Multimed Tools Appl (2017) 76:12941–12964
DOI 10.1007/s11042-016-3709-3

Unsupervised video co-segmentation based on superpixel co-saliency and region merging

Guoheng Huang[1] · **Chi-Man Pun**[1] · **Cong Lin**[1]

Received: 18 December 2015 / Revised: 4 May 2016 / Accepted: 24 June 2016 /
Published online: 5 July 2016

Abstract Nowadays, fully unsupervised video object segmentation is still a challenge in computer vision. Furthermore, it is more difficult to segment the object from a set of clips. In this paper, we propose an unsupervised and on-line method that efficiently segments common objects from a set of video clips. Our approach is based on the hypothesis, that common or similar objects in multiple video clips are salient, and they share similar features. At first, we try to find out the regions in every clip which are salient and share similar features by proposing a new co-saliency scheme based on superpixels. Then, the most salient superpixels are chosen as the initial object marker superpixels. Starting from these superpixels, we merge neighboring and similar regions, and segment out the final object parts. The experimental results demonstrate that the proposed method can efficiently segment the common objects from a group of video clips with generally lower error rate than some state-of-the-art video co-segmentation methods.

Keywords Co-saliency · Co-segmentation · Region merging · Superpixel · Unsupervised

1 Introduction

Video object segmentation is a fundamental task in multimedia, and it can benefit many applications, e.g., surveillance and video editing. Provided with some priors to indicate what or which the object regions are, some supervised or interactive object extraction and tracking approaches can achieve good performance [5, 6, 33]. However, they are hard to be extended:

✉ Chi-Man Pun
cmpun@umac.mo

Guoheng Huang
yb27405@umac.mo

Cong Lin
yb17403@umac.mo

[1] Department of Computer and Information Science, University of Macau, Macau, SAR, China

first, the users need to provide human interaction or manually label the training data as a matter of experience; second, it is hard to operate for user to mark the regions of interest on a set of video clips. Therefore, unsupervised video object segmentation in a set of clips is a challenging problem, and it has much applications in large scale video tagging and retrieval, generation of training sets for supervised learning, and forensic video analysis [37]. Among them, video object co-segmentation is one of feasible solutions and research highlights.

Video object co-segmentation is an extension of image object co-segmentation, image/video co-saliency and video co-tracking/co-detection. The aim of video object co-segmentation is to unsupervised or semi-supervised extract the common object regions which share similar feature simultaneously on a set of video clips. Liu and Zhang etc. propose a hierarchical segmentation based co-saliency model [25]. Li and Ngan present a method to detect co-saliency from an image pair which is modeled as a linear combination of the single-image saliency map (SISM) and the multi-image saliency map (MISM) [24]. Cao and Tao et al. propose a saliency map fusion framework, which exploits the relationship of different saliency cues and obtains a self-adaptive weight to generate the final co-saliency map [8]. Fu, Cao and Tu employ clustering to represent the global correspondence relationship among a set of images, and generate final co-saliency maps by fusing three effective bottom-up cues [14].

There are many prior trials which are successful on image co-segmentation. Among them, an image co-segmentation model proposed by Hochbaum and Singh which enables obtaining efficient solutions to the underlying optimization model using a maximum flow procedure on an appropriately constructed graph [16]. Joulin, Bach and Ponce propose an image co-segmentation which combines existing tools for underlying image segmentation, with kernel methods commonly used in object recognition [20]. Batra, Kowdle and Parikh develop an algorithm for interactive co-segmentation of a foreground object from a group of related images which sharing common objects [7]. This method only needs little user interaction and can extract the common objects from plentiful images.

Video co-tracking/co-detection is one of the most popular simplifications for video co-segmentation. Tang and Joulin etc. present a joint image-box formulation which can be relaxed to a convex quadratic program for solving the co-localization problem [29]. Tang and Brennan etc. propose an on-line semi-supervised learning framework that the object is represented by independent features and support vector machine (SVM) is built for each feature [13]. Zhang et al. propose a novel approach to extract primary object segments in videos in the ‘object proposal’ domain which can be applied to video co-segmentation [36].

Recently, more and more methods have also been proposed for video co-segmentation. Fu and Xu etc. present a video object co-segmentation method based on category-independent object proposals which is able to extract multiple foreground objects in a video set [18]. Joulin, Bach and Ponce propose a novel energy-minimization approach to co-segmentation that can handle multiple objects [21]. Meng and Li etc. propose a new model considering the co-segmentation problem as the shortest path problem and use the dynamic programming method to solve it [27]. The above two methods are proposed for image co-segmentation, but they both can be also applied to video co-segmentation. Guo and Li etc. propose the trajectory co-saliency measure, which captures the notion that trajectories recurring in all the videos should have their mutual saliency boosted [19]. Chiu and Fritz propose to study video co-segmentation where the number of objects is unknown by formulating a non-parametric Bayesian model [10]. Zhang, Javed and Shah propose a novel approach for object co-segmentation in arbitrary videos by sampling, tracking and matching object proposals via a Regulated Maximum Weight Clique (RMWC) extraction scheme [37].

The existing video co-segmentation algorithms cannot segment out the accurate enough boundaries of the common objects from a set of video clips, and most of them are not on-line algorithm. Therefore, the research interest of this paper is to propose a full automatic and on-line video co-segmentation that can accurately extract the common objects from a set of videos. The main idea of our scheme is to consider the regions where are salient (it is different from most parts of the whole image) and share common features (color, texture etc.) with objects in the other video clips (it may be defined as "co-salient regions") as the initial objet marker regions, meanwhile consider the non-salient regions as background marker regions. And then, merge the non-marker regions to the initial object maker regions or background marker regions to classify all regions as object or background. Especially, the strategies to construct a co-saliency map are different in the first frame and other frames. The overview of the whole procedure of the proposed algorithm is shown by Fig. 1. The main contribution can be summarized as follows:

First, we propose a fully automatic (without any user interaction) initialization stage based on co-saliency for the first frame in each video clip. Especially, we propose a novel co-saliency model on superpixel level to measure each region if it is co-salient or not.

Second, for the purpose to describe the color and texture feature of each superpixel or region, we propose a novel region feature based on hierarchical histogram.

Third, we first present an on-line video co-segmentation based on superpixel trajectory. Similar to the initialization stage, it includes co-saliency, marker prediction and region merging. Differ from the initialization stage, we also consider the superpixel motion trajectory in video co-segmentation.

The structure of this paper is as follows. In section 1, we have offered an introduction to related works. In section 2, the proposed initialization based on superpixel co-saliency is

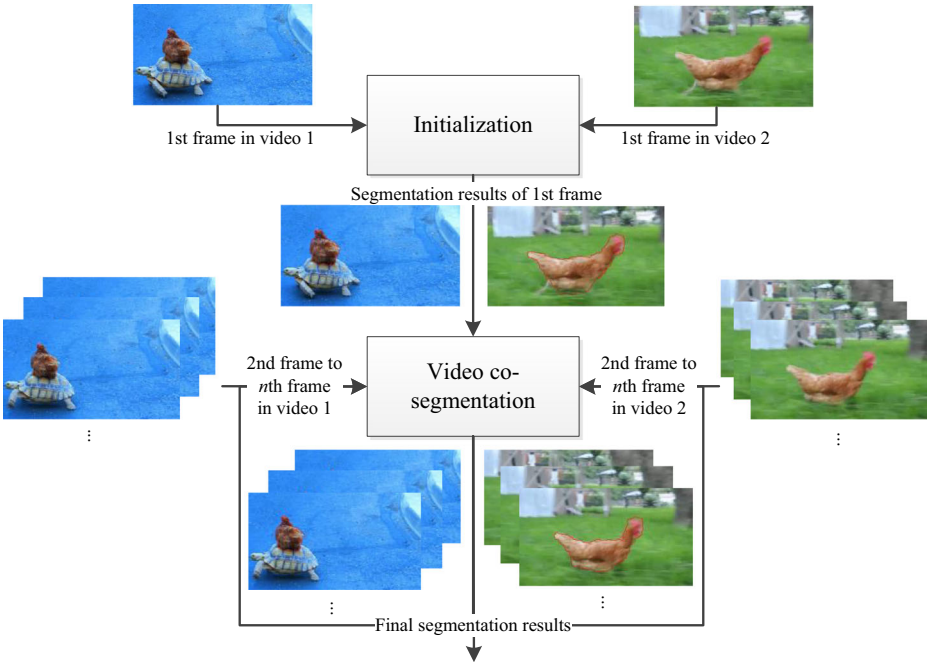

Fig. 1 The overview of the proposed scheme

explained. In section 3, the image co-segmentation based on marker prediction and region merging is presented. In section 4, the proposed video co-segmentation algorithm is explained in detail. In section 5, the experimental results are evaluated based on visual and quantitative comparisons. In section 6, our conclusions and future suggestions are presented.

2 Initialization based on superpixel co-saliency

Co-saliency is a soft problem to co-segmentation. The co-saliency map can provide an initial highlight of potential objects, which can replace the user interaction in segmentation [14]. In our paper, the main idea of the proposed scheme is to merge non-marker superpixels to the initial object marker superpixels (which are salient and share features in each video clip) and the initial background marker superpixels respectively. In this section, we propose an initialization based on superpixel co-saliency which actually is a fully automatic image co-segmentation to initialize the co-segmentation from the first frame, and more details are shown in Fig. 2. In order to reduce the computation and generate the initial segmentation, we over segment the video clips into desired number of superpixels at first. Recently there are many proposed methods which can divide images or video frames into superpixels [3, 11, 30–32]. In our proposed method, we have adopted the SLIC (Simple Linear Iterative Clustering) as the over segmentation method, as the superpixels generated by SLIC are likely equal size and adhere accurately to the boundaries of objects [3]. After superpixels are generated, both the color and texture feature of each superpixel or region will be described by a novel region feature based on hierarchical histogram, which is constructed by multi-layer color histograms.

Then we need to find out the superpixels from the common objects in each video clips. As the definition of co-saliency, the basic assumptions is that a region in an image is co-salient when and only when this region is salient in its image, and it is similar to the other regions which are salient in other images [24]. On one hand, the most co-salient regions should be considered as the initial marker regions if the scheme can describe the co-saliency of a set of images accurately. On the other hand, the least co-salient regions should be considered as the initial background marker regions. Therefore, we try to propose a reliable co-saliency model which includes two stages – superpixel intra saliency and superpixel inter saliency. At first, the superpixel intra saliency is extracted to describe which superpixels are most different from the most of superpixels in a frame. And then, the superpixel inter saliency based on object proposal is calculated to find out the superpixels which are common in a set of video clips. Final, by combining the superpixel intra saliency map and the superpixel inter saliency, we obtain the co-saliency map (on superpixel level). The procedure of the proposed is shown in Fig. 3.

2.1 Region feature extraction based on hierarchical histogram

After the over segmentation stage, we apply a robust region feature based on hierarchical histogram to each superpixel, which allows us to successfully separate the object of interest from the background and compute the similarity in inter saliency stage. Normally, color histogram is popular to apply to region feature representation [28]. However, it is not accurate enough to measure region feature which have texture information with different colors in a region. Therefore, we propose a region feature based on hierarchical histogram which computes different levels of color histogram, to capture not only the color information of each superpixel but also texture information.

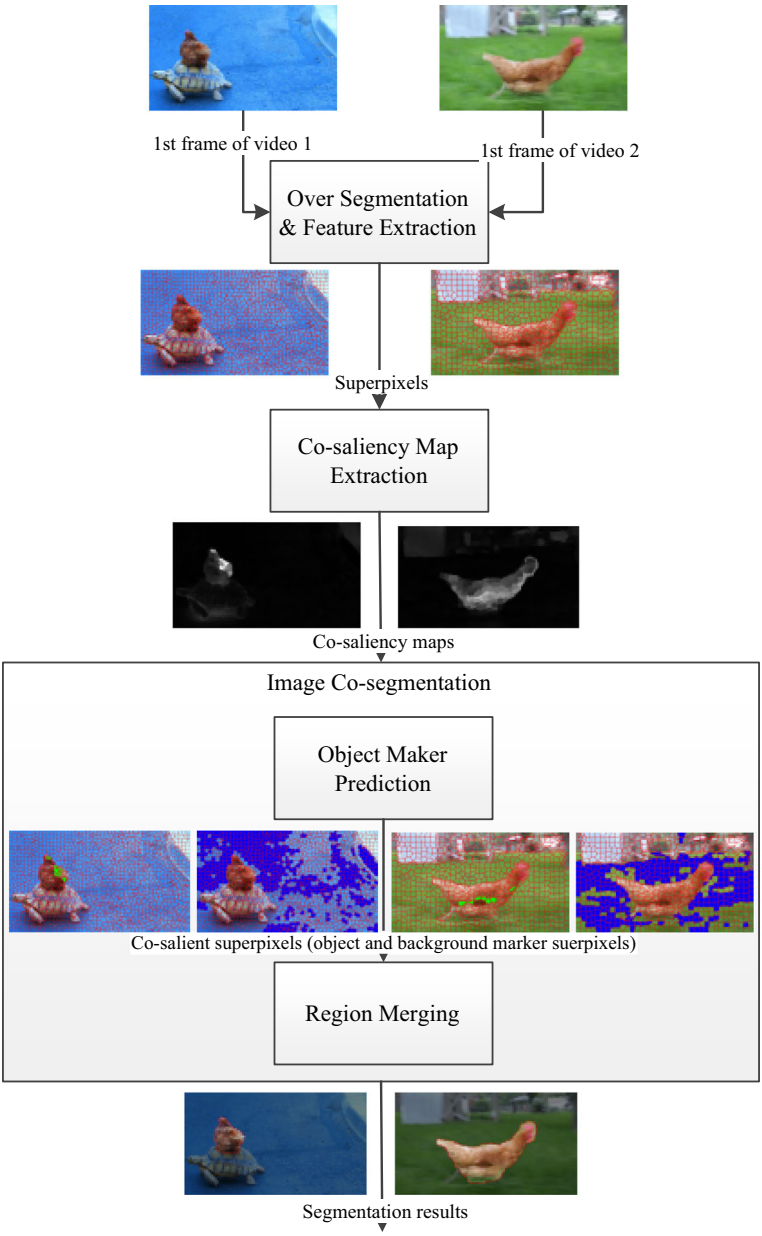

Fig. 2 Initialization and image co-segmentation

In general, the texture feature is constructed by different colors in the same region. We need to extract the color feature in sub-regions in this region. In order to obtain the sub-regions within one region, the basic morphological image erosion operation is a feasible solution. After over segmentation, we erode the boundary of each superpixel/region (represented by $si, i = 1, 2, \ldots k$) m times to obtain m layers of sub-regions of every region (represented by $sij, j = 1, 2, \ldots m$, where $si1 = si$). In this paper, we employ the basic morphological image

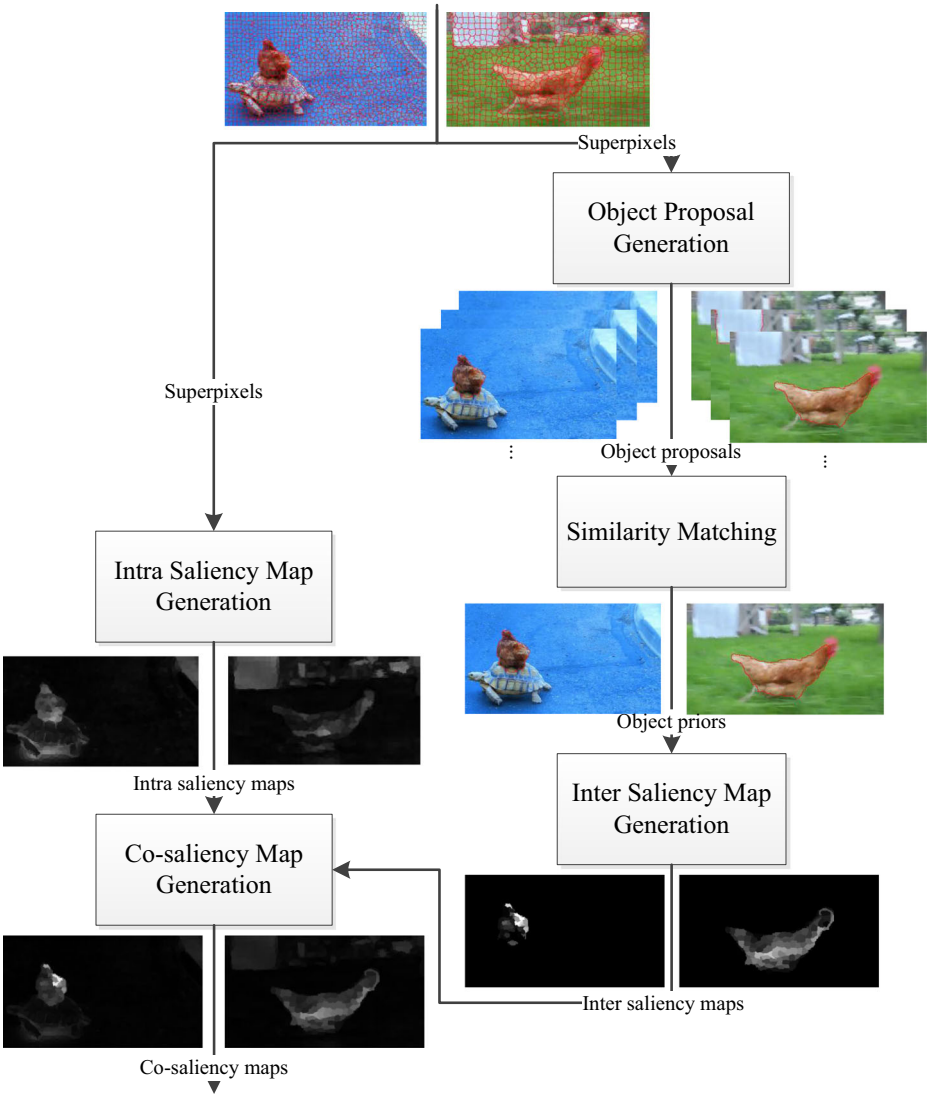

Fig. 3 Co-saliency on superpixel level

erosion operation with circular structure element (with radius 15). For the sub-regions of each region, we then uniformly quantize each color channel into 16 levels, and then, we calculate the histogram of each sub-region in the feature space, yielding $16 \times 16 \times 16 = 4096$ bins. Let Histsij denote the normalized histogram of the jth-layer sub-region of region si. Hence, the proposed region feature of region si is defined as follows:

$$Fsi = \frac{\left(\sqrt{\text{Hist}si1^1}, \ldots, \sqrt{\text{Hist}si1^{4096}}, \ldots, \sqrt{\text{Hist}sim^1} \ldots, \sqrt{\text{Hist}sim^{4096}}\right)}{norm\left(\left(\sqrt{\text{Hist}si1^1}, \ldots, \sqrt{\text{Hist}si1^{4096}}, \ldots, \sqrt{\text{Hist}sim^1} \ldots, \sqrt{\text{Hist}sim^{4096}}\right)\right)} \quad (1)$$

where Histsij^t is the normalized histogram value for bin t of sij. In addition, *norm* denotes the normalized form of a vector. If there is texture information within region si, it may be extracted by the normal histogram of its sub-regions $sij, j = 1, 2, \ldots m$.

An example of the extraction of the proposed region feature based on hierarchical histogram is shown in Fig. 4.

From the Fig. 4, we can see superpixel A and B are similar measured by traditional color histogram (Fig. 4a and d), because these two superpixels are constructed by similar colors. However, actually they are two different textures. Obviously, they are very different in the second and third layer histograms (Fig. 4 (b), (c) and (e), (f)). Therefore, it shows that our hierarchical histogram based region feature can effectively describe the texture and color information within a region.

Then, we use the inner product of the proposed features of two regions si and sj to quantify the similarity between them [22]:

$$\rho(si, sj) = F_{\mathrm{si}} \cdot F_{\mathrm{sj}} \tag{2}$$

Two regions are similar if and only if the proposed similarity is close to 1.

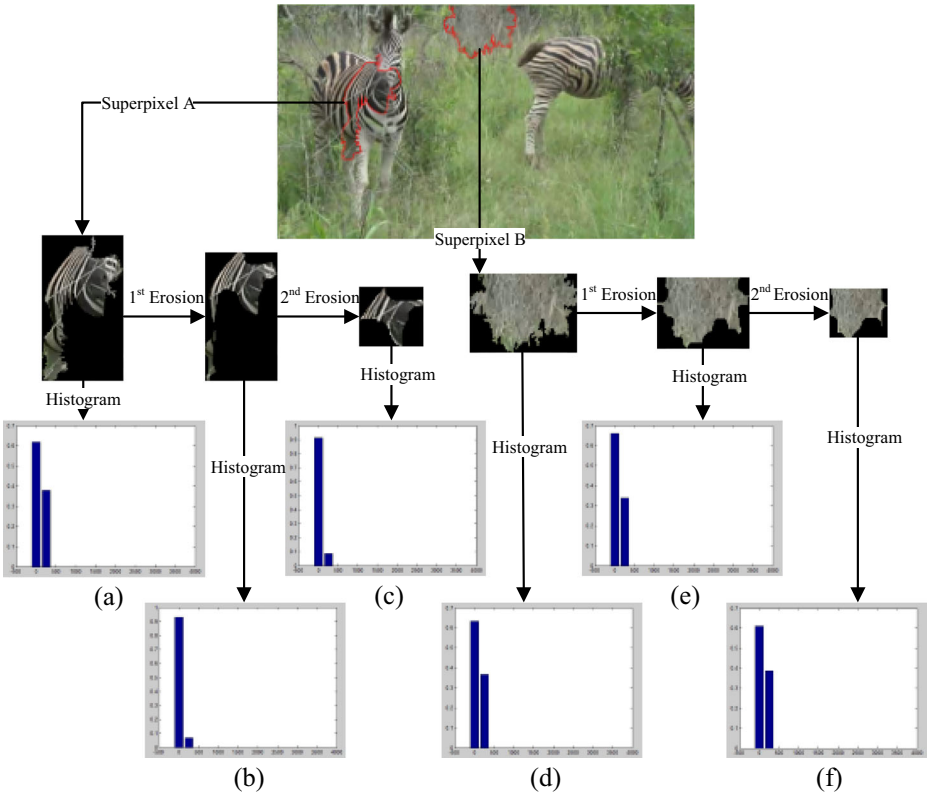

Fig. 4 The proposed region feature based on hierarchical histogram: (**a**) – (**b**) are 3-layer histograms of superpixel A respectively; and (**b**) – (**d**), are 3-layer histograms of superpixel B respectively

2.2 Co-saliency map generation

As mentioned, we assume that the common objects are some regions which are salient and share common region feature. In this paper, we segment out the final common objects through merging the non-marker superpixels to the initial object marker regions. In the co-segmentation case, the regions which share common features and are salient can be considered as the initial object marker regions. In our scheme, we present a new co-saliency method on superpixel level to find out the initial object marker superpixels (for region merging) in each video clip. The proposed co-saliency model includes two stages – superpixel intra saliency and superpixel inter saliency.

2.2.1 Intra saliency map generation

Intra saliency for a single image is a very well developed research area [1, 2, 9, 17, 35]. Among them, Hou and Zhang propose a saliency map by extracting the spectral residual of an image, and they regard the difference between the log amplitude spectrum and the average log amplitude spectrum as the salient parts [17]. Our intra saliency strategy is on superpixel level, and it includes two parts. On one hand, we extend this saliency method [17] to superpixel level as the first part of our intra saliency map, because it can detect the saliency regions with high frequency:

First of all, transfer the current frame to CIELab space. Then, we suppose that *sij* is one of superpixels in the input frame *fi* (in the *i*th frame). Therefore, we extend the spectral residual ([17]) from pixel level to superpixel level:

$$R(sij) = L(sij) - A(sij) \tag{3}$$

where $L(sij)$ is the log amplitude spectrum of *sij*, and $A(sij)$ denotes the general shape of log spectra, which is given as prior information, and it is obtained from 3*3 mean filter of $L(sij)$.

Then, the saliency map [17] on superpixel level is defined as follows:

$$S\text{intra}1(sij) = G * F^{-1}(exp(R(sij) + P(sij)))^2 \tag{4}$$

where $G(\sigma = 8)$ is a Gaussian filter. F^{-1} denotes the Inverse Fourier Transform. $P(sij)$ denotes the phase spectrum of *sij*, which is preserved during the process.

On the other hand, Achanta and Hemami et al. introduce a frequency-tuned approach to estimate center-surround contrast using color and luminance features [2]. This method both consider the color and luminance of the whole image, and it can find out the salient regions which are outstanding in color. Hence we extend this method to superpixel level as the second part of our intra saliency. As same as before, we transfer the frame to CIELab space. Then, the intra saliency map ([2]) on superpixel level is as follows:

$$S\text{intra}2(Sij) = \|\mu(fi) - \omega hc(Sij)\| \tag{5}$$

where $\mu(fi)$ is the mean pixel value (CIELab model) in the input frame *fi*, $\omega hc(sij)$ is the corresponding Gaussian blur (using a 5×5 separable binomial kernel) of *sij*, and $\|\|$ is the L_2 norm (Euclidean distance).

After the two parts of superpixel intra saliency map are collected. We overall consider both saliency maps and add these two saliency maps together by weight. Therefore, the intra saliency map on superpixel level is finalized as below:

$$S\text{intra}(Sij) = w\text{intra}1S\text{intra}1(Sij) + w\text{intra}2S\text{intra}2(Sij) \tag{6}$$

where *S*intra(*sij*) denotes the normalized intra saliency value of *sij*, and the value of *S*intra(*sij*) is within [0, 1]. Here, *w*intra*i*, *i* = 1 or 2. denotes the weight with *w*intra1 + *w*intra2 = 1 (Fig. 5).

2.2.2 Inter saliency map generation

As mentioned, co-saliency is proposed to highlight the superpixels which are salient in their frame and exists in each video clip. On one hand, the saliency map of each superpixel compared with the whole frame is described by intra saliency map. Hence, we present a new scheme to find out the superpixels which share similar features in each clip through determining an inter saliency map.

First of all, object proposal based method is applied to coarsely segment out an initial object (object prior). Given that there is a common object in every video clips, the remaining task in this stage is to find out the common objects. The hypothesis is that if common objects are shared by two video clips, they are similar in region feature in each clip. Hence we match every object proposal with the proposals in other clips which are measured by the proposed hierarchical histogram based region feature. And then, the most similar pairs of object proposals will be chosen as object priors. The research of object proposal has been very mature, and it is very helpful in assistance of video tracking, image segmentation and so on [4, 12, 23, 26]. Therein, a set of segmentations by performing graph cuts based on a seed region and a learned affinity function is generated by category independent object proposals, and then the regions are ranked by various cues [12]. This method has been shown to be quite robust and efficient. Therefore, we have employed it to generate a number of object proposals in this stage.

We can see the object prior from Fig. 6. And then, we define the inter saliency map as follow: given that *sij* is the *j*th superpixel in the chosen object prior (in Fig. 6) *Pi* in the *i*th video (*sij* ∈ *Pi*, and it is a set of superpixels). Then, let *Fsij* be the proposed region feature based on hierarchical histogram of *sij*, and let *FPg* be the region feature of another object prior in the *g*th clip *Pg*.

Afterwards, we use the inner product of the features of two regions *sij* and *Pg* to quantify the similarity between them, hence this similarity is the inter saliency map in this paper:

$$S\text{inter}(Sij) = \rho(Sij, Pg) = Fsij \cdot FPg \tag{7}$$

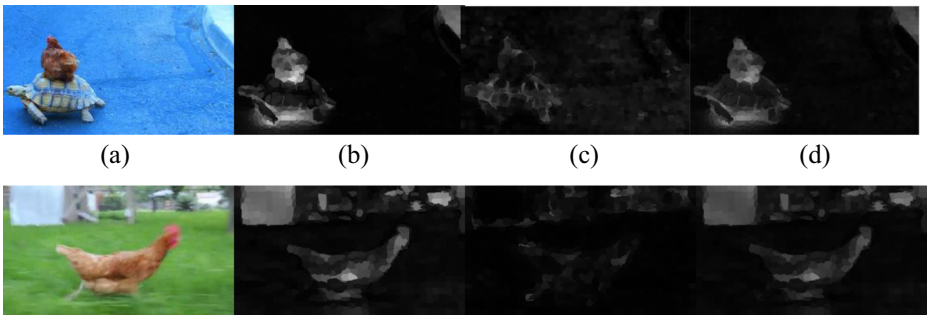

Fig. 5 Intra saliency map on superpixel level: (**a**) & (**e**) are the original frames; (**b**) & (**f**) are the saliency maps from [17] on LAB channel; (**c**) & (**g**) are the saliency maps from [2]; (**c**) & (**g**) are the final saliency map

From Eq. (7), the inter saliency value is within (0, 1]. Two regions are similar if and only if the inner product of their features is close to 1 (Fig. 7).

Sum up, the co-saliency value on every superpixel s is:

$$S(s) = w\text{intra}S\text{intra}(s) + w\text{inter}S\text{inter}(s) \tag{8}$$

Especially, if s is out the range of the object priors, the value of $S\text{inter}(s)$ is 0.

3 Image co-segmentation based on marker prediction and region merging

According to the assumption, the most co-salient superpixels are chosen as object marker superpixels. Some existing algorithms try to directly segment out the objects after saliency map generation. Therein, Fu, Cao and Tu utilize the co-saliency map to extract the salient foreground from the input image by a Markov random field function [14]. However, this algorithm is not robust enough. Therefore, we propose a novel segmentation scheme:

After the superpixel co-saliency maps are generated, initial highlights of a set of objects are provided. In essence, the most salient superpixels which are considered as

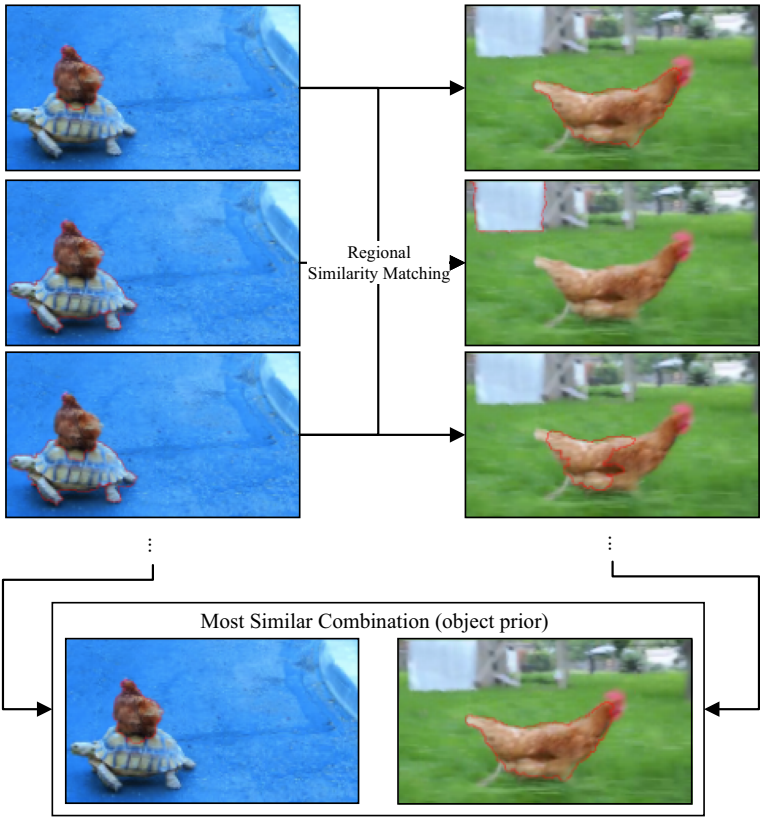

Fig. 6 The similarity comparison in inter saliency

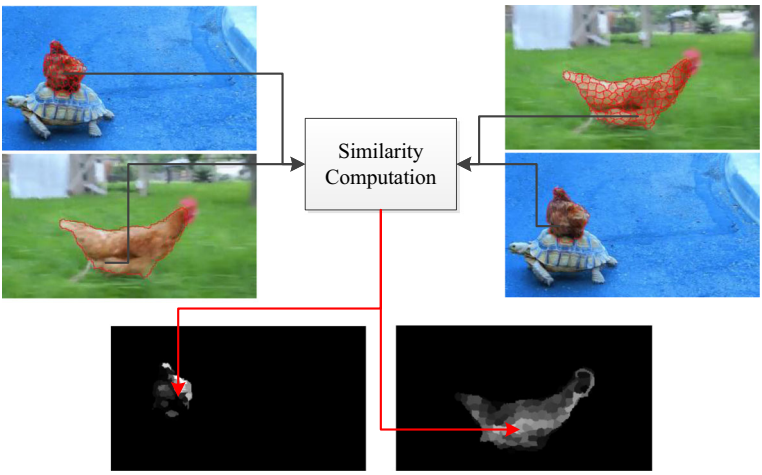

Fig. 7 Inter saliency map generation on superpixel level

the initial object marker superpixels. In the proposed method, we select the *N*object most salient superpixels with highest saliency values computed by our proposed Co-saliency Map Generation algorithms (as in section 2.2) as the object marker superpixels. The superpixels with the saliency value lower than the threshold *T*background will be selected as the background marker superpixels. These two parameters are set by experiments to have overall best performance.

And then, region merging stage is applied to extract the object regions from background. Ning and Zhang etc. propose a region merging based on maximal similarity, which is robust and efficient to interactively segment out the object [28]. The whole MSRM process can be divided into two stages, which are repeatedly executed until no new merging occurs.

Starting the region merging from the object marker superpixels and background regions, object regions and background regions will be merged respectively until there are only two kinds of regions left – object and background. More than one object regions may be segmented out, but they must also share the common region features. The whole merging process will be divided into two stages, which repeat until no new merging occurs.

Before the region merging process, a set of superpixels $S = \{si|si \in f\}$ (f is a frame) are segmented from over segmentation. Besides, the initial object marker superpixels $O = \{oi|oi \in S\}$ and background marker superpixels $B = \{bi|bi \in S\}$ are also input. Moreover, the other superpixels are non-marker superpixels $N = \{ni|ni \in S\}$. Therefore, $O \cup B \cup N = S$. Especially, S and N are always superpixel sets in every loop. However, O and B are merged to regions from superpixels except the beginning of the first loop.

In the first stage, the task is to merge background marker regions B with their adjacent non-marker superpixels N. The object marker regions $O = \{oi|oi \in S\}$, background marker regions $B = \{bi|bi \in S\}$, and non-marker superpixels $N = \{ni|ni \in S\}$ are updated by the merging process in the second stage in the previous iteration. First, for each background region bi, form the set of its adjacent regions or superpixels $\overline{Sbi} = \{ai, i = 1, \ldots r\}$. Second, for each ai and $ai \notin B$, form its set of adjacent regions $\overline{Sai} = \{sj^{ai}, j = 1, \ldots k\}$. There is $bi \in \overline{Sai}$. Third, calculate the

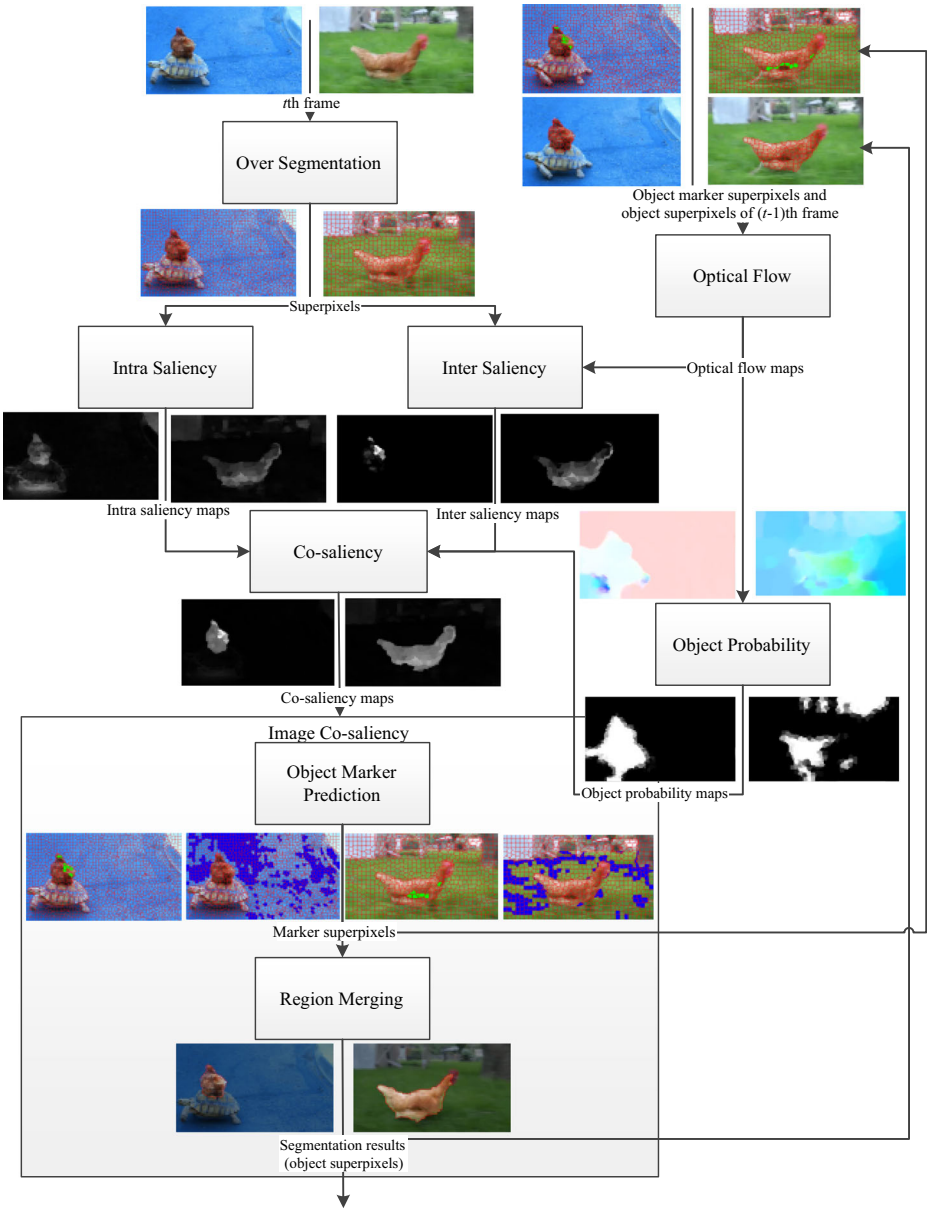

Fig. 8 The process of video co-segmentation

region similarity $\rho(ai, sj^{ai})$ (which is defined in Eq. (2)). If $\rho(ai, bi) = \max j = 1, \ldots k\rho(ai, sj^{ai})$, then $bi = bi \cup ai$. Otherwise, bi and ai will not merge. Fourth, update B and N respectively. Fifth, if the background regions B cannot find out the new merging regions, and the first stage ends. Otherwise, go back to the first step.

In the second stage, the task is to merge the non-marker superpixels N. First of all, the object marker regions $O = \{oi | oi \in S\}$, background marker regions $B = \{bi | bi \in S\}$, and non-

Table 1 The parameters for our experiments

Step	Parameters
Over Segmentation	superpixel size: 180 compactness: 12
Feature Extraction	number of layers: $m = 3$
Initialization	$w$intra1 = 0.5, $w$intra2 = 0.5, $w$intra = 0.3, $w$inter = 0.7, Nobject = 7, Tbackground = 0.1
Video Co-segmentation	$w$intra1 = 0.5, $w$intra2 = 0.5, $w$intra = 0.2, $w$inter = 0.5, $w$probability = 0.3, Nobject = 7, Tbackground = 0.1

marker superpixels $N = \{ni | ni \in S\}$ are updated by the merging process in the first stage of this iteration. First, for each non-marker region ni, form the set of its adjacent regions $\overline{Sni} = \{hi, i = 1, \ldots p\}$. Second, for each hi ($hi \notin O$ and $hi \notin B$), form its set of adjacent regions $\overline{Shi} = \{sj^{hi}, j = 1, \ldots k\}$. There is $ni \in \overline{Shi}$. Third, calculate the region similarity $\rho(hi, sj^{hi})$. If $\rho(hi, ni) = \max j = 1, \ldots k \rho(hi, sj^{hi})$, then $ni = ni \cup hi$. Otherwise, ni and hi will not merge. Fourth, we update N. Fifth, if the non-marker regions N cannot find out new merging region, the second stage stops. Otherwise, go back to the first step. And then, the last iteration ends, and the final segmentation map (finalized object superpixels $O' = \{oi' | oi' \in S\}$ and $B = \{bi | bi \in S\}$, and we have $O' \cup B' = S$) are output.

In addition, the MSRM region merging process can merge different object regions respectively by considering other objects as background regions.

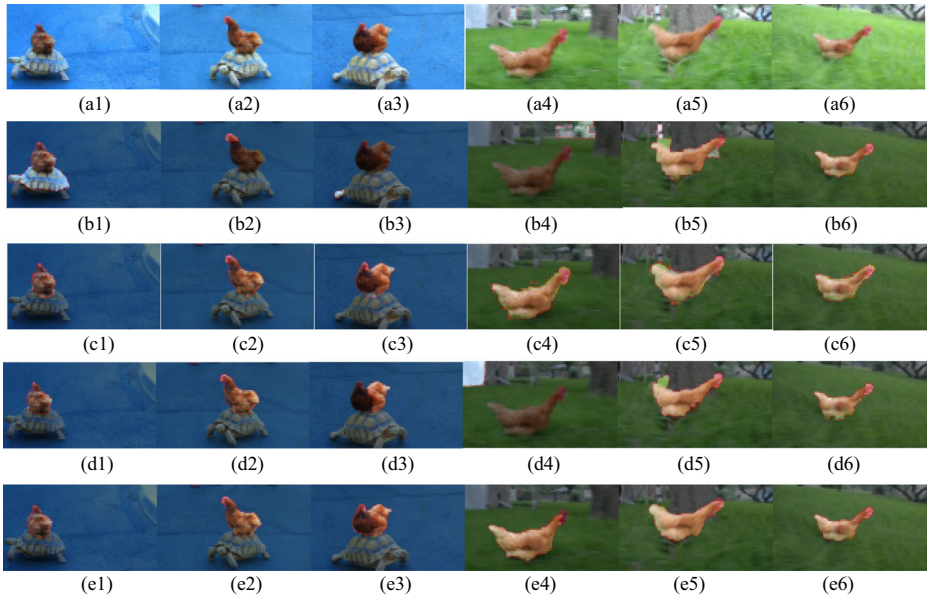

Fig. 9 Visual comparison with DAG, RMWC and ObMiC on "Chicken": (**a1**) - (**a3**) are original frames from *chicken_on_turtle*; (**a4**) - (**a6**) are original frames from *chickenNew*; (**b1**) - (**b6**) are results from DAG; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

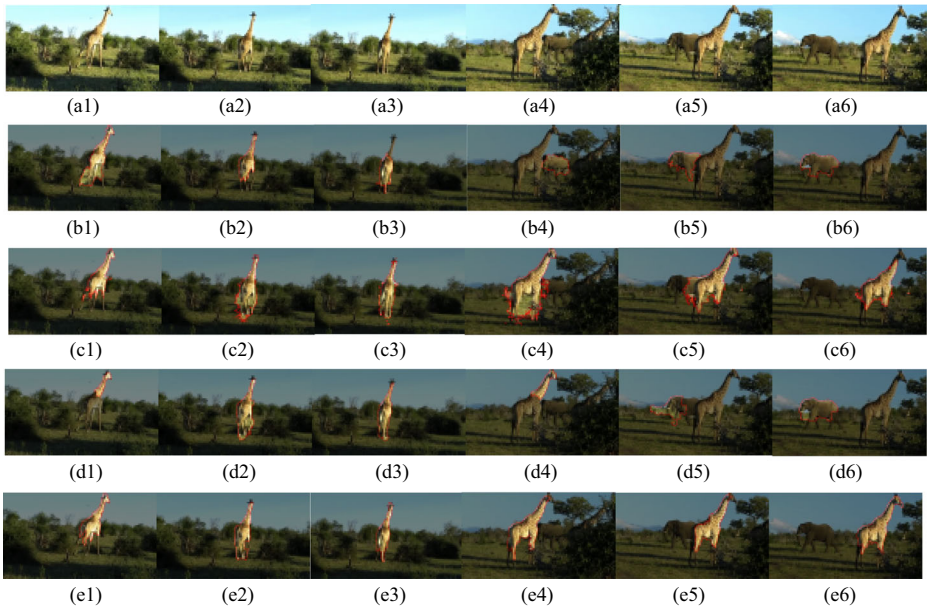

Fig. 10 Visual comparison with DAG, RMWC and ObMiC on "Giraffe": (**a1**) - (**a3**) are original frames from *elephant_giraffe_all1*; (**a4**) - (**a6**) are original frames from *elephant_giraffe_all2*; (**b1**) - (**b6**) are results from DAG; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

4 Video object co-segmentation based on superpixel trajectory

After the initialization stage, the segmentation results of the first frame are output. Based on the results from initialization stage, we propose an on-line video co-segmentation. In our scheme, video co-segmentation can be considered as the image co-segmentation on every frame. Therefore, it is necessary to predict and generate the initial object and background marker regions automatically. In this section, we will present an on-line video co-segmentation based on superpixel trajectory. It is analogous to the initialization stage. In addition, the proposed video co-segmentation scheme is on-line, so the results from previous frames need to guide the segmentation of the current frame.

Actually, the co-segmentation in every frame is an interactive image segmentation process. For the segmentation of the pending frame, we only need to provide the object marker superpixels which are salient and similar with object superpixels in other clips. Analogous to the initialization stage, the object marker superpixels are intra and inter salient. The intra saliency map computation is the same as the initialization. In the inter saliency stage, it is not necessary to generate the object proposal again. First of all, we present a scheme of superpixel motion trajectory to generate the object prior and track the object marker superpixels (Fig. 8).

4.1 Superpixel motion trajectory

At every frame, the segmentation result from the previous frame is chosen as the initial segmentation in inter saliency stage. And then, following the initialization stage, the inter

Fig. 11 Visual comparison with DAG, RMWC and ObMiC on "Lion": (**a1**) & (**a2**) are original frames from *lion_Zebra_all1*; (**a3**) & (**a4**) are original frames from *lion_Zebra_all2*; (**a5**) & (**a6**) are original frames on *lion_Zebra2*; (**b1**) - (**b6**) are results from DAG; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

saliency maps of the current frames in every clips are generated. The number of object marker superpixels is the same in every frame.

For each object marker superpixel, we need to roughly predict its motion. The optical flow based method at the superpixel level is a popular choice for this task and has been extensively researched [15, 34]. In this paper, we implement the optical flow model based on the strategy of [34] because of its simplicity and rapid operation. The optical flow based method is applied to predict the location of object marker superpixels $omi-1j$ and object superpixels $oi-1j$ ($omi-1j \in oi-1j$, starting from $omi-1j$, segment out $oi-1j$ by region merging) in frame $fi-1$.

Let pixel $pi \in s$, $i = 1, 2, \ldots n$, be a pixel in superpixel s, where n is the size of s. The expected displacement of pi from the proposed optical flow is (ui, vi). Finally, we define the expected trajectory of s as $(\overline{u}, \overline{v}) = median((u1, v1), (u2, v2), \ldots, (un, vn))$, where *median* is the median operator.

Therefore, we have

$$\begin{aligned} tmij &= omi{-}1j + \left(\overline{u}, \overline{v}\right)ij \\ tij &= oi{-}1j + \left(\overline{u}, \overline{v}\right)ij \end{aligned} \tag{9}$$

where $tmij$ and tij ($tmij$ may not in tij) are predicted location in frame fi from the object marker superpixel $omi-1j$ and object superpixel $oi-1j$.

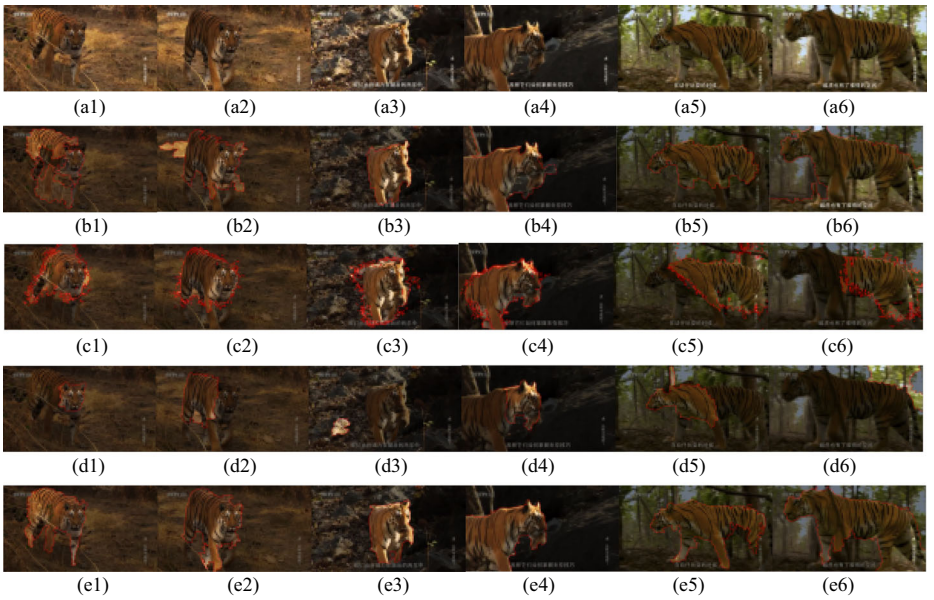

Fig. 12 Visual comparison with VCS, RMWC and ObMiC on "Tiger": (**a1**) & (**a2**) are original frames from *tiger1_all8*; (**a3**) & (**a4**) are original frames from *tiger1_all9*; (**a5**) & (**a6**) are original frames on *tiger1_all10*; (**b1**) - (**b6**) are results from VCS; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

4.2 Video co-saliency

Familiar with section 2.2.1, our intra saliency are the same in every frame.

$$S\text{intra}(Sij) = w\text{intra}1S\text{intra}1(Sij) + w\text{intra}2S\text{intra}2(Sij) \tag{10}$$

In inter saliency stage, the object proposal is replaced with the expected location which is from the set of the trajectories *tij* of the object superpixels $oi-1j$. Therefore, the object prior in the frame *fi* is within this area:

$$oprior = \cup oi{-}1j \tag{11}$$

Table 2 Quantitative comparisons with the state of the art methods on the MOViCS dataset

Video Set/Method	DAG [14]		RMWC [4]		ObMiC [15]		Ours	
Chicken	0.68	206.49 s	0.83	269.39 s	0.87	420.02 s	**0.85**	**87.93 s**
Giraffe	0.56	334.33 s	0.64	273.03 s	0.63	427.05 s	**0.71**	**88.04 s**
Lion	0.66	293.59 s	0.64	146.34 s	0.71	295.00s	**0.82**	**73.39 s**
Tiger	0.55	244.80s	0.34	319.03 s	0.54	570.18 s	**0.56**	**93.83 s**
Overall	0.61	269.80s	0.62	251.95 s	0.68	428 s	**0.74**	**69.30s**

The first column of each method is the average intersection-over-union metric per frame; the second column of each method is the average execution time per frame (the unit is second); the values in bold are best performance

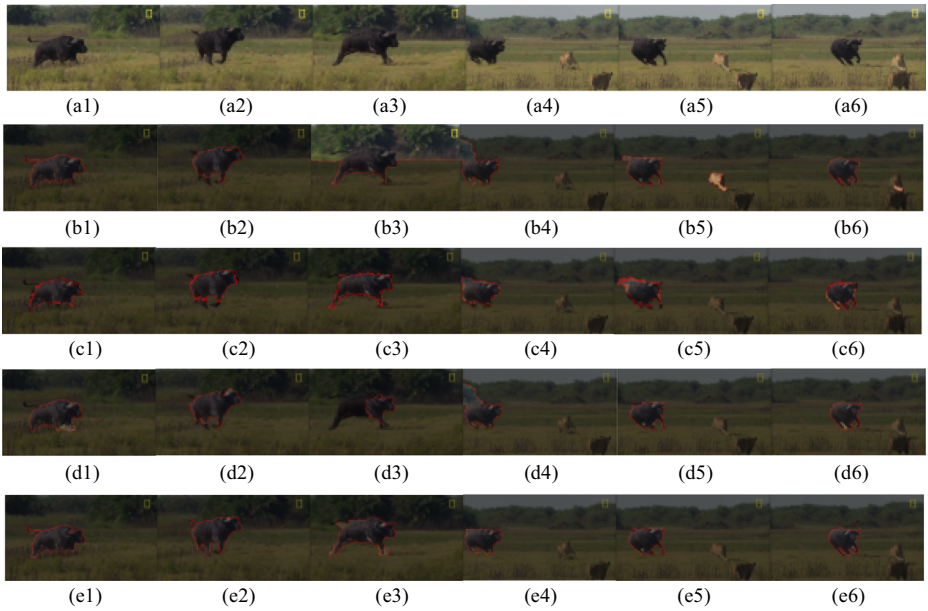

Fig. 13 Visual comparison with DAG, RMWC and ObMiC on "Buffalo": (**a1**) - (**a3**) are original frames from *buffalo*; (**a4**) - (**a6**) are original frames from *buffalo_lion*; (**b1**) - (**b6**) are results from DAG; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

Then, following with

$$Sinter(Snij) = \rho(snij, opriormi) = Fsnij \cdot Fopriormi \tag{12}$$

where *snij* is the *j*th superpixel in the *i*th frame *fni* in the *n*th clip, and *opriormi* is the object prior in *i*th frame *fmi* in the *m*th clip from Eq. (11). *Fsnij* and *Fopriormi* are respectively the proposed region feature (from section 2.1) of *snij* and *opriormi*.

Normally, the superpixels between two frames do not shift too far. According to this assumption, we propose an object superpixel probability as follows:

$$Sprobability(Sij) = \begin{cases} 1, sij \in \{\cup oi-1j\} \cap \{\cup omi-1j\} \\ 0.8, sij \in \cup omi-1j, and sij \notin \cup oi-1j \\ 0.6, sij \in \cup oi-1j, and sij \notin \cup omi-1j \\ 0, sij \notin \cup omi-1j, and sij \notin \cup oi-1j \end{cases} \tag{13}$$

Sum up, the co-saliency map of the superpixel *sij* is:

$$S(sij) = wintraSintra(sij) + winterSinter(sij) + wprobabilitySprobability(sij) \tag{14}$$

where *w*intra, *w*inter and *w*probability are respectively the weight of *S*intra, *S*inter and *S*probability, and *w*intra + *w*inter + *w*probability = 1.

After the co-saliency maps are generated, the object and background marker superpixels will be generated the same as the initialization stage.

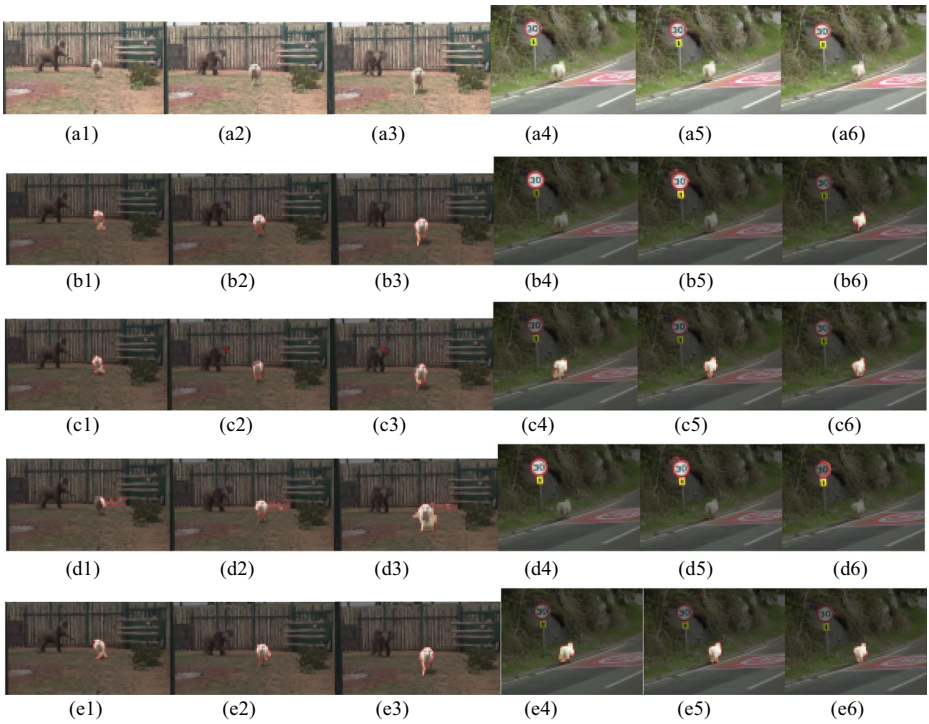

Fig. 14 Visual comparison with DAG, RMWC and ObMiC on "Sheep": (**a1**) - (**a3**) are original frames from *elephant_sheep*; (**a4**) - (**a6**) are original frames from *sheep*; (**b1**) - (**b6**) are results from DAG; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

5 Experimental results

The proposed method was tested on three challenging video co-segmentation dataset (MOViCS dataset [10], Safari dataset [37] and ObMiC dataset [15]) and was compared with other several state-of-the-art methods (DAG [14], RMWC [4] and ObMiC [15]). We test the single object co-segmentation on MOViCS dataset [10] and Safari dataset [37] and the multi object co-segmentation on ObMiC dataset. The results show that our method performs better both qualitatively and quantitatively. All experiments are launched in windows 7 64 bit, Intel Core i7-3517U CPU (dual core and quad threading: 1.9 GHz and 2.4 GHz) with 4 GB RAM. Especially, our method is an on-line segmentation.

Table 3 Quantitative comparisons with the state of the art methods on the Safari dataset

Video Set/Method	DAG [14]		RMWC [4]		ObMiC [15]		Ours	
Buffalo	0.68	436.61 s	**0.87**	228.04 s	0.78	412.98 s	0.74	**86.48 s**
Elephant	0.47	389.24 s	0.35	137.93 s	**0.57**	394.34 s	0.51	**78.93 s**
Lion	**0.49**	479.48 s	0.32	264.83 s	0.43	347.89 s	0.43	**79.33 s**
Sheep	0.38	500.20s	0.36	287.83 s	0.33	578.50s	**0.75**	**80.98 s**
Overall	0.51	451.38 s	0.48	229.66 s	0.53	433.43 s	**0.61**	**81.43 s**

The values in bold are the best results

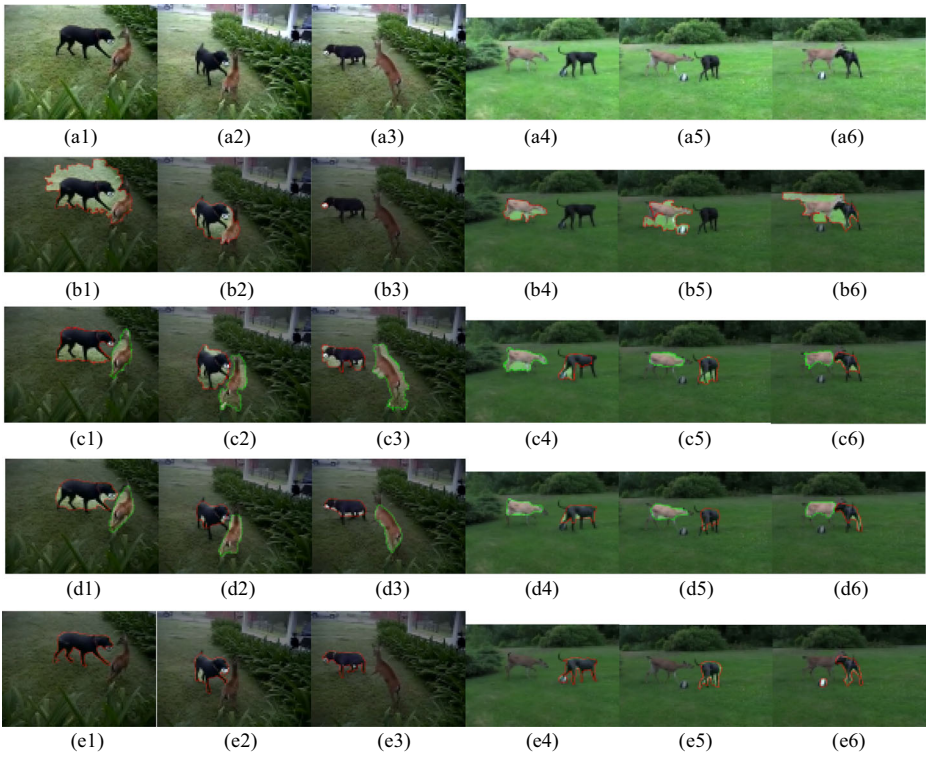

Fig. 15 Visual comparison with DAG, RMWC and ObMiC on "Dog": (**a1**) - (**a3**) are original frames from *DogDeer_2*; (**a4**) - (**a6**) are original frames from *DogDeer_4*; (**b1**) - (**b6**) are results from DAG; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

Therefore, the user can handle with the video frame by frame. In the experiments, we do not only compare with other video co-segmentations (RMWC [4] and ObMiC [15]), but also compare with a traditional unsupervised single video segmentation which cannot segment out the common objects from a set of video clips simultaneously (DAG [14]). The user needs to segment clip by clip by using DAG. Moreover, the execution time of our method is shorter than other 3 competing methods. For all experiments, the parameters are the same and shown as Table 1. In over segmentation stage (superpixel generation), theoretically, the smaller and more distortion each superpixel is, the more accurate the segmentation result is. We only apply three-layer hierarchical histograms to extract the region feature, because the texture information is relatively simple in all experiments. From the results of many experiments on different datasets, we find that our proposed method has overall best performance with the parameters as shown in Table 1.

5.1 Tests of MOViCS dataset

The MOViCS dataset ([10]) is a popular video co-segmentation dataset which has the ground truth annotations for quantitative analysis. It contains 4 video sets which totally has 11 videos, 5 frames of each video have pixel-level annotations for the object labels. In this section, we evaluate our method and three state-of-the-art method on the MOViCS

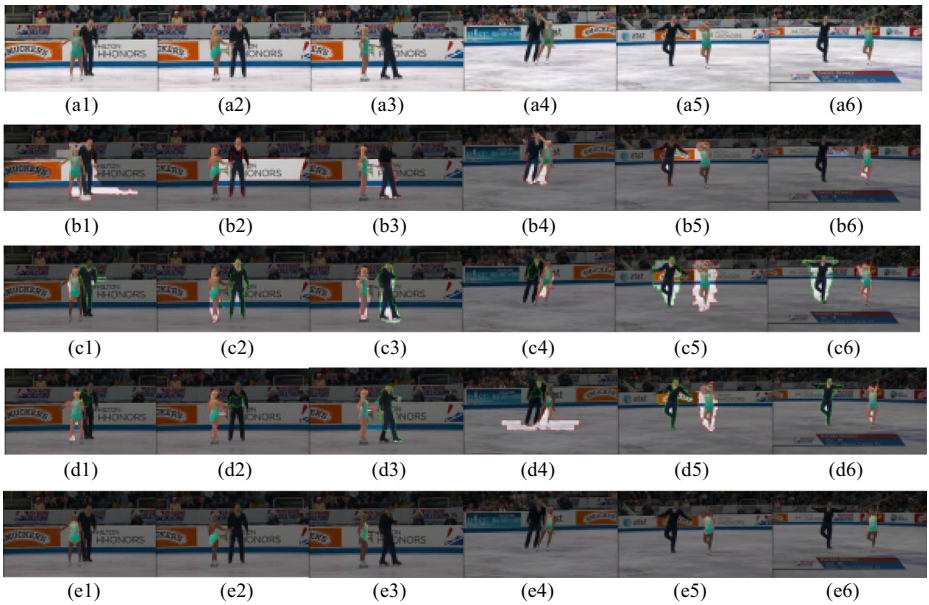

Fig. 16 Visual comparison with DAG, RMWC and ObMiC on "Skating": (**a1**) - (**a3**) are original frames from *Skating_2*; (**a4**) - (**a6**) are original frames from *Skating_3*; (**b1**) - (**b6**) are results from DAG; (**c1**) - (**c6**) are results from RMWC; (**d1**) - (**d6**) are results from ObMiC; (**e1**) - (**e6**) are results from our method

dataset to test single object co-segmentation. For fair comparison, we keep all parameters (for every method) constant in every experiment.

From the above figures, we can see that the results from our method are stably attached with the boundaries of objects. Unlike prior methods, our method does not have the property to break objects into a number of fragments and the proposed method also produces better contours. For example, the only common object (chick) (Fig. 9) is segmented out accurately by our method without any background part. Two videos clips in which the object regions are not accurately segmented are *tiger1_all8*, *tiger1_all10* in video set "tigers" and *elephant_giraffe_all1*, *elephant_giraffe_all1* in video set "giraffe". This is due to the features of objects in these clips are similar to the background (Figs. 10, 11 and 12).

In order to quantify our results, we adopt the intersection-over-union metric as below:

$$M(S,G) = \frac{S \cap G}{S \cup G} \tag{15}$$

where S is a set of segments and G is the ground truth.

We compare our method with 3 state-of-the-art methods (DAG [14], RMWC [4] and ObMiC [15]) as shown in Table 2. Note that our method is an unsupervised method, and it outperforms the other methods except for the set "Chicken" where it is a close second. The results in Table 2 are the average per-frame intersection-over-union metric compared to the ground truth. The definition is as Eq. (15).

Table 4 Quantitative comparisons with the state of the art methods on the ObMiC dataset

Video Set/Method	DAG [14]		RMWC [4]		ObMiC [15]		Ours	
Dog	0.37	238.54 s	0.55	117.40s	**0.69**	295.03 s	0.43	**29.49 s**
Person	0.65	250.38 s	0.78	109.34 s	**0.83**	286.25 s	0.68	**39.02 s**
Monster	0.56	324.45 s	**0.70**	192.54 s	0.66	254.53 s	0.52	**35.93 s**
Skating	0.23	238.90s	0.64	194.89 s	**0.72**	343.85 s	0.33	**50.45 s**
Overall	0.45	263.07 s	0.67	153.54 s	**0.73**	294.92 s	0.49	**38.72 s**

The values in bold are the best results

5.2 Tests of safari dataset

Video object co-segmentation problem is new. To our best knowledge, there are two publicly available dataset with ground truth for testing single object co-segmentation (MOViCS dataset [10] and Safari dataset [37]). This Safari dataset contains 5 classes of animals and a total of 9 videos. We show the visual comparisons between our method and other three state-of-the-art methods (DAG [14], RMWC [4] and ObMiC [15]) in Figs. 13 and 14, and quantitative results are shown in Table 3.

Figure 13 show that our method yields best performance on the set "Buffalo". We can compare with Fig. 13 (b4), (c4), (d4) and (e4), and the background region can be eliminated only by our method. In Fig. 14, the performance of our method is not good enough at the clip *sheep* from "Sheep". On one hand, it is partly because the features of common objects (sheep) from *elephant_sheep* and *sheep* are not similar enough (especially the color). On the other hand, it is due to the boundaries generated by superpixel generation are not accurate enough.

5.3 Tests of ObMiC dataset

The above two datasets (MOViCS dataset [10] and Safari dataset [37])) are collected to test the single common object co-segmentation. For evaluating the performance on multi object co-segmentation, we test all methods on ObMiC Dataset [18]. The ObMiC dataset contains 4 pairs of video clips (total 8 clips), and there are two common objects in every clip. We show the visual comparisons in Figs. 15 and 16, and quantitative comparisons are shown in Table 4.

In Figs. 15 and 16, our method sometimes cannot segment out both two common objects effectively, because our method is designed for single object co-segmentation. Besides, in Fig. 15, only the blue skirt of the stating player is segmented out, because the skirt is very salient and shape. For this case, it also seems that our method is not robust enough to the object with many different colors.

6 Conclusion

In this paper, we propose a robust video object co-segmentation method based on co-saliency and region merging. We show that the proposed co-saliency scheme based on superpixel is applied to initialize the segmentation, which include two stage – superpixel intra co-saliency and superpixel inter co-saliency. Some most co-salient superpixels are chosen as the object marker superpixels. And then, region merging will start from these object marker superpixels until the object regions are segmented out. Besides, a novel region feature based on

hierarchical histogram is proposed to represent each region and superpixel. Numerous experimental results and evaluations demonstrate the proposed method performs favorably against existing state-of-the-art algorithms in the literature in some datasets for video co-segmentation. We demonstrated that the proposed method efficiently segments the common objects out from a set of video clips. However, we also found the performance of our proposed method may not be robust enough if there are multiple objects in the video clip. Besides, these segmentation results can be further improved by establishing the more robust relevance between objects in different clips and using more spatial-temporal information. As the most time consuming parts of the proposed algorithm are region merging and object proposal process, we will explore to develop some efficient and effective alternatives even propose new algorithms. Since a novel region feature based on hierarchical histogram is applied to describe each superpixel in our paper, better region features can be incorporated to further improve the results.

References

1. Achanta R, Estrada F, Wils P, Süsstrunk S (2008) Salient region detection and segmentation. In: Gasteratos A, Vincze M, Tsotsos J (eds) Computer Vision Systems. vol. 5008, Springer Berlin Heidelberg, pp. 66–75
2. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: 2009 I.E. Conference on Computer Vision and Pattern Recognition, pp. 1597–1604
3. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34:2274–2282
4. Alexe B, Deselaers T, Ferrari V (2012) Measuring the objectness of image windows. IEEE Trans Pattern Anal Mach Intell 34:2189–2202
5. Badrinarayanan V, Budvytis I, Cipolla R (2013) Semi-supervised video segmentation using tree structured graphical models. IEEE Trans Pattern Anal Mach Intell 35:2751–2764
6. Bai X, Wang J, Simons D, Sapiro G (2009) Video SnapCut: robust video object cutout using localized classifiers. ACM Trans Graph 28:1–11
7. Batra D, Kowdle A, Parikh D, Jiebo L, Tsuhan C (2010) iCoseg: interactive co-segmentation with intelligent scribble guidance. In: 2010 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176
8. Cao X, Tao Z, Zhang B, Fu H, Feng W (2014) Self-adaptively weighted co-saliency detection via rank constraint. IEEE Trans Image Process 23:4175–4186
9. Cheng M-M, Zhang G-X, Mitra NJ, Huang X, Hu S-M (2011) Global contrast based salient region detection. In: 2011 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 409–416
10. Chiu W-C, Fritz M (2013) Multi-class video co-segmentation with a generative multi-video model. In 2013 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 321–328
11. Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24:603–619
12. Endres I, Hoiem D (2010) Category Independent Object Proposals. In: Daniilidis K, Maragos P, Paragios N (eds) Computer Vision – ECCV 2010. vol. 6315, Springer Berlin Heidelberg, pp. 575–588
13. Feng T, Brennan S, Qi Z, Hai T (2007) Co-tracking using semi-supervised support vector machines. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pp. 1–8
14. Fu H, Cao X, Tu Z (2013) Cluster-based co-saliency detection. IEEE Trans Image Process 22:3766–3778
15. Golland P, Bruckstein AM (1997) Motion from color. Comput Vis Image Underst 68:346–362
16. Hochbaum DS, Singh V (2009) An efficient algorithm for Co-segmentation. In: 2009 I.E. 12th International Conference on Computer Vision, pp. 269–276
17. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: 2007 I.E. Conference on Computer Vision and Pattern Recognition, pp. 1–8
18. Huazhu F, Dong X, Bao Z, Lin S (2014) Object-based multiple foreground video co-segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2014 I.E. Conference on, pp. 3166–3173
19. Jiaming G, Zhuwen L, Loong-Fah C, Zhou SZ (2013) Video co-segmentation for meaningful action extraction. In: 2013 I.E. International Conference on Computer Vision (ICCV), pp. 2232–2239
20. Joulin A, Bach F, Ponce J (2010) Discriminative clustering for image co-segmentation. In: 2010 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1943–1950

21. Joulin A, Bach F, Ponce J (2012) Multi-class cosegmentation. In 2012 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 542–549
22. Kailath T (1967) The divergence and bhattacharyya distance measures in signal selection. IEEE Trans Commun Technol 15:52–60
23. Krähenbühl P, Koltun V (2014) Geodesic object proposals. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision – ECCV 2014. vol. 8693, Springer International Publishing, pp. 725–739
24. Li H, Ngan KN (2011) A co-saliency model of image pairs. IEEE Trans Image Process 20:3365–3375
25. Liu Z, Zou W, Li L, Shen L, Meur OL (2014) Co-saliency detection based on hierarchical segmentation. IEEE Signal Process Lett 21:88–92
26. Manen S, Guillaumin M, Gool LV (2013) Prime object proposals with randomized prim's algorithm. Presented at the Proceedings of the 2013 I.E. International Conference on Computer Vision
27. Meng F, Li H, Liu G, Ngan KN (2012) Object co-segmentation based on shortest path algorithm and saliency model. IEEE Trans Multimedia 14:1429–1441
28. Ning J, Zhang L, Zhang D, Wu C (2010) Interactive image segmentation by maximal similarity based region merging. Pattern Recogn 43:445–456
29. Tang K, Joulin A, Li-Jia L, Li F-F (2014) Co-localization in real-world images. In: 2014 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1464–1471
30. Van den Bergh M, Boix X, Roig G, de Capitani B, Van Gool L (2012) SEEDS: superpixels extracted via energy-driven sampling. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (eds) Computer Vision – ECCV 2012. vol. 7578, Springer Berlin Heidelberg, pp. 13–26
31. Vedaldi A, Soatto S (2008) Quick shift and kernel methods for mode seeking. In: Forsyth D, Torr P, Zisserman A (eds) Computer Vision – ECCV 2008. vol. 5305, Springer Berlin Heidelberg, pp. 705–718
32. Vincent L, Soille P (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans Pattern Anal Mach Intell 13:583–598
33. Wang T, Han B, Collomosse J (2014) TouchCut: fast image and video segmentation using single-touch interaction. Comput Vis Image Underst 120:14–30
34. Willert V, Eggert J, Clever S, Körner E (2005) Probabilistic color optical flow. In: Kropatsch W, Sablatnig R, Hanbury A (eds) Pattern Recognition. vol. 3663, Springer Berlin Heidelberg, pp. 9–16
35. Zhai Y, Shah M (2006) Visual attention detection in video sequences using spatiotemporal cues. Presented at the Proceedings of the 14th annual ACM international conference on Multimedia, Santa Barbara, CA, USA
36. Zhang D, Javed O, Shah M (2013) Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. Presented at the Proceedings of the 2013 I.E. Conference on Computer Vision and Pattern Recognition
37. Zhang D, Javed O, Shah M (2014) Video object co-segmentation by regulated maximum weight cliques. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision – ECCV 2014. vol. 8695, Springer International Publishing, pp. 551–566

Guoheng Huang received his B.Sc. degree in Applied Mathematics and M.E degree in Software Engineering from South China Normal University. He is currently a Ph. D. student majoring at software engineering at the University of Macau. His research interests include Image/Video Processing and Pattern Recognition.

Chi-Man Pun received his B.Sc. and M.Sc. degrees in Software Engineering from the University of Macau in 1995 and 1998 respectively, and Ph.D. degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2002. He is currently an Associate Professor at the Department of Computer and Information Science of the University of Macau. He has investigated several funded research projects and published more than 100 refereed scientific papers in international journals, books and conference proceedings. Dr. Pun has served as the General Chair for the 10th International Conference Computer Graphics, Imaging and Visualization (CGIV2013), and program / session chair for several other international conferences. He has also served as the editorial member / referee for many international journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, Pattern Recognition, etc. His research interests include Digital Image Processing; Digital Watermarking; Pattern Recognition and Computer Vision; Intelligent Systems and Applications. He is also a senior member of the IEEE and a professional member of the ACM.

Cong Lin received his B.Sc. degree in Information & Computational Science from Guangdong University of Technology M.Sc. degree in Software Engineering from University of Macau. He is currently a Ph. D. student majoring at software engineering at the University of Macau. His research interests include Image/Video Processing and Pattern Recognition.