Multimed Tools Appl (2018) 77:10589–10614
DOI 10.1007/s11042-017-4561-9

Adaptive video object proposals by a context-aware model

Wenjing Geng[1] · Chunlong Zhang[1] · Gangshan Wu[1]

Received: 23 December 2016 / Revised: 21 February 2017 / Accepted: 27 February 2017/
Published online: 17 April 2017

Abstract Most previous works focus on image object proposals while few on video object proposals. Besides, the existing explorations about video object proposals mainly concentrate on localizing the dominant object. In this paper, we aim at exploring a uniform framework for proposing multi-objects in videos no matter they are in the foreground or background. The method is derived from image object proposals, and makes best use of video characteristics. To achieve this task, we propose an adaptive context-aware model for video object proposals. First, spatial candidate windows are generated by the image method for acquiring the adequate bounding box samples. Temporal boxes are calculated by the motion based mapping. Considering the mapping loss, we define a box confidence coefficient contributing to keeping the proposal consistency and restraining the motion blur. The output proposal bounding boxes are ranked based on the scores calculated by the weighted scoring system. The proposed method is separately evaluated on the proposed multi-object dataset and the public dataset. The results compared with several state-of-the-arts show that our method has the most satisfactory overall performance for multi-object proposals in videos.

Keywords Multi-object proposals · Spatial and temporal windows · Adaptive context-aware model · Proposal consistency · Multi-object dataset

✉ Gangshan Wu
gswu@nju.edu.cn

Wenjing Geng
jenngeng@gmail.com

Chunlong Zhang
clzhang.nju@gmail.com

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

1 Introduction

Object proposal generation, i.e., proposing the object-like regions from millions of sliding windows, has become a promising and helpful technique both in multimedia and computer vision [44, 56, 57]. It is derived from some visual cognitive and neuropsychological evidence that human can quickly and accurately identify objects without recognizing them [15, 16]. By object proposals, multimedia tasks can concentrate on a few of proposal bounding boxes probably containing objects, instead of starting from millions of sliding windows. Utilizing object proposals as a pre-processing procedure, benefits many applications both in efficiency and effectiveness, e.g. object tracking [26, 67], image/video segmentation [21, 43] and classification [13, 31, 59, 65], video summarization [22, 38], activity recognition [6, 66], object retrieval [47, 58, 63], multimodality retrieval [10, 12, 61], landmark recognition [11, 36, 50], and image/video storytelling [14, 35, 62]. Furthermore, it is better to consider domain knowledge when applying object proposals to some specific or novel applications.

For most video object proposals start from image object proposals [27, 40, 56], there are generally two categories of object proposals for both image and video, segment-based proposals [3, 51] and window-based proposals [24, 32]. The former model usually starts from generating multiple segments and then merges them into the proposed regions. To get better segments, the sophisticated algorithm is prerequisite which brings more computational consumptions. In contrast, the goal of window-based proposals aims at assigning high scores to the bounding boxes that probably contain objects. Due to the lightweight design, even some low-level features can achieve good results both in accuracy and efficiency [9, 68]. Considering the procedures of human vision to recognize objects, roughly localizing and accurately recognizing, the latter model is more intuitive and suitable for pre-processing. In this paper, we focus on applying the latter model to videos because of its simplicity, which contributes to making object proposals much more efficient, especially running as a pre-processing procedure in videos.

Although many works concentrate on object proposals [8, 33, 34], few previous works focus on video object proposals in recent years. To the best of our knowledge, most existing video object proposals mainly devote to proposing moving or dominant objects [27, 40, 53]. Merely locating moving objects has not achieved the goal set by object proposals. It is similar to the technique of moving object segmentation and tracking [28, 54]. To achieve multi-object proposals in videos, it is straightforward to apply object proposals [2, 9, 45, 68] frame by frame. But experiments find that applying image methods frame by frame may lead to proposal inconsistency. This experimental phenomenon is illustrated in Fig. 1b, obviously showing that proposal inconsistency even exists within frames with similar content and structure. Despite the high detection rate of image object proposals, directly applying these methods to videos still results in omitting objects. Two reasons lead to this omitting. One is that there are no dynamic cues in image object proposals because they are designed for detecting static objects. The other is that motion blur and color ambiguities will degrade the edge or contour based proposal results. Inspired from all the above, we further explore the criteria that should be considered when applying object proposals to videos.

Good extendibility Object proposals have been studied a lot and remarkable achievements have been made. It is wasteful to leave image achievements alone and demonstrate a different path for videos. Therefore, it is better to seek for a good scheme to extend image methods to videos.

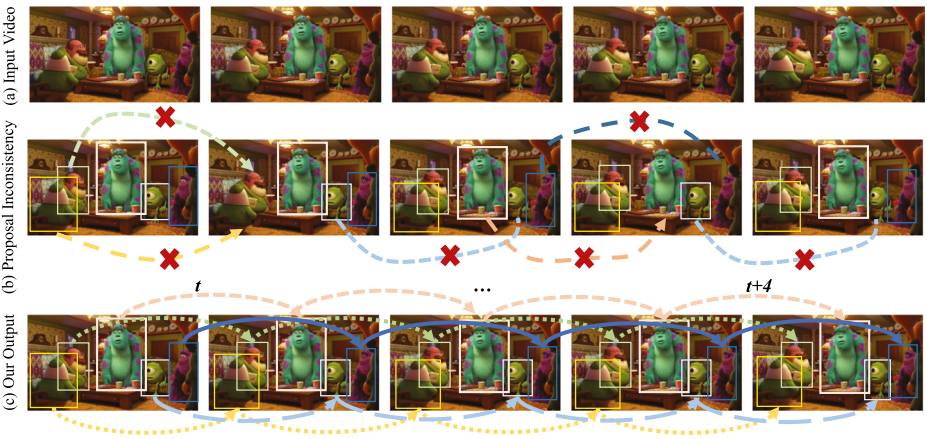

Fig. 1 Given (**a**) an unlabeled video, our model produces (**c**) a set of spatio-temporal bounding box proposals for both foreground and background objects at the same time. (**b**) shows the proposal inconsistency caused by utilizing image object proposals frame by frame

Multi-object proposals As a pre-processing procedure, a good video object proposal method should propose all the object-like regions no matter they are in the foreground or not, dominant or not. Besides, proposing multi-objects will benefit more applications.

Proposal consistency constraint Directly applying object proposals frame by frame may lead to three defects. First, omitting objects is inevitable even in consecutive or similar frames. Second, motion blur and color ambiguities may degrade the edge or contour based proposal results. Third, temporal information preserved in neighboring frames is not being used.

Based on the above considerations, we propose an adaptive context-aware model for multi-object proposals in videos. Image methods are extended into videos by considering motion cues and spatial-temporal evolutions in our model. The increasing computational complexity lies in the motion cues' calculation, which is inevitable in video processing. To achieve multi-object proposals, both spatial and temporal bounding box generations are considered. By adaptively integrated spatial and temporal proposals, the proposal consistency constraint is kept as much as possible. To evaluate the efficiency of the proposed method, we build a multi-object dataset specially for video object proposals. 30 shots are collected from five famous movies. This benchmark is suitable for evaluating multi-object proposals in videos, for the average number of objects achieves 3.34 and the ground truth is offered frame by frame. For a comprehensive evaluation, we also compare our method with the state-of-the-art on a public motion segmentation dataset, called Freiburg-Berkeley motion segmentation dataset [5, 39], including 30 shots with keyframe ground truths. We extend this motion segmentation dataset to video object proposal dataset by offering annotations per frame in bounding boxes. The proposed method achieves good performance on both datasets, showing that our method is competent for video object proposals. A similar work to this paper was proposed in [18], but the context-aware model is absolutely bidirectional and no classifications are considered when conducting motion estimation based mapping. Different to [18], we present an adaptive context-aware model with more elaborate processings. It can be transformed into a unidirectional model according to the temporal sequence by omitting the temporal scoring refinement. Therefore, it is more efficient while

with improved detection rate, which means that it is much easier to apply to real time applications. Besides, we also expand the proposed dataset and validate the effectiveness of our method in a public dataset.

The contributions of this paper can be briefly stated as follows.

- We propose an adaptive context-aware model for video object proposals, which contributes to proposing both still and moving objects no matter they are in the background or foreground in videos;
- We integrate the spatial and temporal boxes by introducing an adaptive and classified motion based mapping, which can be extended to other applications. On account of no complicated computations, the efficiency is high enough;
- We employ a temporal scoring refinement mechanism to further improve the detection rate;
- We build a challenging dataset for multi-object proposals in videos, which is collected from five famous movies with 30 shots in total (about 3.34 objects per frame), and has bounding box annotations frame by frame.

The rest of this paper is organized as follows. In Section 2, we give a review of related works on both frame-based and sequence-based object proposals in videos. Then we introduce the main body of the proposed context-aware model in Section 3 and the temporal scoring refinement in Section 4. Next, we demonstrate the experimental results and give some discussions of the results in Section 5. At last, we give a brief conclusion and perspective of our work in Section 6.

2 Related work

Few methods are specially designed for video multi-object proposals. Most of them usually start from per-frame object proposals which are then generalized in the temporal domain. According to our survey on the related work, we generally divide video object proposals into frame-based object proposals and sequence-based object proposals.

Frame-based object proposals The concept of object proposals is firstly presented by Alexe et al. in [1] aiming at reducing the number of true negative sliding windows. They further explored the solution and introduced a generic objectness measure to quantify the possibility of containing objects for the candidate windows. Rahtu et al. [45] scored the windows by utilizing an effective linear feature combination, which achieved better results compared with [1]. By adopting low-level features, such as gradient, saliency, and superpixel straddling [1], most methods can achieve good performance in both efficiency and accuracy. Along with maturity of current techniques, image methods can be directly applied into video frame by frame. Cheng et al. [9] proposed a very fast method to filter the initial sliding windows at 300fps by merely utilizing the gradient feature. It seems that this method has already fulfilled the demands of real time applications in videos, but it can only achieve better results on 0.5 intersection over union (IoU), which is not applicable in practical applications. Zitnick et al. [68] leveraged accuracy and efficiency very well by using the edge feature. Although it has the best performance even over the challenging overlap among the window-based proposals, the computing time has no competitive advantages compared with [9]. As to segment-based proposals, such as [3, 37, 46, 52], though achieving accurate segmented results to some extent, complicated computations may bring much more time-consuming when applying to videos, making them unsuitable for running as preprocessing procedures. In short, although frame-based object proposals can achieve the task

of video multi-object proposals, frame-by-frame usage may lead to proposal inconsistency among temporal sequences according to experiments. Therefore, temporal information should be subtly adopted into video object proposals with limited increase of computational efforts.

Sequence-based object proposals Few methods specially serve video multi-object proposals. Most related works about video object proposals mainly aim at proposing dominant objects in the temporal domain, which we call them sequence-based object proposals. Gilad et al. [48] aimed at finding the dominant objects in the scene and obtaining rough, yet consistent segmentations thereof. Due to the usage of multiple segments, it is inapplicable to serve as a pre-processing procedure. Van den Bergh et al. [53] 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, it is similar to the task of object tracking. Oneata et al. [40] explored the problem of generating video tube proposals for spatio-temporal action detection. This research is a branch of action detection in videos, while our method devotes to proposing the category independent bounding boxes that probably contain objects no matter they are still or not. Perazzi et al. [42] performed an SVM-based pruning step to retain high quality foreground proposals. Xiao et al. [56] presented an unsupervised approach to generate spatio-temporal tubes that localize the foreground objects. Though considering the importance of proposal consistency, these methods aim at keeping the proposal consistency of foreground objects. In brief, most related methods [27, 41, 60] are explicitly defined to propose dominant objects or moving objects for video object detection. These methods seem to be moving object segmentations rather than video object proposals.

With the emergence of deep learning, increasing works turn to deep architecture for help. There is no exception in the task of object proposals [19, 20, 30]. Zhang et al. [64] leveraged a Convolutional-Neural-Network model to generate location proposals of salient objects. Kong et al. [29] presented a deep hierarchical network for handling region proposal generation and object detection jointly. Hayder et al. [23] proposed an approach to co-generate object proposals in multiple images by introducing a deep structured network that jointly predicted the objectness scores and the bounding box locations of multiple object candidates. Though most of these methods have achieved pleasing results, Chavali et al. [7] reported the gameability of the current object proposal evaluation protocol especially for learning-based methods, for they argued that the choice of using a partially annotated dataset for evaluation of object proposals is problematic. Learning-based methods define an object as the set of annotated classes in the dataset, which obscures the boundary between a proposal algorithm and an object detector. In order to generalize object proposals as a pre-processing procedure in videos and localize category independent objects as much as possible, the low-level feature is more receivable and explicable.

3 Adaptive context-aware object proposal model

Given a video, we aim at generating a series of spatio-temporal bounding box proposals for both foreground and background objects at the same time by leveraging advantages of image object proposals and the basic feature in videos. 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 with the frame-by-frame usage of image object proposals. The main procedures are outlined in Fig. 2, including spatial candidate

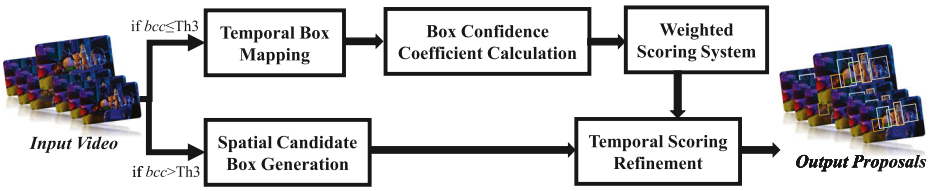

Fig. 2 The framework of the proposed method. *bcc* is the box confidence coefficient

box generation, temporal box mapping, box confidence coefficient calculation and weighted scoring system. The temporal scoring refinement will be introduced in Section 4.

3.1 Spatial candidate box generation

The standard practice for generating initial bounding box is starting from densely sampled sliding windows. Millions of these windows are filtered by well-designed selective rules. In fact, there is no need to generate so many boxes for every frame. Proposal boxes generated by image methods can be used as initial candidate boxes. There are three benefits of generating spatial candidate boxes by image methods. First, image methods can generate proposals including both foreground and background objects. Second, the detection rate of image methods has increased a lot for meeting the demands of applications. Third, starting from image methods will significantly increase the computing efficiency because of saving time for handling so many boxes. Let t represent the t_{th} frame f_t of one video. n is the number of generated spatial candidate boxes $B_n^{(t)}$ which can be shown as:

$$B_n^t = \{b_i | b_i \in I(f_t, M), n \leq M\}, \tag{1}$$

where M is the maximum number of generated bounding boxes. The computing consuming can be adjusted by setting M.

3.2 Temporal box mapping

As a pre-processing procedure, there is no need to pay much attention to feature extraction or bounding box matching. In order to make an effective temporal box mapping, we classify the surrounded relationship between the bounding box and the object based on the motion fields. Each frame has a corresponding motion field, e.g., calculated by an optical flow method. We manage to use motion distribution to guide the temporal box mapping. It is obvious that the bounding box should be moved according to the motion of main part surrounded by it, which is regarded as the displacement of the object. Therefore, it is significant to find the exact displacement of the object. To achieve this target, we firstly optimize the initial bounding box set B_n^t of frame t by making each box approach to the boundary of objects. Let $b_i (b_i \in B_n^t)$ represent one generated bounding box, the coordinates of b_i can be denoted as:

$$c_{b_i} = \{(x_l^i, y_l^i) | l \in [1, P], l \in N^+, P \geq 2\}, \tag{2}$$

where (x_l^i, y_l^i) are the coordinates of bounding box b_i. l is the numerical order of sampling points which is greater than 2. $l \in N^+$ means that l is a positive integer. We use four corners as sampling points in our experiments.

Let f_{b_i} represent the corresponding motion field of bounding box b_i, the optimization can be illustrated as (3).

$$c_{b_i^o} = \arg\min_{x_l^i, y_l^i} \sum [f_{b_i}(\Gamma(\Phi(x_l^i, y_l^i, \beta))) - f_{b_{i_c}}(x_c^i, y_c^i)], \tag{3}$$

where $\Gamma(.)$ represents a transformation of coordinates, and $\Phi(.)$ does shrink to the bounding box's coordinates. β is the step size rate for each shrink which equals 0.1 in our experiments. In optimization process, we utilize $\Gamma(.)$ transformation to find the midpoint of each edge of the rectangle with the bounding box's shrinking. Because of comparing with the four corners, midpoints have more possibilities to approach the object. It is more helpful than relying on the four corners to map the temporal box. Our new temporal box is mapped based on its corresponding motion map. Not every bounding box can be optimized within fixed iterations. If the box can be converged, the new coordinates $c_{b_{i,1}^{t+1}}$ can be denoted as shown in (4).

$$c_{b_{i,1}^{t+1}} = Mapping(c_{b_i^{t,o}},\ \omega f_{b_i^{t,o}}(\Gamma(x_l^i, y_l^i)) + (1-\omega) f_{b_{i_c}^{t,o}}(x_c^i, y_c^i)), \tag{4}$$

where $Mapping(.)$ is a motion field based mapping function that transforms pairs of coordinates to another. ω is used to weight the object's displacement for suppressing the noisy motion as much as possible. As to the bounding boxes that cannot be optimized, motion mapping is performed on the four corners based on the corresponding motion. It is noted that we utilize a median filter with $s \times s$ patch to filter the motion vector of each corner to do denoising. Then the mapped coordinates of bounding box $b_{i,2}^{t+1}$ can be denoted as:

$$c_{b_{i,2}^{t+1}} = Mapping(c_{b_i^t},\ \omega f_{b_i^t}(\Gamma(x_l^i, y_l^i)) + (1-\omega) f_{b_{i_c}^t}(x_c^i, y_c^i)), \tag{5}$$

The difference of (4) and (5) lies in the referred bounding box. If the initial generated bounding box can be optimized, then the input for mapping temporal box is the optimized box b_i^o, while the input is the original box b_i. This tactic contributes to making the proposed bounding box fit the object's boundary as much as possible. The final mapped temporal box set can be described as:

$$B_n^{t+1} = \{b_i^{t+1} \mid b_i^{t+1} \in b_{i,1}^{t+1}\ or\ b_i^{t+1} \in b_{i,2}^{t+1},\ i \in [1, n]\}, \tag{6}$$

Merely guided by the motion field, not every bounding box can be successfully optimized. The strategy is that keeping the bounding box containing the background object moving with the object yet without obvious appearance change. Meanwhile, making the bounding box containing the moving object shifting with the object yet with approaching change to its surrounding object. We also give a detailed algorithm description in Algorithm 1.

Algorithm 1 Procedure for the Temporal Box Mapping

procedure TBOXMAPPING(TBM)
  INPUT: A set of bounding box B_n^t of t_{th} frame, and corresponding motion field $f_{B_n}^t$.
  OUTPUT: A set of temporal mapped bounding box B_n^{t+1} of $(t+1)_{th}$ frame.
  Initialize $iter, Th1, diff$.
  while not converged **do**
    Shrink each bounding box by $b_i^t = \Phi(x_l^i, y_l^i, \beta)$ with a fixed step size β.
    Transform (x_l^i, y_l^i) to center point of each boundary by $(x_l^i, y_l^i)' = \Gamma_1(x_l^i, y_l^i)$.
    Transform (x_l^i, y_l^i) to box center by $(x_c^i, y_c^i)' = \Gamma_2(x_l^i, y_l^i)$.
    Calculate the corresponding motion fields $f(x_l^i, y_l^i)'$ and $f(x_c^i, y_c^i)'$.
    Update $diff = \| f(x_l^i, y_l^i)' - f(x_c^i, y_c^i)' \|$.
    $iter = iter + 1$.
    if $diff \leq Th1$ **then**
      Select the optimized $b_i^{t,o}$.
      Calculate the optimized temporal box $b_{i,1}^{t+1}$ by (4), and put them in B_n^{t+1}.
    end if
  end while
  Mapping the remaining box b_i^t by (5), and put them in B_n^{t+1}.
end procedure

3.3 Box confidence coefficient calculation

Due to the motion blur, not every temporal mapping can bring the pleasing result. For example, inaccurate motion fields may lead to ambiguous displacements. In order to reduce the impact of obscure moving, not every frame is suitable for temporal box mapping. Therefore, an adaptive strategy should be introduced to determine whether the bounding boxes of the current frame can be temporally mapped or not. Different from directly evaluating the accuracy of motion fields, we introduce a concept of box confidence coefficient bcc, which is calculated by making statistics of the box loss of each frame after achieving the temporal mapping, as shown in (7).

$$bcc = \frac{\mathcal{N}_{b_{loss}^t}}{\mathcal{N}_{B_n^t}}, \tag{7}$$

where b_t^{loss} represents the set of lost bounding boxes, which can be denoted as (8):

$$b_{loss}^t = \{b_i^t \mid w_{b_i^t} h_{b_i^t} \leq Th2\}. \tag{8}$$

$w \cdot h$ is the area of one bounding box, i.e., the number of pixels. It is assumed that the mapping error may increase along with the increment of smaller bounding boxes. We utilize (9) to make a decision whether the bounding boxes of current frame can be mapped or not. If $D = 1$, we recommend generating bounding boxes by temporal mapping. The detailed procedures for achieving adaptive context-aware temporal mapping are illustrated in Algorithm 2. It is different from Algorithm 1. Algorithm 1 describes the procedure for generating temporal bounding boxes, while Algorithm 2 emphasizes the procedure when to generate boxes by spatial methods.

$$D^{t+1} = \begin{cases} 1, & bcc \leq Th3 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Algorithm 2 Procedure for the Adaptive Context-aware Temporal Mapping

procedure ADAPTIVECTM(*ACTM*)
  INPUT: A set of temporal mapped bounding boxes B_n^t of t_{th} frame.
  OUTPUT: Object proposals $\mathbf{B}_n$ of t_{th} frame.
  Initialize $Th2$, $Th3$, M.
  Calculate the area $w_{b_i^t} h_{b_i^t}$ for each bounding box b_i^t in B_n^t.
  Compare each $w_{b_i^t} h_{b_i^t}$ with $Th2$, and keep statistics of $\mathcal{N}_{b_{loss}^t}$.
  Count the number of mapped bounding boxes $\mathcal{N}_{B_n^t}$ for t_{th} frame.
  Calculate $bcc = \frac{\mathcal{N}_{b_{loss}^t}}{\mathcal{N}_{B_n^t}}$
  if $bcc \leq Th3$ **then**
    $\mathbf{B}_n = B_n^t$.
  else
    $\mathbf{B}_n = I(f_t, M)$ by (1).
  end if
end procedure

3.4 Weighted scoring system

Bounding box based object proposals are ranked based on window scores. For spatial boxes are generated by image methods, window scores are assigned by the scoring system of image method. The spatial score $s_{b_i}^{t,s}$ of one bounding box b_i for t_{th} frame can be denoted as:

$$s_{b_i}^{t,s} = IS(f_t, b_i^t), \tag{10}$$

where $IS(\cdot)$ is the scoring system of image method. It is different from the function $I(\cdot)$ described in Section 3.1. The former is used for assigning a score for each window, while the latter serves as the spatial bounding box generation by the image object proposals.

As to temporal mapped bounding boxes, there are two steps to get the final scores. First, these mapped windows should be scored by (10) to get their spatial scores. Second, considering that these windows are mapped from the previous frame, temporal impacts should be considered into the scoring system. To simplify the scoring procedure, we adopt a linear weighted scoring strategy for temporal mapped bounding boxes, as shown in (11).

$$s_{b_i}^{t,tm} = \lambda s_{b_i}^{t-1} + (1-\lambda) s_{b_i}^{t,s}, \tag{11}$$

where $s_{b_i}^{t,tm}$ means the temporal score $s^{t,tm}$ for the bounding box b_i. Although the scores for temporal mapped windows are assigned between two neighboring frames, the mapping relationship can occur within several frames. The bounding boxes and the scores of those boxes are synthetically considered and acquired in the temporal sequences by the proposed context-aware model because of the global and temporal strategies.

4 Temporal scoring refinement

Although the improved results can be acquired by merely utilizing the above processings, further refinement can also be applied to the generated window-based proposals. There are two ways to do the refinement. One is adjusting the shape of generated proposals. The other is refining the scores of generated proposals to make sure that the proposals with objects could get top ranks. We choose to do the temporal scoring refinement for we have achieved

the pleasing improvement in the previous processing steps. Besides, our method aims at presenting a pre-processing routine, which means that the refinement strategy should be designed as simple as possible.

Our strategy is derived from temporal consistency constraint. Good methods should meet the demands of temporal consistency when they are applied into videos [4, 25, 55]. As to scores of neighboring frames, the consistency constraint is also essential, i.e., the mapped box should have scoring consistency compared with its neighboring boxes. We utilize a centered moving average filter to do the temporal scoring refinement. Therefore, the refined score $s^r_{b_i}(t)$ of the bounding box b_i in t_{th} frame can be updated by (12).

$$s^r_{b_i}(t) = \frac{s(t - \frac{\varpi-1}{2}) + s(t - \frac{\varpi-1}{2} + 1) + ... + s(t) + ... + s(t + \frac{\varpi-1}{2})}{\varpi}, \quad (12)$$

where ϖ is an adaptive moving window size to eliminate the score noise across the temporal domain separated by (9). It is noted that our refinement is performed on the temporal domain, centered by the current frame. Therefore, it can be used as a post-processing procedure to further improve the detection rate. Meanwhile, due to only relying on temporal score denoising, the improvement is limited.

5 Experiment and analysis

5.1 Dataset

The proposed method is evaluated on two datasets. One is designed for multi-object proposals and built in this paper. The other is a public dataset specially for motion segmentation, called Freiburg-Berkeley motion segmentation dataset [5, 39]. The former dataset is built from five famous movies, Mission Impossible, Monsters University, Kung Fu Panda, X-Men and Toy Story. We randomly select six shots from each movie, forming 30 shots in total. Five subjects, three men and two women, are invited to annotate the dataset. They firstly annotated some keyframes and then mapped the bounding boxes to the other frames by motion-based mapping. Finally, they adjusted the annotations with obvious offsets. By this way, we offered bounding box annotations for every frame in the proposed dataset. For a multi-object proposal dataset, the average number of the proposed dataset achieves 3.34. The detailed description of our dataset is depicted in Table 1.

As to the FBMS-59 dataset [5, 39], there are 29 shots for training set and 30 shots for testing set. For there is no learning process in the proposed method, i.e., the proposed framework is an unsupervised method, we only adopt the testing set to do the experiments. Because of being designed for motion segmentation task, the ground truths are object segmentations. Besides, only few keyframes are labeled. In fact, video frames are different from image sets. Although containing consecutive frames, objects may not occur in every

Table 1 Description of the proposed dataset. Num. Shot is the number of shots, Ave. Obj. is the number of average objects, and Ave. Frame is the number of average frames of all shots in each movie

Shot Source	Resolution	Num. Shot	Ave. Obj.	Ave. Frame
Mission Impossible	640*268	6	2.7	41
Monsters University	640*360	6	4.2	81
Kung Fu Panda	640*272	6	2.2	96
X-Men	640*266	6	3.3	55
Toy Story	960*540	6	4.3	61

frame. Because the task of our method focuses on proposing the object-like windows, we pay much attention to the frames with objects. It is noted that some keyframes only contain a tiny part of objects, and some of neighboring frames contain no objects. It is impossible to recognize the objects from those frames. Therefore, we re-annotated this dataset in the same way with labeling the proposed dataset, and removed several frames without obvious main objects. To make an overall illustration, we classified this testing set, FBMS-30, based on its categories to 13 classes. Table 2 shows the details of our annotated FBMS-30 dataset.

5.2 Experimental setting

Our approach is implemented using Matlab on a desktop PC with an Intel i5 4590 CPU and 8GB memory. To show the efficiency of the proposed method for eliminating the proposal inconsistency in sequential frames, we compare our method with the state-of-the-art bounding box based object proposals: Edgebox [68], Bing [9], Rahtu [45] and Objectness [2]. Considering the efficiency and fairness, the authors' public source codes with optimized parameters in their papers are adopted in all the experiments. Three popular evaluation metrics are utilized to quantitatively evaluate the performance of the proposed method, the same as [17]. They are the detection rate (DR) with given number of windows (#WIN) (DR-#WIN), DR with variational IoU threshold covered by ground truth annotations for a fixed number of proposals (DR-IOU), and the average detection rate (ADR), i.e., average recall (AR) [24] between 0.5 and 1 by averaging over the overlaps of the images' annotations with the closest matched proposals (ADR-#WIN). Let #GT represent the number of the annotative ground truth for one image, o be the IoU overlap, the DR-#WIN and ADR are separately calculated according to (13) and (14).

$$\text{DR-\#WIN} = \frac{\#(o > \epsilon)@\#\text{WIN}}{\#\text{GT}} \quad \epsilon \in \{x | 0.5 \leq x \leq 1\}, \tag{13}$$

$$ADR = 2 \int_{0.5}^{1} DR(o) \mathrm{d}o, \tag{14}$$

Table 2 Description of FMBS dataset [5, 39]. Num. Shot is the number of shots, Ave. Obj. is the number of average objects, and Ave. Frame is the number of average frames of all shots in each category

Shot Source	Resolution	Num. Shot	Ave. Obj.	Ave. Frame
Camel	680*540	1	1	100
Cars	640*480	4	1.8	35
Cats	640*480,640*360,450*253	3	1.7	197
Dogs	600*400	2	1	310
Farm	720*405	1	2	252
Giraffes	600*400	1	2	320
Goats	500*333	1	3	280
Horses	720*405,480*360,620*349	3	3.7	499
Lion	720*405	1	1	156
Marple	450*350,350*288	6	1.5	303
People	600*338,640*480	3	1.7	83
Rabbits	480*270,680*400,600*338	3	2	226
Tennis	530*380	1	2	466

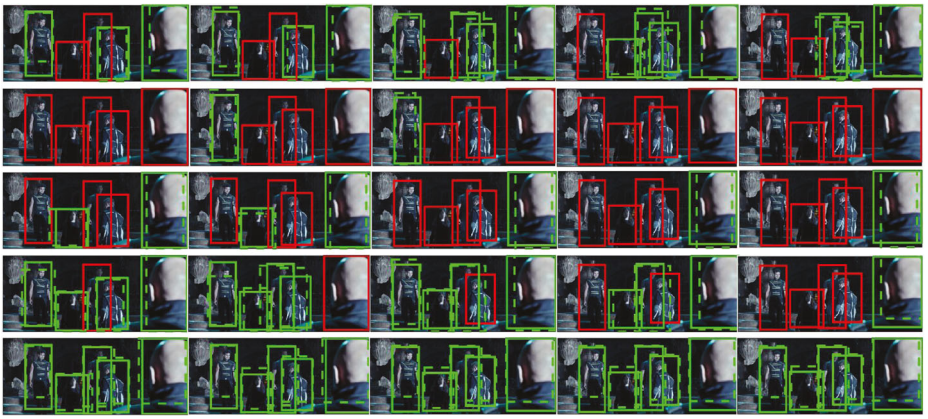

Fig. 3 Comparisons of spatio-temporal bounding box proposals for 5 sequential frames from one shot of our dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. *Green solid rectangles* are the annotated ground truths, while *green dashed rectangles* are the hit proposals. The *red solid rectangles* are the missed ground truths. ($o = 0.7$ and #WIN=1000)

where DR-#WIN is curved by a fixed IoU threshold ϵ between 0.5 and 1 with incremental number of windows, while DR-IOU is plotted based on the different IoU between 0.5 and 1 with a fixed number of windows. And the ADR is calculated according to the different DR on distinct IoU with changing the number of proposals.

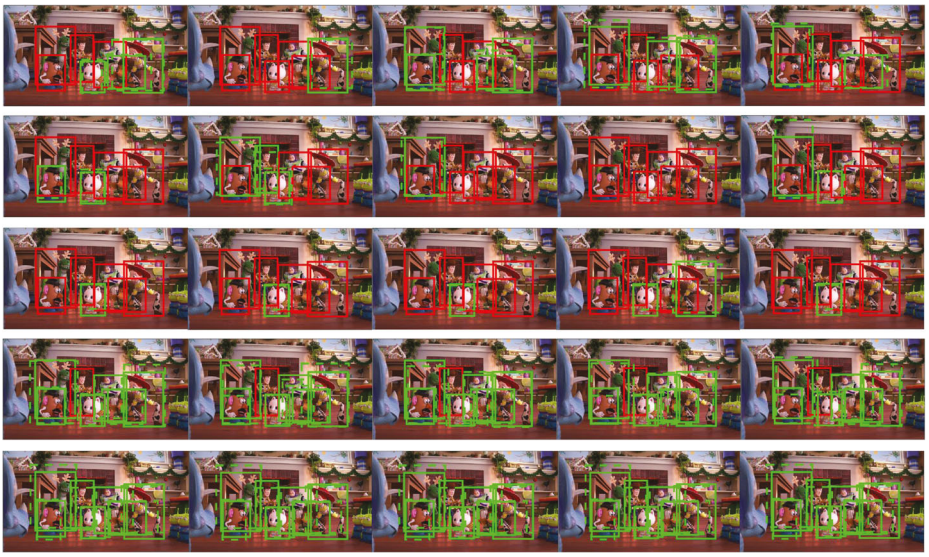

Fig. 4 Comparisons of spatio-temporal bounding box proposals for 5 sequential frames from one shot of our dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. *Green solid rectangles* are the annotated ground truths, while *green dashed rectangles* are the hit proposals. The *red solid rectangles* are the missed ground truths. ($o = 0.7$ and #WIN=1000)

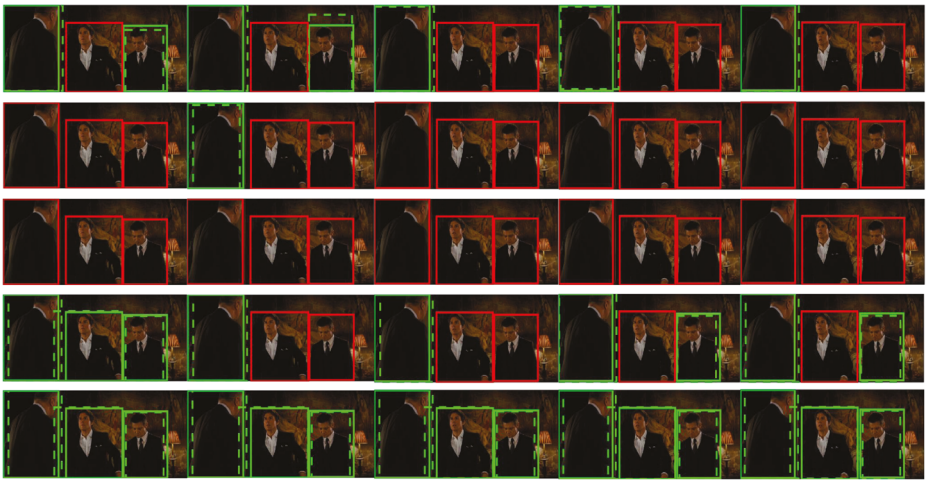

Fig. 5 Comparisons of spatio-temporal bounding box proposals for 5 sequential frames from one shot of our dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. *Green solid rectangles* are the annotated ground truths, while *green dashed rectangles* are the hit proposals. The *red solid rectangles* are the missed ground truths. ($o = 0.8$ and #WIN=1000)

As to parameter settings, they are set as $\{M, \omega, s, Th1, Th2, Th3, \lambda\} = \{10^4, 0.5, 5, 3, 100, 0.01, 0.5\}$. We use $M = 10^4$ as the upper bound for the number of generated bounding boxes. ω and λ are the weight value, we set both of them 0.5. $s \times s$ is the

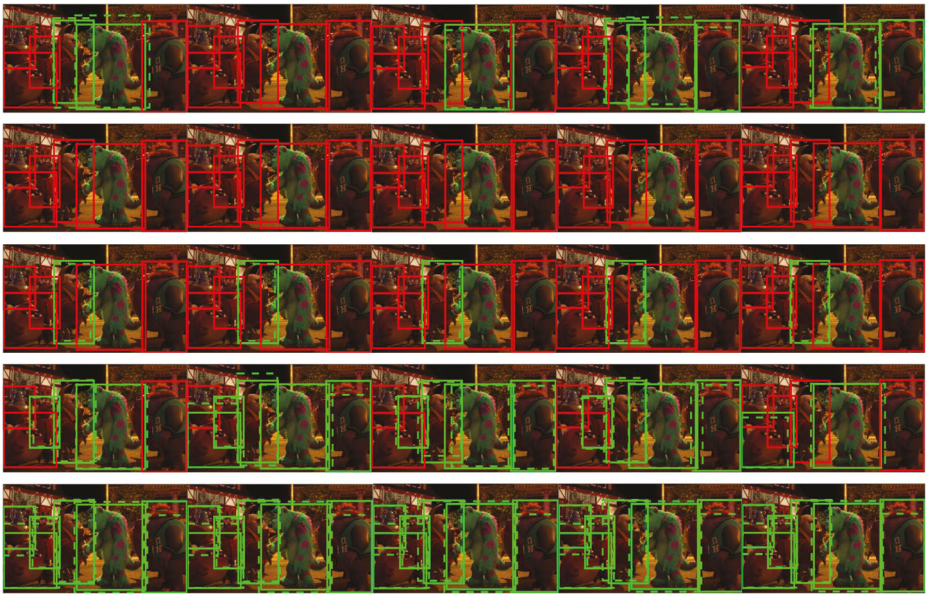

Fig. 6 Comparisons of spatio-temporal bounding box proposals for 5 sequential frames from one shot of our dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. *Green solid rectangles* are the annotated ground truths, while *green dashed rectangles* are the hit proposals. The *red solid rectangles* are the missed ground truths. ($o = 0.8$ and #WIN=1000)

Fig. 7 Comparisons of spatio-temporal bounding box proposals for 7 sequential frames from one shot of FBMS dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. Green solid rectangles are the annotated ground truths, while green dashed rectangles are the hit proposals. The red solid rectangles are the missed ground truths. ($o = 0.7$ and #WIN=1000)

window size that we use to filter the motion fields, and we set $s = 5$. $Th1$ is used to define the similar motion difference in pixel and set as 3. $Th2$ is the area of the bounding box regarded as the lost one, set as 100. $Th3$ is the threshold contributing to determining temporal mapping or not, set as 0.01. Besides, we utilize [49] to calculate the motion fields in

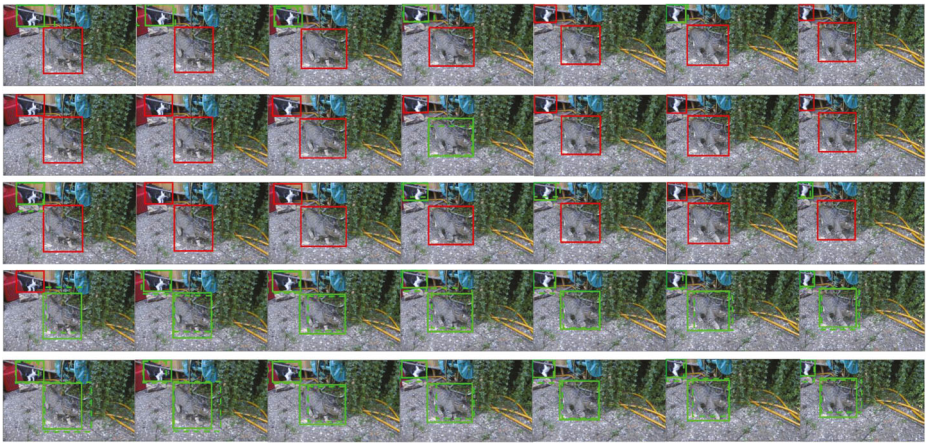

Fig. 8 Comparisons of spatio-temporal bounding box proposals for 7 sequential frames from one shot of FBMS dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. *Green solid rectangles* are the annotated ground truths, while *green dashed rectangles* are the hit proposals. The *red solid rectangles* are the missed ground truths. ($o = 0.7$ and #WIN=1000)

Fig. 9 Comparisons of spatio-temporal bounding box proposals for 7 sequential frames from one shot of FBMS dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. *Green solid rectangles* are the annotated ground truths, while *green dashed rectangles* are the hit proposals. The *red solid rectangles* are the missed ground truths. ($o = 0.8$ and #WIN=1000)

our experiments for its accuracy and efficiency. In fact, any motion field type with adequate accuracy and high efficiency can be utilized in our framework. As to the video's basic feature, it is better to be pre-computed, while it can also be integrated into our model if it is efficient enough.

Fig. 10 Comparisons of spatio-temporal bounding box proposals for 7 sequential frames from one shot of FBMS dataset. These proposals are generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and our method from the first line to the fifth line. *Green solid rectangles* are the annotated ground truths, while *green dashed rectangles* are the hit proposals. The *red solid rectangles* are the missed ground truths. ($o = 0.8$ and #WIN=1000)

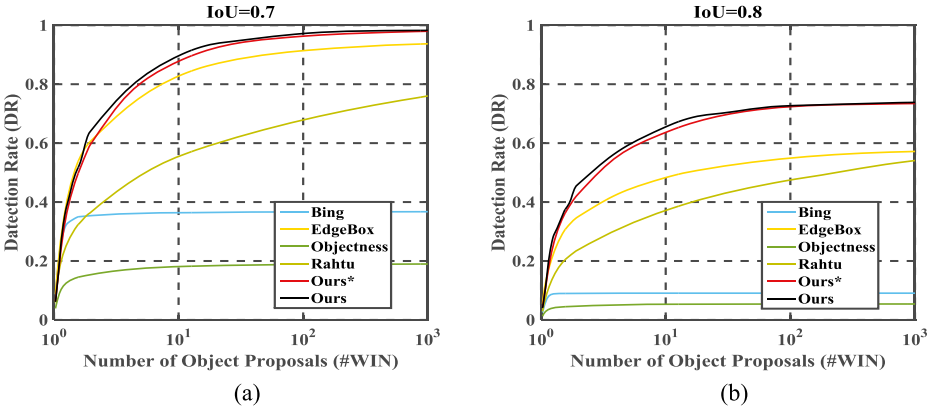

Fig. 11 Detection rate curves of different methods with (**a**) $o = 0.7$ and (**b**) $o = 0.8$ on our dataset

5.3 Comparison

Our method focuses on eliminating the proposal inconsistency when applying object proposals frame by frame, and manages to yield twice the results with half the effort by introducing image object proposals into videos. Besides, we aim at presenting a framework suitable for pre-processing and probably extending to real time applications by improving hardware configurations. Considering that few methods are specially for spatio-temporal bounding box based multi-object proposals in videos, we compare the proposed method with the bounding box based state-of-the-arts [2, 9, 45, 68], according to the survey in [24], performed on the temporal sequences frame by frame. Considering both accuracy and efficiency of the existing object proposals, we recommend the frame-by-frame usage of Edgebox [68] as the baseline in achieving the task of video multi-object proposals.

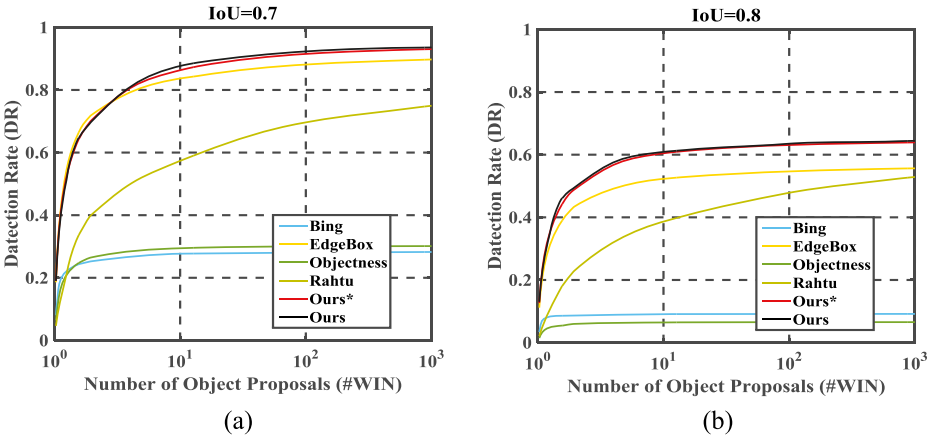

Fig. 12 Detection rate curves of different methods with (**a**) $o = 0.7$ and (**b**) $o = 0.8$ on FBMS dataset

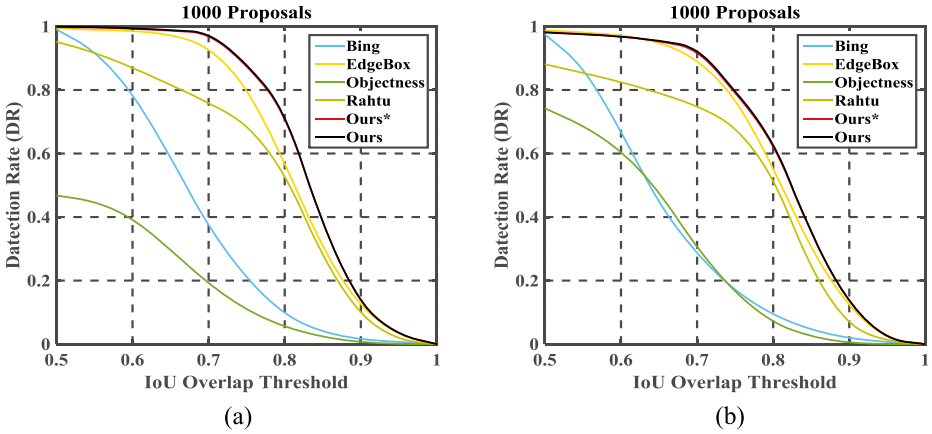

Fig. 13 DR-IOU curves of different methods on (**a**) our dataset and (**b**) FBMS dataset with #WIN=1000

Qualitative evaluation For we evaluate our method on two datasets, we separately exhibit qualitative comparisons with different methods from Figs. 3 to 10. The green solid rectangle is the ground truth, and the green dashed rectangle is the hit proposal. Those red solid rectangles are the missing matched ground truths. In order to show the performance of our method on IoU 0.7 and 0.8, both of which are the accepted intersection over union with the bounding box ground truth in real applications, we present two kinds of comparative results for each dataset. Figures 3 and 4 exhibit the five consecutive proposals for two shots in our dataset when #WIN=1000 and IoU $o = 0.7$ in the proposed dataset, and Figs. 5 and 6 exhibit the five consecutive proposals for two shots in our dataset when #WIN=1000 and IoU $o = 0.8$ in the proposed dataset. Figures 7 and 8 show seven sequential proposal results of two shots in FBMS dataset when #WIN=1000 and IoU $o = 0.7$, and Figs. 9 and 10 show seven sequential proposal results of two shots in FBMS dataset when #WIN=1000

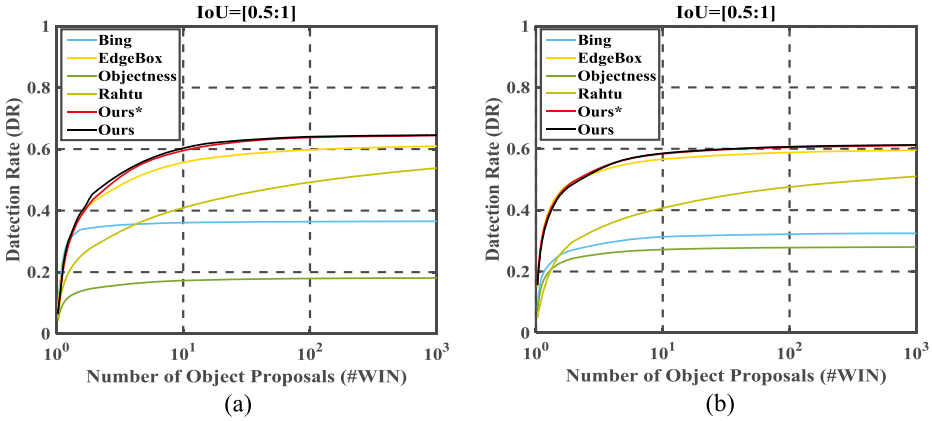

Fig. 14 ADR-#WIN curves of different methods on (**a**) our dataset and (**b**) FBMS dataset with $o \in [0.5, 1]$

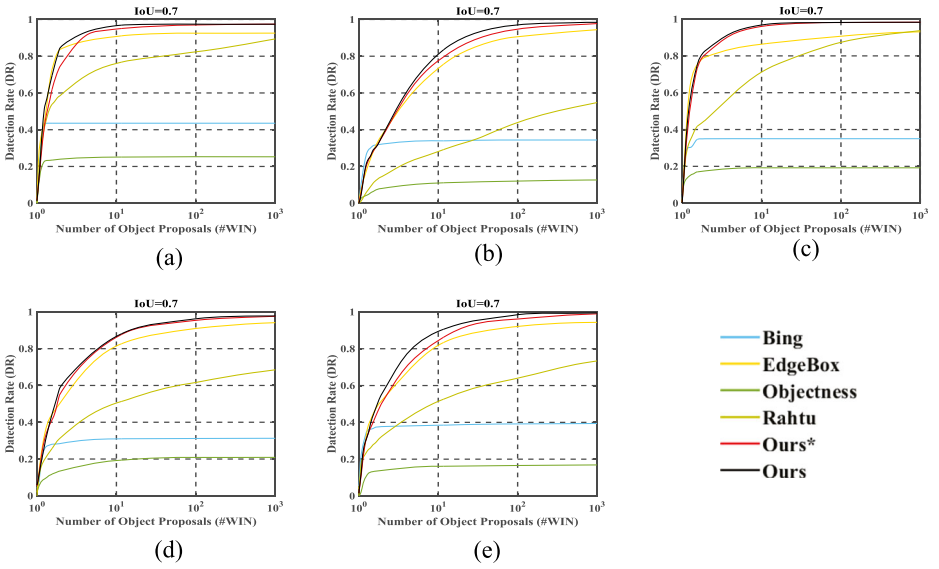

Fig. 15 Detection rate curves of the shots from different movies in our dataset. **a–e** are separately from Mission Impossible, Monsters University, Kung Fu Panda, X-Men and Toy Story with $o = 0.7$

and IoU $o = 0.8$. The spatio-temporal bounding box proposals generated by Rahtu [45], Objectness [2], Bing [9], Edgebox [68] and Ours are placed in rows. Obviously, our method achieves the best performance in eliminating the proposal inconsistency both in IoU 0.7 and 0.8 on different datasets.

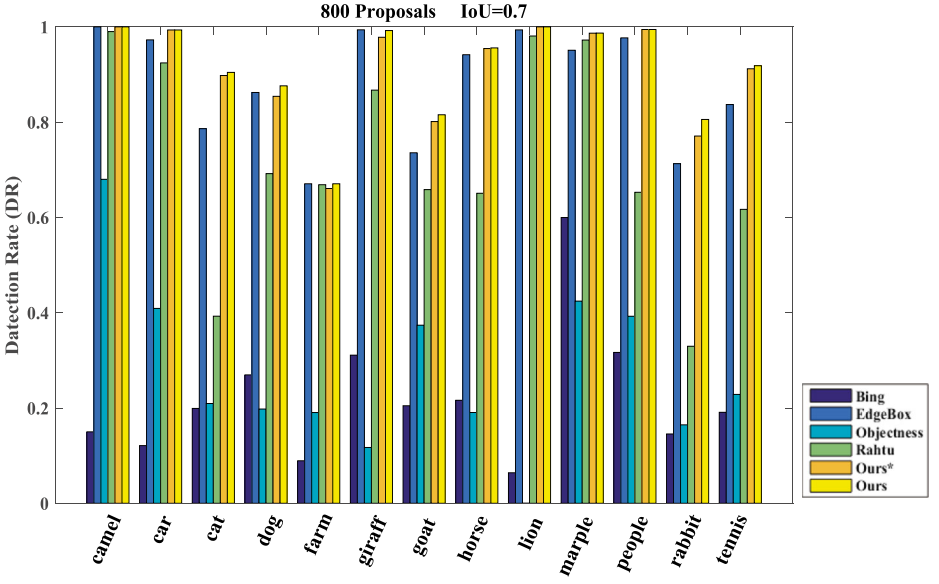

Fig. 16 The comparison of detection rate distribution on different category of FBMS dataset with $o = 0.7$ and #WIN=800

Table 3 Comparison of our method and the frame usage proposal method with different #WIN under $o = 0.7$, $o = 0.8$ and average IoU on our dataset

Method	Type	#WIN=300			#WIN=1000		
		0.7-DR	0.8-DR	ADR	0.7-DR	0.8-DR	ADR
Rahtu [45]	Window	0.54	0.36	0.40	0.76	0.54	0.54
Objectness [2]	Window	0.18	0.05	0.17	0.19	0.05	0.18
Bing [9]	Window	0.36	0.09	0.36	0.37	0.09	0.37
Edgebox [68]	Window	0.81	0.47	0.55	0.94	0.57	0.61
Ours*	Window	0.86	0.62	0.59	0.98	0.73	0.64
Ours	Window	0.88	0.64	0.59	0.98	0.74	0.65

Quantitative evaluation In order to present an overall performance of the proposed method, we make a comprehensive quantitative evaluation by utilizing three popular metrics in object proposals, DR-#WIN, DR-IOU and ADR-#WIN defined in Section 5.2. Figure 11a and b show the detection rate on our dataset with IoU=0.7 and IoU=0.8, and Fig. 12a and b are drawn for FBMS dataset with the same settings as Fig. 11. We give the DR-IOU curve for our dataset and FBMS dataset in Fig. 13a and b. Figure 14a and b illustrate the ADR-#WIN curve for both our dataset and FBMS dataset. For our dataset consists of five different movies, we also give the separate quantitative evaluations on the shots in different movies to show the improvement distributions in Fig. 15. It is shown that the proposed method can achieve the best results on each movie set compared with others. As to FBMS dataset, we give an overall evaluation on different classes classified from FBMS dataset in Fig. 16. The height of the bar represents the detection rate on IoU $= 0.7$ and #WIN=800. It is shown that the proposed method can achieve good performance on different categories, i.e., there is no obvious category bias. To make a further comparison, we also present the detailed comparison of the detection rate between our method and the state-of-the-art on our dataset and FBMS dataset with different proposal numbers under IoU=0.7, IoU=0.8 in Tables 3 and 4.

Running time comparison For our contribution lies in eliminating proposal inconsistency occurring among the temporal sequences, we only compare the running time in generating object proposals. As to the motion field calculation, it has been pre-computed in

Table 4 Comparison of our method and the frame usage proposal method with different #WIN under $o = 0.7$, $o = 0.8$ and average IoU on FBMS dataset

Method	Type	#WIN=300			#WIN=1000		
		0.7-DR	0.8-DR	ADR	0.7-DR	0.8-DR	ADR
Rahtu [45]	Window	0.56	0.37	0.40	0.75	0.53	0.51
Objectness [2]	Window	0.29	0.06	0.27	0.30	0.07	0.28
Bing [9]	Window	0.28	0.09	0.31	0.28	0.09	0.32
Edgebox [68]	Window	0.83	0.52	0.56	0.90	0.56	0.60
Ours*	Window	0.85	0.60	0.58	0.93	0.64	0.61
Ours	Window	0.87	0.61	0.58	0.94	0.64	0.61

Table 5 Average running time comparison on our dataset

	Rahtu[45]	Objectness[2]	Bing[9]	Edgebox[68]	Ours*	Ours
Code	Matlab, C++	Matlab, C++	C++	Matlab, C++	Matlab, C++	Matlab, C++
Time(s)	5.5	4.66	0.03	0.91	0.57	0.58

our framework for motion is the basic feature in videos and many achievements on motion calculations exist. In a word, our model only relies on the calculated motion fields, while not on motion calculation methods. Table 5 shows the comparison of our method and the state-of-art in running time for generating temporal object proposals. Although not efficient as Bing [9], it is shown that the proposed method achieves better performance both in accuracy and efficiency. In addition, our running time is the average computational time across all the resolutions of our dataset.

5.4 Discussion

Our method manages to extend image object proposals to videos, by proposing multi-objects instead of only focusing on dominant objects in videos. It is designed for multi-object proposals compared with those moving object segmentation methods. Besides, the proposed method can eliminate the proposal inconsistency caused by frame-by-frame usage of image object proposals. In general, the proposed method has both advantages and disadvantages. The good characters can be summarized as good performance, category independent and unsupervised, meanwhile we also give the limitations of our method.

Good performance Figure 13 shows that the proposed method has low IoU drop. Our method achieves good results on both IoU 0.7 and 0.8, while some state-of-the-art methods

Fig. 17 More results about spatio-temporal bounding box proposals generated by the proposed method for our dataset. The *green solid rectangle* is the annotated ground truth and the *green dashed rectangle* is the hit proposal. ($o = 0.8$ and #WIN=1000)

Fig. 18 More results about spatio-temporal bounding box proposals generated by the proposed method for FBMS dataset. The *green solid rectangle* is the annotated ground truth and the *green dashed rectangle* is the hit proposal. ($o = 0.8$ and #WIN=1000)

can only achieve the improvement on a fixed IoU or a lower IoU value. Considering the requirements from real applications, IoU 0.7 and 0.8 are sufficient enough to leverage accuracy and practicability. Figures 17 and 18 illustrate more results on some temporal frames in our dataset and FBMS dataset. Although without any complicated calculation, our method can localize the object as much as possible.

Category independent From the experimental dataset perspective, there are many kinds of objects including regular and irregular shapes. Figures 15 and 16 show the detection rate on different movies and categories. It could be seen that there is no obvious category bias. No matter the object is real or imaginary, the proposed method can make a further improvement, though there are differences in the improved values. Therefore, our method is category independent which is suitable for applying to practical applications.

Unsupervised method There is no learning stage and no category tendentiousness in our method. Therefore, it is an unsupervised method independent of datasets. Because of no prerequisite, our method is more propitious to be utilized as a pre-processing procedure.

Limitations Our method is an extension of image object proposals in videos as a pre-processing procedure. Therefore, some defects in image method may be inherited. But this issue can be fixed with the boosting of image object proposals. Furthermore, because of motion blur and the inaccurate motion field, bounding box proposals cannot be accurately mapped for every frame. If most bounding boxes cannot be temporally or accurately mapped, our method may degrade to frame-by-frame usage. That's why we did not achieve

significant improvements on every shot. Fortunately, with the advance of motion field estimation, the problem will be hopefully solved in the near future.

6 Conclusion

An adaptive context-aware model is proposed for video object proposals in this paper. It aims at eliminating proposal inconsistency when applying image methods frame by frame, while taking advantages of both image methods and video features. By introducing the proposed context-aware model, image object proposals can be successfully migrated into video processing, yielding twice the results with half the effort. To evaluate the efficiency of proposed video multi-object proposals, we build a specific multi-object dataset with bounding box based ground truths annotated frame by frame, and we also annotate one public dataset in the same way. Experiments on these challenge datasets demonstrate that the proposed approach outperforms the performance by utilizing the state-of-the-art method on the single frame in sequences.

Our future work will focus on the refinement strategy of our method to make a further improvement on the object detection rate rather than only refining on the ranking temporal scores. We will also manage to explore the parameter optimization of our model to provide more targeted parameter settings.

Acknowledgements 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. Alexe B, Deselaers T, Ferrari V (2010) What is an object? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 73–80. IEEE
2. Alexe B, Deselaers T, Ferrari V (2012) Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11):2189–2202
3. Arbeláez P., Pont-Tuset J, Barron JT, Marques F, Malik J (2014) Multiscale combinatorial grouping. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 328–335. IEEE
4. Bai T, Li YF, Zhou X (2015) Learning local appearances with sparse representation for robust and fast visual tracking. IEEE Transactions on Cybernetics 45(4):663–675
5. Brox T, Malik J (2010) Object segmentation by long term analysis of point trajectories. In: Proceedings of the european conference on computer vision, pp 282–295. Springer
6. Caba Heilbron F, Carlos Niebles J, Ghanem B (2016) Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1914–1923. IEEE
7. Chavali N, Agrawal H, Mahendru A, Batra D (2016) Object-proposal evaluation protocol is' gameable'. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–10. IEEE
8. Chen X, Ma H, Wang X, Zhao Z (2015) Improving object proposals with multi-thresholding straddling expansion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2587–2595. IEEE
9. Cheng MM, Zhang Z, Lin WY, Torr P (2014) Bing: Binarized normed gradients for objectness estimation at 300fps. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3286–3293. IEEE
10. Cheng Z, Li X, Shen J, Hauptmann AG (2016) Which information sources are more effective and reliable in video search. In: Proceedings of the international conference on research on development in information retrieval, pp 1069–1072. ACM

11. Cheng Z, Shen J (2016) On very large scale test collection for landmark image search benchmarking. Signal Process 124:13–26
12. Cheng Z, Shen J, Miao H (2016) The effects of multiple query evidences on social image retrieval. Multimedia Systems 22(4):509–523
13. Choi MK, Wang Z, Lee HG, Lee SC (2016) A bag-of-regions representation for video classification. Multimedia Tools and Applications 75(5):2453–2472
14. Chu WT, Yu CH, Wang HH (2015) Optimized comics-based storytelling for temporal image sequences. IEEE Transactions on Multimedia 17(2):201–215
15. Endres I, Hoiem D (2010) Category independent object proposals. In: Proceedings of the european conference on computer vision, pp 575–588. Springer
16. Endres I, Hoiem D (2014) Category-independent object proposals with diverse ranking. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(2):222–234
17. Geng W, Li S, Ren T, Wu G (2016) Object proposals using svm-based integrated model. In: Proceedings of the international joint conference on neural networks, pp 4154–4161. IEEE
18. Geng W, Wu G (2016) Context-aware video object proposals. In: Proceedings of the IEEE conference on parallel and distributed systems, pp 1203–1206. IEEE
19. Ghodrati A, Diba A, Pedersoli M, Tuytelaars T, Van Gool L (2015) Deepproposal: Hunting objects by cascading deep convolutional layers. In: Proceedings of the IEEE international conference on computer vision, pp 2578–2586. IEEE
20. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587. IEEE
21. Girshick R, Donahue J, Darrell T, Malik J (2016) Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(1):142–158
22. Gygli M, Grabner H, Van Gool L (2015) Video summarization by learning submodular mixtures of objectives. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3090–3098. IEEE
23. Hayder Z, He X, Salzmann M (2016) Learning to co-generate object proposals with a deep structured network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–10. IEEE
24. Hosang J, Benenson R, Dollár P., Schiele B (2016) What makes for effective detection proposals? IEEE Transactions on Pattern Analysis and Machine Intelligence 38(4):814–830
25. Hu JF, Zheng WS, Ma L, Wang G, Lai J (2016) Real-time rgb-d activity prediction by soft regression. In: Proceedings of the european conference on computer vision, pp 280–296. Springer
26. Hua Y, Alahari K, Schmid C (2015) Online object tracking with proposal selection. In: Proceedings of the IEEE international conference on computer vision, pp 3092–3100. IEEE
27. Jain M, Van Gemert J, Jégou H, Bouthemy P, Snoek CG (2014) Action localization with tubelets from motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 740–747. IEEE
28. Jang WD, Lee C, Kim CS (2016) Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 696–704. IEEE
29. Kong T, Yao A, Chen Y, Sun F (2016) Hypernet: Towards accurate region proposal generation and joint object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–10. IEEE
30. Kuo W, Hariharan B, Malik J (2015) Deepbox: Learning objectness with convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 2479–2487. IEEE
31. Li Y, Lu H, Li J, Li X, Li Y, Serikawa S (2016) Underwater image de-scattering and classification by deep neural network. Comput Electr Eng 54:68–77
32. Liu J, Ren T, Bao BK, Bei J (2016) Depth-aware layered edge for object proposal. In: Proceedings of the IEEE international conference on multimedia and expo, pp 1–6. IEEE
33. Liu J, Ren T, Bei J (2016) Elastic edge boxes for object proposal on rgb-d images. In: Proceedings of the international conference on multimedia modeling, pp 199–211. Springer
34. Liu J, Ren T, Wang Y, Zhong SH, Bei J, Chen S (2016) Object proposal on rgb-d images via elastic edge boxes. Neurocomputing 236:134–146
35. Liu Y, Mei T, Chen CW (2016) Automatic suggestion of presentation image for storytelling. In: Proceedings of the IEEE international conference on multimedia and expo, pp 1–6. IEEE
36. Lowry S, Sünderhauf N, Newman P, Leonard JJ, Cox D, Corke P, Milford MJ (2016) Visual place recognition: a survey. IEEE Trans Robot 32(1):1–19

37. Manen S, Guillaumin M, Van Gool L (2013) Prime object proposals with randomized prim's algorithm. In: Proceedings of the IEEE international conference on computer vision, pp 2536–2543. IEEE
38. Meng J, Wang H, Yuan J, Tan YP (2016) From keyframes to key objects: video summarization by representative object proposal selection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1039–1048. IEEE
39. Ochs P, Malik J, Brox T (2014) Segmentation of moving objects by long term video analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(6):1187–1200
40. Oneata D, Revaud J, Verbeek J, Schmid C (2014) Spatio-temporal object detection proposals. In: Proceedings of the european conference on computer vision, pp 737–752. Springer
41. Papazoglou A, Ferrari V (2013) Fast object segmentation in unconstrained video. In: Proceedings of the IEEE international conference on computer vision, pp 1777–1784. IEEE
42. Perazzi F, Wang O, Gross M, Sorkine-Hornung A (2015) Fully connected object proposals for video segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 3227–3234. IEEE
43. Pont-Tuset J, Marques F (2016) Supervised evaluation of image segmentation and object proposal techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(7):1465–1478
44. Pont-Tuset J, Van Gool L (2015) Boosting object proposals: From pascal to coco. In: Proceedings of the IEEE international conference on computer vision, pp 1546–1554. IEEE
45. Rahtu E, Kannala J, Blaschko M (2011) Learning a category independent object detection cascade. In: Proceedings of the IEEE international conference on computer vision, pp 1052–1059. IEEE
46. Rantalankila P, Kannala J, Rahtu E (2014) Generating object segmentation proposals using global and local search. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2417–2424. IEEE
47. Savelonas MA, Pratikakis I, Sfikas K (2015) An overview of partial 3d object retrieval methodologies. Multimedia Tools and Applications 74(24):11,783–11,808
48. Sharir G, Tuytelaars T (2012) Video object proposals. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops, pp 9–14. IEEE
49. Sun D, Roth S, Black MJ (2014) A quantitative analysis of current practices in optical flow estimation and the principles behind them. Int J Comput Vis 106(2):115–137
50. Sunderhauf N, Shirazi S, Jacobson A, Dayoub F, Pepperell E, Upcroft B, Milford M (2015) Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. Proceedings of Robotics: Science and Systems XII
51. Uijlings JR, van de Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
52. Van de Sande KE, Uijlings JR, Gevers T, Smeulders AW (2011) Segmentation as selective search for object recognition. In: Proceedings of the IEEE international conference on computer vision, pp 1879–1886. IEEE
53. Van den Bergh M, Roig G, Boix X, Manen S, Van Gool L (2013) Online video seeds for temporal window objectness. In: Proceedings of the IEEE international conference on computer vision, pp 377–384. IEEE
54. Wang W, Shen J, Porikli F (2015) Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3395–3402. IEEE
55. Wang W, Shen J, Shao L (2015) Consistent video saliency using local gradient flow optimization and global refinement. IEEE Trans Image Process 24(11):4185–4196
56. Xiao F, Lee YJ (2016) Track and segment: an iterative unsupervised approach for video object proposals. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–10. IEEE
57. Xu X, Ge L, Ren T, Wu G (2015) Adaptive integration of depth and color for objectness estimation. In: Proceedings of the IEEE international conference on multimedia and expo, pp 1–6. IEEE
58. Xu X, Geng W, Ju R, Yang Y, Ren T, Wu G (2014) Obsir: Object-based stereo image retrieval. In: Proceedings of the IEEE international conference on multimedia and expo, pp 1–6. IEEE
59. Yang G, Zhang Y, Yang J, Ji G, Dong Z, Wang S, Feng C, Wang Q (2015) Automated classification of brain images using wavelet-energy and biogeography-based optimization. Multimedia Tools and Applications 75(23):15,601–15,617
60. Zhang D, Javed O, Shah M (2013) Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 628–635. IEEE
61. Zhang H, Shang X, Luan H, Wang M, Chua TS (2016) Learning from collective intelligence: Feature learning using social images and tags. ACM Trans Multimed Comput Commun Appl 13(1):1–23
62. Zhang H, Shang X, Yang W, Xu H, Luan H, Chua TS (2016) Online collaborative learning for open-vocabulary visual classifiers. In: Proceedings of the IEEE international conference on computer vision and pattern recognition, pp 2809–2817. ACM

63. Zhang H, Zha ZJ, Yang Y, Yan S, Gao Y, Chua TS (2013) Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In: Proceedings of the ACM international conference on multimedia, pp 33–42. ACM
64. Zhang J, Sclaroff S, Lin Z, Shen X, Price B, Mech R (2016) Unconstrained salient object detection via proposal subset optimization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–10. IEEE
65. Zhang Y, Phillips P, Wang S, Ji G, Yang J, Wu J (2016) Fruit classification by biogeography-based optimization and feedforward neural network. Expert Systems 33(3):239–253
66. Zhou Y, Ni B, Hong R, Wang M, Tian Q (2015) Interaction part mining: a mid-level approach for fine-grained action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3323–3331. IEEE
67. Zhu G, Porikli F, Li H (2016) Robust visual tracking with deep convolutional neural network based object proposals on pets. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 26–33. IEEE
68. Zitnick CL, Dollár P. (2014) Edge boxes: Locating object proposals from edges. In: Proceedings of the european conference on computer vision, pp 391–405. Springer

Wenjing Geng is currently a Ph.D. student of Department of Computer Science and Technology at Nanjing University. Her current research interests include image/video content analysis and stereo media processing.

Chunlong Zhang is currently a master student of Department of Computer Science and Technology at Nanjing University. His current research interests include object proposals and image analysis.

Gangshan Wu received his Ph.D. degree from Nanjing University in 2000. He is currently a Professor of Department of Computer Science and Technology at Nanjing University. His current research interests include stereo image/video analysis and information retrieval.