Story. The red curve are the results generated by our method, while the black curve are the results generated by [5] with $o = 0.7$.

should be firstly initialized. Considering the computing efficiency, these boxes can be generated by image proposals. In order to illustrate the context-aware framework more clearly, we formulate the key procedure for each processing module in a concise way. Let n represent the n_{th} frame f_n of one video shot V, the generated candidate bounding box C_n can be shown as:

$$C_n = \{b_i | b_i \in I(f_n, \alpha M) \ , \alpha \geq 1\}, \tag{1}$$

where M is the initial number of object proposals, and α is a loose parameter used for adjusting M. Experiments show that the smaller α will not degrade the detection rate too much, so we adopt $\alpha = 1.2, M = 1000$ in the experiments. $I(\centerdot)$ represents image object proposal method which can return some bounding boxes with corresponding scores.

B. Motion estimation based mapping

As to video processing, motion flow should always be calculated. Therefore, it is intuitional to apply optical flow into video object proposals for it reflects the movement of pixels. Considering the computational efficiency, a beyond pixels method proposed in [13] is utilized in our experiments. In fact, any optical flow algorithm which has a good performance in leveraging commuting efficiency and accuracy can be adopted. Because of bidirectional contextual information preserved in video, bidirectional optical flow should be calculated at the same time. Let the context length of frame n be $2k$, then the bidirectional motion-based contextual mapping bounding boxes C_{n-k} and C_{n+k} can be formulated as:

$$C_{n-k} = \{b'_i | b_i = Mapping(b_i, m_{n-k}) \ \& \ b_i \in C_n\}, \tag{2}$$

$$C_{n+k} = \{b'_i | b_i = Mapping(b_i, m_{n+k}) \ \& \ b_i \in C_n\}, \tag{3}$$

where m_{n-k} is the optical flow calculated from n_{th} frame to its previous k_{th} frame, and m_{n+k} is the optical flow calculated from n_{th} frame to its post k_{th} frame. $Maping(\centerdot)$ is the optical flow based motion mapping function.

C. Contextual re-scoring and re-ranking

The key procedure for window scoring based object proposals depends on scoring strategy. Experiments show that evaluating bounding boxes frame by frame will lead to object omitting. To avoid this kind of deficiency, contextual information should be introduced into the video object proposals in a proper way. To achieve this target, we build a context-aware framework. Firstly, generating limited window candidates as described in Sec. II-A. Secondly, mapping the bounding boxes of the current frame forward and backward as described in Sec. II-B. The objective of mapping is to get a series of corresponding boxes and acquire new scores based on contextual frame information. The contextual re-scoring can be calculated as follows.

$$S^i_{n-k} = I(f_{n-k}, C_{n-k}), \tag{4}$$

$$S^i_{n+k} = I(f_{n+k}, C_{n+k}). \tag{5}$$

The final score S_n of bounding box b_i in frame f_n is calculated by a weighted Gaussian, which is described in Eq. (6).

$$S^i_n = w_p \centerdot S_{n-k} + ... + w_q \centerdot S_n + ... + w_r \centerdot S_{n+k}, \tag{6}$$

where the subscript weighted parameters are in the subset of $\{p, ...q, ...r \in [3, k]\}$.

The advantage of the proposed framework lies in that we did not take much attention to design a new scoring strategy for video. It is noted that we manage to design a tactics by using the image scoring strategy to score itself. Re-ranking is performed on the weighted scores calculated between bounding boxes generated by image methods and the motion-based mapping boxes propagating among neighboring sequences. Therefore, the proposed framework can achieve better detection rate on the same recall compared to utilize image methods on single frame.

III. Experiments

A. Experimental settings

The proposed framework is validated on the dataset built in this paper, for there is no suitable dataset for window based video multi-object proposals. Our dataset collected from five famous movies. Each movie randomly contributes 5 shots, annotated by 12 subjects. The details of the built dataset are depicted in Table I. GT is short for ground truth, annotated by the rectangle frame by frame. Objects in each

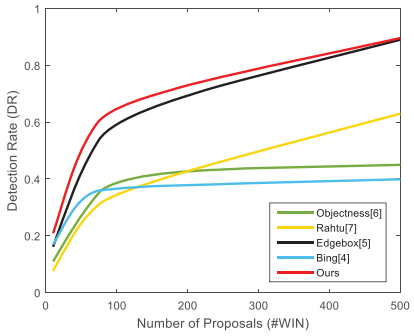


Figure 4. The detection rate curves of different methods on the proposed dataset with IoU=0.7 and #WIN=500.

frame are labeled by drawing bounding boxes around them. Ave. Obj. is the average number of objects contained in each shot. Ave. Frame is the average number of frames in each shot. In performance evaluation, detection rate (DR) with given number of windows (#WIN) (DR-#WIN) is used to illustrate the improvement, which is defined in Eq. (7).

$$\text{DR-\#WIN} = \frac{\#(o > \epsilon)@\text{\#WIN}}{\#\text{GT}} \quad \epsilon \in \{x | 0.5 \leq x \leq 1\}, \quad (7)$$

where o is the criterion of intersection over union. We adopt $o = 0.7$ in our experiments for it is sufficient in the real applications. Furthermore, the Edgebox image proposals presented in [5] is utilized to generate window candidates, for it achieves the best performance over the state-of-the-art both in accuracy and efficiency.

B. Result and discussion

As the built dataset is collected from five movies, we firstly evaluate on the different sources. Fig. 3 shows the separate quantitative evaluation results of the shots from different movies. We found that the improvement varies from different sources, because the degree of motion blur differs from each shot. As Edgebox image proposals mainly rely on the edge feature, the fuzzy boundary caused by motion will definitely degrade the proposal results. That is to say, the proposed method can achieve better results if the video quality is not good. Fig. 3 (a), (d) and (e) present big improvements because these shots contain more motion blur or have lower color contrast. On the contrary, Fig. 3 (b) and (c) illustrate less improvements for the scenes of these shots have clear edge boundaries. Fig. 4 depicts the comparisons of the detection rate on the proposed dataset. Comparing to the state-of-the-art, it is obviously shown that the proposed method can indeed improve the single frame results under the IoU 0.7. The only increasing calculation is the computing of optical flow, which can be controlled within an acceptable consumption by adopting the methods leveraging both accuracy and efficiency.

Table I
DISCRIPTION OF THE PROPOSED DATASET

Shot Source	Resolution	GT	Ave. Obj.	Ave. Frame
Mission Impossible	640*268	Y	2.4	67
Monsters University	640*360	Y	3	59
Kung Fu Panda	640*272	Y	2	66
X-Men	640*266	Y	2.4	34
Toy Story	960*540	Y	4.2	77

IV. CONCLUSIONS

Motivated by introducing image object proposal methods into video and yielding twice the result with half the effort, a context-aware framework by utilizing contextual information preserved in video frames is proposed. To evaluate the efficiency of video multi-object proposal, we build a specific multi-object dataset with bounding box based ground truth annotated frame by frame. Experiments on this challenge dataset show that the proposed approach outperforms the performance by utilizing the state-of-the-art method on single frame in sequence.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation of China under Grant No. 61321491 and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *TPAMI*, 2014.

[2] J. Liu, T. Ren, and J. Bei, "Elastic edge boxes for object proposal on rgb-d images," in *MMM*, 2016.

[3] X. Chen, H. Ma, X. Wang, and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," in *CVPR*, 2015.

[4] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014.

[5] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[6] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *TPAMI*, 2012.

[7] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *ICCV*, 2011.

[8] G. Sharir and T. Tuytelaars, "Video object proposals," in *CVPRW*, 2012.

[9] M. Van den Bergh, G. Roig, X. Boix, S. Manen, and L. Van Gool, "Online video seeds for temporal window objectness," in *ICCV*, 2013.

[10] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, "Spatio-temporal object detection proposals," in *ECCV*, 2014.

[11] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, and C. G. Snoek, "Action localization with tubelets from motion," in *CVPR*, 2014.

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

[13] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, 2009.