



# Strong target constrained video saliency detection

Shoudong Han<sup>1,2</sup> · Yujun Liu<sup>1</sup> · Zhuo Hu<sup>3</sup> 

© Springer Nature Switzerland AG 2019

## Abstract

Aiming at the problem of video saliency detection, a strong target constrained video saliency detection method is proposed in this paper. In order to detect the salient region fast and effectively, strong target constraints forcing by the location, scale and color model are introduced into the video saliency detection. First, a target locating strategy for obtaining the location and scale information is proposed by correcting the result of video tracking with the optical flow result and the segmentation result of the last frame. Second, the estimated color model of the target is also calculated by the obtained segmentation results. Finally, the strong target constraints are integrated into the saliency model in the way of extending the significance hypothesis, and a high quality saliency map is obtained, where segmentation is employed for constrained parameters updating. In details, Densecut is initialized by the obtained saliency map to calculate the segmentation result of the last frame. Compared with some state-of-art saliency detection methods, our proposed method performs outstandingly, and the results on DAVIS dataset are significantly improved in terms of accuracy and robustness.

**Keywords** Saliency target detection · Target tracking · Video saliency · Video segmentation

## 1 Introduction

As the most common way to connect the outside world for human beings, vision can achieve many image-based functional purposes, such as face recognition, scene segmentation, target tracking and so on, while receiving a large amount of information [1–5]. The human visual system focuses on the “saliency” area, so the brain burden is greatly reduced by the processing and storage of information in the small part of the “saliency” area. There is no doubt that this way of focusing on the “saliency” area has a unique advantage in visual information processing.

In addition to extensive research in neurobiology [6, 7], cognitive psychology [8, 9] and other fields, the concept of saliency detection [10–12] also has a very important research value in computer vision field. For example, using salient region as prior information in image segmentation can automatically locate foreground targets,

and transform the interactive segmentation algorithm into automatic segmentation algorithm [13]. Otherwise, the extraction of the salient region can reduce the data to be processed, thereby the efficiency of the algorithm can be improved. Color contrast is the main feature of salient region and other features such as texture features, shape features, and space locations. By comparing the features, the algorithm determines the saliency of each region. For feature contrast, there are two main methods: local contrast and global contrast.

The saliency detection algorithm for local contrast is mainly based on the comparison between local regions. When contrasting features, the current region is compared only with some adjacent regions to find out the difference. Therefore, the algorithm tends to produce high saliency at areas such as the edge of the foreground object and noise. This kind of methods have the advantage of high execution efficiency. However, due to the locality of the

✉ Zhuo Hu, huzhuohust@163.com | <sup>1</sup>Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China. <sup>2</sup>Research Institute of Huazhong University of Science and Technology in Shenzhen, Shenzhen 518057, China. <sup>3</sup>School of Automation, Wuhan University of Technology, Wuhan 430070, China.



comparison areas, the algorithm is susceptible to noise and cannot form a salient region with stable connectivity. In order to detect salient regions more accurately, Liu [14] proposed a series of new features, including multiscale contrast, center-surround histogram, and color spatial distribution, and salient regions are detected by forming a conditional random field to efficiently combine these features.

On the contrary, the global contrast detection algorithm calculates the global contrast relationship of the image. It is obvious that the time complexity of the global contrast algorithm is very high. In applications, the saliency results are usually used as preprocessing operations, and the high time complexity certainly greatly reduces the practicability of the algorithm, even if the effect of the saliency detection is quite excellent. Cheng [15] proposed the global contrast based saliency detection algorithm by simplifying the color feature. The original three channel color is quantified from 256 values to 12 values, and color with lower probability in the image is ignored. By simplifying the operation, the color contrast in the image can be calculated quickly. This saliency method based on global contrast has achieved quite excellent results.

As an extension of image saliency in the field of video, video saliency [16, 17] has also received extensive attention. According to the features, the current video saliency detection methods can be divided into three categories: methods based on space domain [18], methods based on time domain [19, 20] and methods based on spatio-temporal domain [21]. It is generally believed that image saliency represents the most salient region in the current image, and the region most likely to be labeled as foreground. For images, salient region usually satisfies two hypotheses: the color difference between saliency area and any other region in the image is larger; the salient region is closer to the image center than the other regions [22]. However, the features of videos are richer than images. The image can be regarded as a single frame of the video, and the target in video often has the motion characteristics. After introducing multiple frames in videos, we can obtain not only the appearance information but also the temporal context information of the target between frames.

In this paper, we detect the saliency in the video by adding strong target constraints consist of location, scale and color to image saliency hypotheses, and then a strong target constrained video saliency detection (STCVSD) method is proposed. First, the location and scale information of the target are obtained by correcting the results of video tracking with the optical flow result and the segmentation result of the last frame. Second, the color model of the target is estimated by the fore/background Gaussian mixture model with the existing segmentation results.

Finally, the location information, scale information and the color model are integrated into the image saliency in the form of strong target constraints, and the strong target constrained video saliency results can be calculated. At the same time, Densecut [23] is initialized by the obtained saliency map to calculate the segmentation result of the last frame, which is employed for constrained parameters updating, and the ground-truth is used to initialize the key frame. Through the use of optical flow and target tracking, the video saliency detection range is narrowed and consequently the speed is accelerated. Besides, by means of location information, scale information and color model, the corresponding robustness and accuracy are also greatly improved. The flow diagram of our proposed STCVSD is shown in Fig. 1.

## 2 Calculation of location and scale information

### 2.1 Optical flow information with contour feature

The optical flow algorithm is used to obtain the optical flow field in each frame. According to the results of the optical flow field, there are two obvious facts in the motion region. On one hand, all pixels in the motion region are consistent and the region has obvious contour, but pixels in the non-moving regions are chaotic, and there is no visible contour of the object. On the other hand, the motion direction of the edge pixels in the motion region is very different from that of the non-moving regions in the neighborhood. Therefore, this paper needs to pretreat the optical flow according to the above facts, so as to obtain more effective optical flow regions [24].

First, the gradient of optical flow field is calculated. In this paper, the gradient value at the pixel  $q$  is defined as  $\vec{f}_q$ , which represents the speed of the pixel, and the intensity of the optical flow gradient at the pixel  $q$  can be expressed as

$$b_q^m = 1 - \exp\left(-\lambda^m \|\vec{f}_q\|\right) \quad (1)$$

where  $\lambda^m$  is the intensity coefficient.

According to the result of the gradient intensity  $b_q^m$ , it is possible to distinguish between pixels on the contours of most moving regions and pixels that are not on the contours. In particular, the pixel whose gradient is greater than a certain threshold  $T_m$  is sufficiently salient, so it can be immediately determined as the contour pixel of the moving region. For pixels whose gradient strength is less than  $T_m$ , their attribution need to be further determined by the difference of motion directions. So, the maximum value of the motion direction (angle) difference between

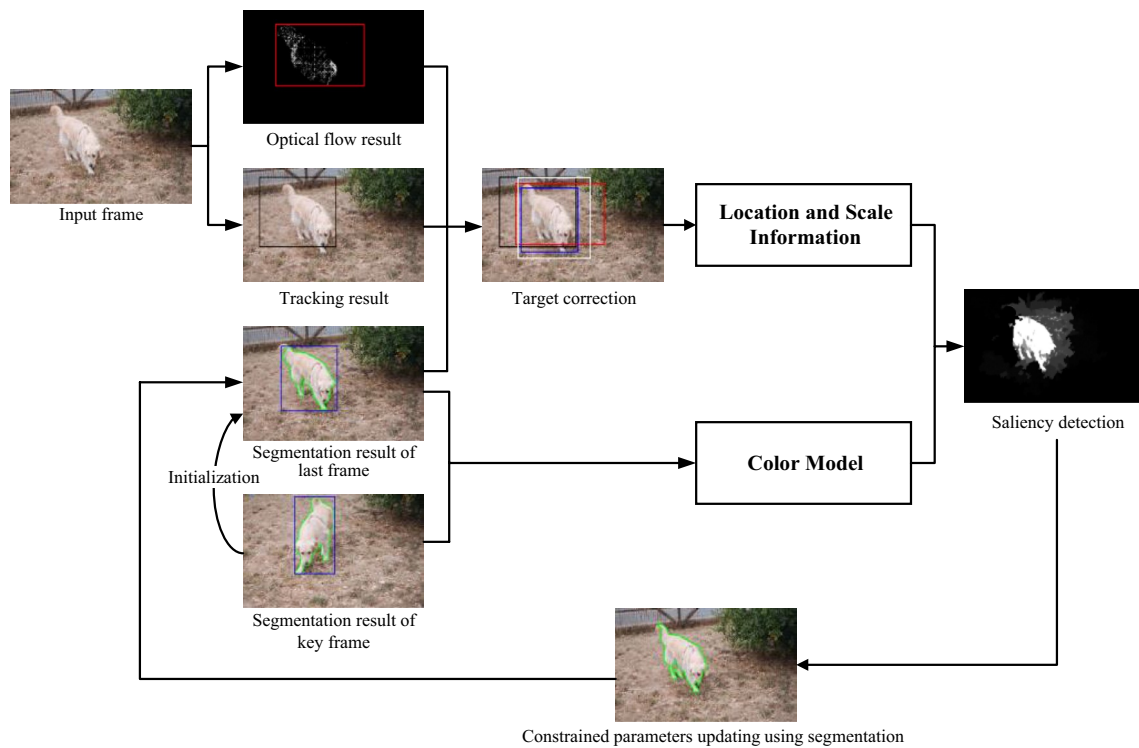


Fig. 1 The flow diagram of our proposed STCVSD

each pixel and the neighborhood pixels also needs to be calculated as

$$b_q^\theta = 1 - \exp\left(-\lambda^\theta \max_{q' \in Q_q} \delta\theta_{q,q'}^2\right) \tag{2}$$

where  $\delta\theta_{q,q'}^2$  represents the angle difference between pixel  $q$  and pixel  $q'$ ,  $Q_q$  represents the neighborhood of pixel  $q$ , and  $\lambda^\theta$  is the angle difference coefficient.

According to the gradient intensity value of the optical flow field and the difference of motion directions between the pixels, the difference of pixel velocity can be got as follows.

$$b_q = \begin{cases} b_q^m, & \text{if } b_q^m > T_m \\ b_q^m * b_{q'}^\theta, & \text{if } b_q^m \leq T_m \end{cases} \tag{3}$$

Because the value of threshold  $T_m$  is different in different scenes, iterative best threshold based on histogram [25] is used to get the threshold  $T_m$ . For pixels whose  $b_q$  is larger than 0.5, we identify them as pixels belonging to moving region contours.

According the above method, a rough contour of the motion region can be obtained. In a video scene where the distant view and the near view alternate, the noise interference in the rough contour cannot be eliminated by the morphological processing [26]. Therefore, more accurate

motion information needs to be further extracted on the basis of the rough contour image. It has been observed that there are usually no obvious contours of the noise, so the noise can be removed according to the contour feature. Specifically, we take any pixel in the video frame as the benchmark, start from 12 o'clock direction, and launch a ray clockwise every 45 degrees. A total of 8 rays can be obtained, and the intersecting times between the rays and the contours of the pixels are counted. Obviously, when the pixel is in the enclosed area, the rays and the contour should be intersected odd times, as in Fig. 2a. When the pixel is located outside the enclosed area, the rays intersects with the contour should be even times, such as Fig. 2b. If there are more than 4 rays from the pixel with an odd number of times intersected by the contour, the pixel is judged to be the inner pixel of the contour, and it is regarded as an effective moving pixel. Otherwise, the pixel is judged to be an outer pixel and a noisy pixel. The integral graph algorithm [27] is used to get the result quickly. After this processing, the optical flow motion information in the video frame can be obtained. Some examples are shown in the first column of Fig. 4.

### 2.2 Scale-variable KCF with APCE

Because the optical flow is easily affected by the motion background, even if contour feature is used to eliminate





**Fig. 3** Comparison between KCF and scale-variable KCF with APCE. The yellow rectangles represent groundtruth, the green rectangles represent the tracking results of the original KCF, and the black rec-

tangles represent the tracking results of our proposed scale-variable KCF with APCE

In the process of target detection, the feedback of the tracking results is often used to determine whether the model needs to be updated. In this section, the average peak correlation energy metric (APCE) [29] is introduced into the KCF tracking algorithm. And the tracking model is not updated when the target is obscured or the target is temporarily not tracked. APCE reflects the volatility of the response map and the confidence coefficient of the target detection. If the wave peak is sharp and the noise is small, the target obviously appears in the detection range, and the APCE value will become larger. If the target is obscured or disappearing, the APCE value will be significantly smaller. The calculation process is as Algorithm 1, and several examples of tracking are shown in Fig. 3.

### 2.3 Target correction

The target location and scale information cannot be obtained accurately through the target tracking algorithm or the optical flow algorithm simply. For example, when the movement of the target region is slow, the optical flow field is chaotic and disordered, while the tracking algorithm can effectively get the location and scale of the foreground. When the target region is moving intensely, the tracking algorithm is easy to lose or offset the target. Therefore, optical flow motion information and the segmentation result of the last frame are used to verify and correct the tracking result after obtaining the target location and scale information by scale-variable KCF with APCE preliminarily.

ALGORITHM II

THE CORRECTION PROCESS OF TARGET'S LOCATION AND SCALE

**Input:**

The result of optical flow,  $O$ ;  
 Tracking result,  $R$   
 Bounding-box of last frame's segmentation binary,  $F$ ;  
 Reserve scale,  $\theta$ , set as 1.05 usually.

**Output:**

Correction box of target,  $B$

```

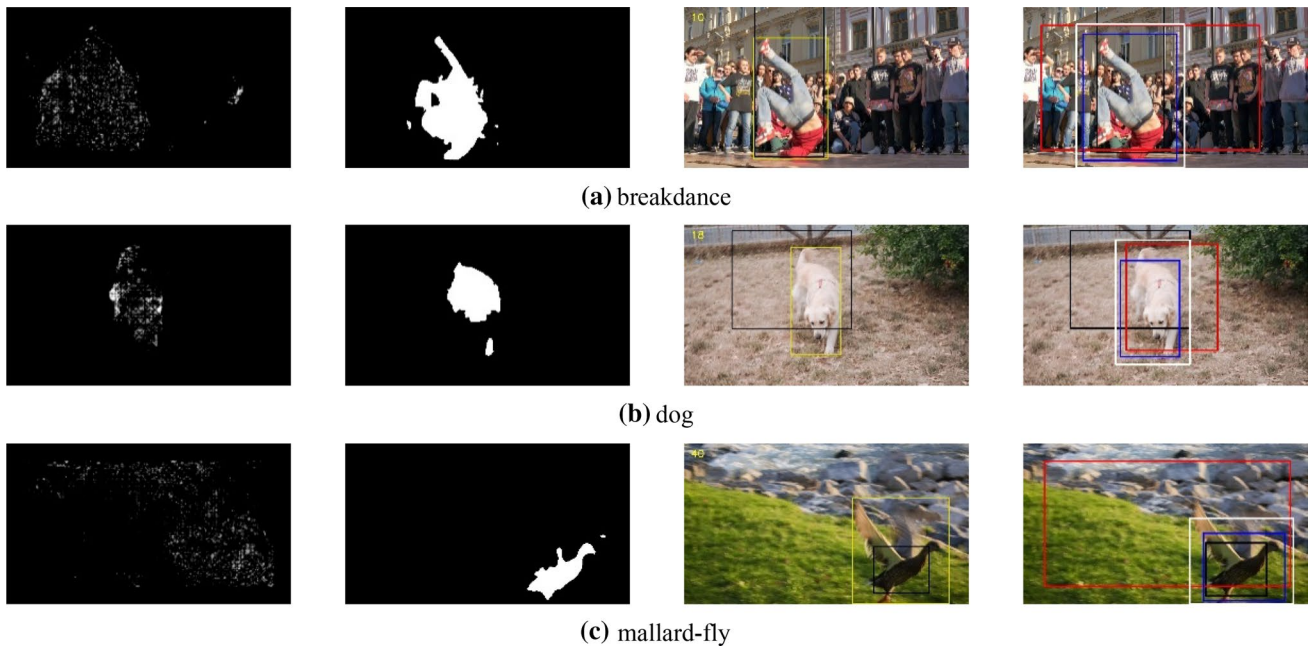
1:  if  $O$  locates in  $R$  mostly then
2:    compute motion box  $C$  via  $O$ 
3:    if  $C \approx F$  or  $R \approx F$  then
4:       $B_t \leftarrow$  the one more similar to  $F$  between  $C$  and  $R$ 
5:    else
6:       $B_t = F$ 
7:    end if
8:  else
9:    if  $R \approx F$  then
10:      $B_t = R$ 
11:   else
12:      $B_t = F$ 
13:   end if
14: end if
15:  $B = B_t * \theta$ 
16: return  $B$ 
    
```

First, the tracking result is compared with the optical flow motion contour, if the difference is small, the correction is continued in next step. Otherwise, the result of

video tracking is regarded as the correction result. Second, the tracking result and the optical flow result are compared with the segmentation results of the last frame respectively, and the one with smaller difference is judged to be the result of correction. If the difference of the two results is very large, the segmentation result of the last frame is used as the correction result. Finally, the correction result obtained by the above operation is enlarged to a certain proportion as an effective correction result. As a result, we can obtain the accurate target location and scale information. The correction process of target's location and scale is as Algorithm II, and some examples are shown in Fig. 4.

**3 Estimation of target color model**

In addition to the location and scale information of the target, the color model of the target is also estimated based on the existing target segmentation results. At first, an interactive segmentation is performed on the key frame, and the accurate target segmentation result is obtained. The color model of the target in key frame almost represents the color model of the target in the whole video. At the same time, it is noticed that the color model of the current frame target is the closest to the target color model of the last frame. Therefore, the color model of the key frame foreground object is used as the basic color model, and color model of the last frame is



**Fig. 4** The process and results of information correction for target location and scale. Optical flow results are shown in column 1, segmentation results of the last frame are shown in column 2, tracking

results of scale-variable KCF with APCE (black boxes) are shown in column 3, and the final correction results (white boxes) are shown in column 4

weighted into it to estimate the color model of the current frame.

A color model  $H_{fore-1}$  can be obtained by establishing a Gaussian mixture model for the target in the key frame. For the last frame, the color model of the foreground is  $H_{fore-2}$  and the color model of the background is  $H_{back-2}$ . The foreground color model  $H_{fore}$  in the current frame can be calculated as

$$H_{fore} = H_{fore-1} + \zeta H_{fore-2} \tag{4}$$

where  $\zeta$  is the weight coefficient of color model, and it is set to 0.3 in this paper.

Two consecutive frames in the same video usually have similar scenes. Therefore, the background color model in the last frame is used to estimate the background color model of the current frame.

$$H_{back} = H_{back-2} \tag{5}$$

According to the estimated foreground color model  $H_{fore}$  and background color model  $H_{back}$ , the probability  $H_{fore}(q_k)$  of pixel  $q_k$  belonging to the foreground and the probability  $H_{back}(q_k)$  of pixel  $q_k$  belonging to the background can be calculated. In order to better determine the probability of pixel label assignment, the obtained probability value of pixels belonging to foreground or background are both normalized as follows.

$$H_{fore-F}(q_k) = \frac{H_{fore}(q_k)}{H_{fore}(q_k) + H_{back}(q_k)} \tag{6}$$

$$H_{back-F}(q_k) = 1 - P_{fore}(q_k) \tag{7}$$

Foreground probability  $H_{fore-F}(q_k)$  of pixel  $q_k$  is obtained through (6), and the background probability  $H_{back-F}(q_k)$  is calculated by (7). As a result, the target color model constraint can be effectively estimated.

#### 4 Computation of strong target constrained video saliency

On the basis of the traditional image saliency, we extend the assumption that the closer the region to the center of the image is, the more salient the region is in image saliency. For the video saliency, according to the location and scale information, it is proposed that the closer the region to the target center is, the more salient the region is in this paper. And the target color model is further included here that the closer the region to the target color model is, the more salient it is. Therefore, video saliency  $S_v(r_k)$  can be obtained as

$$S_v(r_k) = w_s(r_k)w_o(r_k) \sum_{r_k \neq r_i} e^{\frac{D_s(r_k, r_i)}{-\sigma_s^2}} w(r_i)D_r(r_k, r_i) \tag{8}$$

where  $D_s(r_k, r_i)$  represents the centroid distance between region  $r_k$  and  $r_i$ , and  $D_r(r_k, r_i)$  represents the color distance between region  $r_k$  and  $r_i$ . It is supposed  $b$  as the target box,  $t$  is the empirical weight in which  $t_1, t_2, t_3$  is set as 1.5, 1.0 and 0.5 respectively.  $T$  is the empirical threshold for any region  $r_k$ , and  $T_1, T_2$  is set to 0.5 and 1.0 respectively. The interregional distance function  $dis(r_k, b)$  is used to obtain the distance weight  $w_s(r_k)$ .

$$w_s(r_k) = \begin{cases} t_1, & \text{if } dis(r_k, b) < T_1 \\ t_2, & \text{if } T_1 < dis(r_k, b) < T_2 \\ t_3, & \text{if } T_2 < dis(r_k, b) \end{cases} \tag{9}$$

A piecewise function is used here to increase the saliency values in the scale and to effectively reduce the saliency values outside the scale. Further, the weight  $w_o(r_k)$  of the color model can be obtained as

$$w_o(r_k) = \frac{1}{e^{-(H_{fore-F}(r_k)-0.5)}} \tag{10}$$

where the foreground probability value  $H_{fore-F}(r_k)$  of the region  $r_k$  is estimated based on the target color model.

#### 5 Constrained parameters updating using segmentation

In videos, the feature information that can be extracted is richer than that in the image, because of the relationship between the last frame and the current frame. In this paper,

the segmentation results of the target are extracted frame by frame, so the use of the segmented target information can provide more help for the calculation of video saliency. Due to the space-time context information between frames in video scene is closely related, the segmentation result of the last frame can provide effective information for the determination of target location and scale by correcting the tracking result of the current frame. Simultaneously, saliency results do not have connectivity and obvious boundaries, so existing segmentation results can be used to calculate the color model constraints of the target in the current frame.

The traditional video segmentation method usually simplifies the video segmentation problem into two parts: the extraction of the prior information and the target segmentation. The common video segmentation methods calculate the prior information according to the target color model, the contour constraint, the motion information and the simple saliency [30]. And the Graph-cut is usually selected to perform the segmentation operation. However, only fusing a particular feature or a simple feature cannot



**Fig. 5** Segmentation results in DAVIS. The green lines are results of our proposed method and the blue lines are the results of global Densecut

quickly extract effective prior information. At the same time, the segmentation method building graph models for all video frame pixels is inefficient.

Before calculating the location and scale information of the target in the current frame, we have obtained the saliency detection result of the last frame. Then the Densecut [23] is initialized with the obtained saliency map to calculate the segmentation result of the last frame, while the groundtruth is used to initialize the key frames. As a result, the unknown area is compressed, the computation of the target segmentation algorithm is reduced, and the accuracy of the segmentation result is improved. It also provides reference for accurate calculation of location and scale information at the same time. Parts of segmentation results are shown in Fig. 5.

## 6 Experimental analysis

### 6.1 Environment and dataset

For experimental analysis, the Visual Studio 2013 and the OpenCV image library are selected as the development tools and the experiment program is written in C++. The program runs in the hardware environment of Intel (R) Xeon (R) CPU E5-2699 v3, 128 GB RAM.

For dataset, the DAVIS [31] is selected as the test case. The DAVIS contains 50 test videos, including a variety of challenging video segmentation test sets such as occlusions, motion blur and appearance changes. And the DAVIS also comes with the standard segmentation results of all 50 videos, which are the manually calibrated Ground-Truth. Under this dataset, the saliency detection experiments are carried out on STCVSD (our method), RC [15],



PISA [32] and CA [33], and the experimental results are compared and analyzed qualitatively. In addition, the segmentation results that are produced during saliency detection process can also be compared with the state-of-art video segmentation methods including BVS [34], CVOS [35], FCP [36] and FST [24] on the DAVIS quantitatively. Therefore, DAVIS is chosen as the data set in the following experiments.

## 6.2 Results and analysis

### 6.2.1 Verification of scale-variable KCF with APCE

Experiments on KCF algorithm and scale-variable KCF with APCE proposed in Sect. 2.2 are carried out. The improved KCF algorithm has better tracking results compared with the original KCF for the video whose target is from far to the near or from the near to the distant. In Fig. 3, the tracking results of the several frames in car-shadow, drift-straight and motocross-bumps are given. The yellow rectangles in images are groundtruth, the green rectangles are the tracking results of the original KCF, and the black rectangles are the tracking results of our proposed scale-variable KCF with APCE.

The target in car-shadow as Fig. 3a is from the near to the distant. The original KCF is not able to reduce the scale of the tracking box in time when the target is far away, resulting in that green boxes are too large in the second image and the third image in Fig. 3a. As our improved method increases the possible forms of scale change, the scale of the tracking box can be reduced in time, and the target is located more accurately. The target in drift-straight as Fig. 3b, motorbike as Fig. 3c is from far to the near. For the original KCF algorithm, the smaller scale of the tracking box located in the first frame does not adjust in time, so that the target cannot be tracked or only a small part of the target can be located. While our improved KCF can complete enveloping the tracking targets (black rectangles) by adjusting the height and width separately. The original KCF tracking result (green rectangle) of the third image in Fig. 3b is far from the real target, and APCE is used to exclude the situation of losing target in our improved KCF so that our result is still accurate as the black rectangle. It can be proved that our proposed scale-variable KCF with APCE has certain advantages in determining the location and scale of the target.

### 6.2.2 Target location and scale correction

In Fig. 4, it shows the process and results of information correction for target location and scale, including optical flow binary images, segmentation results of last frame,

improved KCF tracking boxes and the correction results. For improved KCF tracking results, the yellow rectangles represent groundtruth, and the black rectangles are our tracking results. In correction process, the red box is the target box obtained directly from the result of the optical flow, which may quite different from the real target. The blue box is the bounding rectangle of last frame segmentation result, and the white box is the final result after correcting through the method mentioned in Sect. 2.3. According to the optical flow results of Fig. 4a, c, the result of optical flow is easy to be affected when there some motion interference in the background, so that the red box is far larger than the real larger. For the segmentation mask of the last frame as shown in Fig. 4b, c, if the segmentation result of the last frame is incomplete, the blue box cannot be accurately located in the current frame to consider the offset of the motion. According to results of improved KCF of Fig. 4b, c, even though the tracking algorithm has been improved, it is still not sensitive enough to the drastic scale change, so the black box in Fig. 4b is larger than the target, and the black box in Fig. 4c is smaller than the target. Finally, the improved KCF tracking results are corrected by combining the optical flow result and the last frame segmentation result, and the appropriate target location and scale corrections (white boxes) are obtained. The final accurate target location and scale information not only provide accurate feature information for saliency detection, but also greatly compress the region to be detected. Most of the background regions are eliminated, and then the accuracy and efficiency of our proposed video saliency detection algorithm are further improved owing to the reduction of redundant information.

### 6.2.3 Video saliency detection with strong target constraints

**6.2.3.1 Qualitative analysis** Accurate foreground color estimation models are obtained through the computation of video saliency, so accurate segmentation results can also be acquired. In Fig. 5, segmentation results (green lines) that are produced during saliency detection are given to compare with the results (blue lines) of global Densecut. According to the segmentation results of bear video sequences as Fig. 5a, we can find that for the videos with small differences between foreground and background, a high quality of segmentation is achieved for our proposed method and it avoids the jitter of segmentation results. Besides, for the video scenes of target occlusion (such as the bus in Fig. 5b and lucia in Fig. 5c and fast moving (such as the car-roundabout in Fig. 5d and paragliding-launch in Fig. 5e), global Densecut without saliency maps is easy to segment the background into the foreground according to the results (blue lines), because of the lack



**Fig. 6** Comparisons of video saliency detection results with other saliency detection methods, **a** the original image, **b** our STCVSD, **c** RC [15], **d** PISA [32], and **e** CA [33]

of more accurate target constrained information, such as location, scale, color model and shape.

Aiming at the comparisons of video saliency detection, in Fig. 6, the results of our proposed STCVSD method are obtained to compare with the results of other saliency detection algorithms, which are the state-of-art methods on the basis of traditional algorithms, such as RC [15], PISA [32], and CA [33]. From the results of Fig. 6c–e, we can find that these compared existing state-of-art saliency extraction algorithms are all mostly in a global chaotic trend. In contrast, after using the strong target information constraints as proposed in our STCVSD method, as shown in

Fig. 6b, the saliency is rapidly rising in the target scale to obtain more accurate saliency detection results. The reason for that is the traditional saliency detection methods usually have no definite scale restriction, maybe just some weak central constraints are included, so the corresponding results of them cannot effectively represent the saliency regions when the current target is not in the image center or the target is similar to the background. However, the results of our proposed STCVSD method show a connectivity trend because of the strong constraints of target location, scale information and color model.

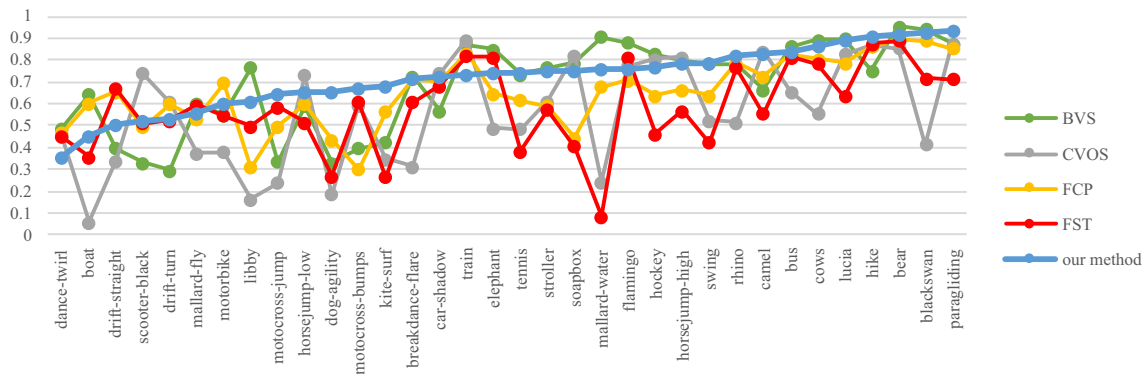
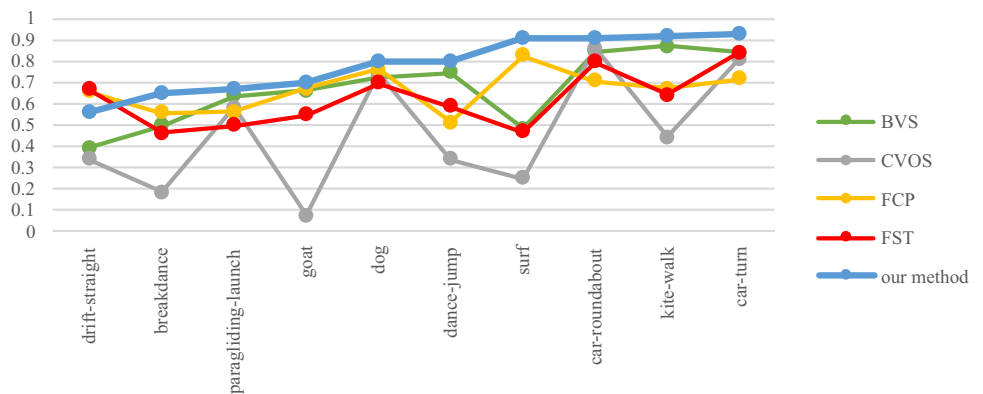


Fig. 7 Experimental results and quantitative comparisons (the middle frame as the key frame)

Fig. 8 Experimental results and quantitative comparisons (the first frame as the key frame)



**6.2.3.2 Quantitative analysis** Due to the lack of groundtruth for video saliency, the quantitative analysis is not feasible. The quantitative data of segmentation results in the DAVIS is calculated to compare with the state-of-art video segmentation methods such as BVS [34], CVOS [35], FCP [36], FST [24], so that the validity of STCVSD can be verified.

The evaluation criteria IoU (Intersection over Union) in the field of image/video segmentation is selected in this paper as (11), where  $Area_V$  represents the obtained segmentation result, and  $Area_{GT}$  represents the foreground in Ground-Truth. The average of IoU is calculated in (12), where  $IOU_i$  is the IoU value of the  $i$ th video,  $Num_i$  represents the number of available frames (except the key frame) in the  $i$ th video, and  $videos$  is the set of test videos.

$$IoU = \frac{Area_V \cap Area_{GT}}{Area_V \cup Area_{GT}} \tag{11}$$

$$IoU_{average} = \frac{\sum_{i \in videos} IoU_i \times Num_i}{\sum_{i \in videos} Num_i} \tag{12}$$

Table 1 Averages of quantitative comparisons (the middle frame or the first frame as the key frame)

	Our method	BVS	CVOS	FCP	FST
MAverage	0.728	0.716	0.585	0.662	0.595
FAverage	0.706	0.682	0.459	0.659	0.619

For different test videos, the experiments are carried out by selecting the middle frame and the first frame as the key frame respectively. For most videos, selecting the middle frame as the key frame can improve the accuracy of the tracking algorithm and reduce the influence of the far frame on the current frame. There is a certain improvement in the effect of saliency extraction and segmentation.

In Fig. 7, the IoU values of our proposed method by selecting the middle frame as the key frame are shown, and they are compared with other segmentation algorithms. It can be seen that although the accuracy of other methods in a few videos is higher than ours, they usually go up and down dramatically. By contrast, the accuracy of our method undulates slightly, and the overall performance is more stable. In other words, our method

has superior performance in terms of adaptability and robustness.

Since the targets of a few videos are incomplete in the middle frame because of the occlusion, blur, drawing and so on, it is not appropriate to select the middle frame as the key frame for them. In order to reduce the artificial interaction and simplify the algorithm, the first frames of these videos are selected as the key frames, and the quantitative comparisons of IoU results are drawn in Fig. 8. It is clearly shown that our method in this experiment has obvious advantages for almost all the videos, in addition to being more universal and stable, in general, our method also achieves the best accuracy.

As shown in Table 1, the MAverage denotes the IoU average value of the videos with the middle frame as the key frame, which is 0.728, exceeding the BVS algorithm 0.12. And the FAverage denotes the IoU average value of the videos with the first frame as the key frame, which is 0.706, exceeding the BVS algorithm 0.024. Besides, compared with the deep learning based RFC-VGG [37] whose IoU is 0.6984, our method is also more superior. Therefore, our segmentation results are better than these state-of-art video segmentation algorithms on the whole. And it is also proved that the STCVSD method proposed in this paper is effective.

## 7 Conclusion

In this paper, a video saliency detection method based on strong target constraints is proposed by fusing location, scale and color information. Traditional optical flow algorithm is included to extract contour features, KCF tracking method is improved to be scale-variable and APCE is used to enhance accuracy, and these improved algorithms are used to correct location and scale of the target with previous segmentation results. The color model is calculated from the segmentation results of the key frame and the last frame. Finally, these location, scale information and the color model are fused to constrain the saliency calculation. According to experimental results, our proposed video saliency detection method STCVSD can effectively extract the real saliency region in video sequences. For the qualitative results, our method shows a connectivity trend which is better than other saliency detection methods, and the intermediate segmentation results are superior to Densecut without saliency. For the quantitative results, the average IoU of our segmentation results is higher than other state-of-art video segmentation methods including the deep learning based method on DAVIS dataset. As a result, our proposed STCVSD method is verified to be excellent.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China under Grant Nos. 61703317 and 61105006; the Shenzhen Strategic Emerging Industry Development Special Fund under Grant No. JCYJ20170307172130906; the Aerospace Science and Technology Foundation under Grant No. 2018-HT-HZ; the Open Fund of Key Laboratory of Image Processing and Intelligent Control (Huazhong University of Science and Technology), Ministry of Education under Grant No. IPIC2019-01; the Fundamental Research Funds for the Central Universities under Grant Nos. WUT:2018IVB072, 2018IVA110.

## Compliance with ethical standards

**Conflict of interest** The author(s) declare that they have no conflict of interests.

## References

1. Latif A, Rasheed A, Sajid U et al (2019) Content-based image retrieval and feature extraction: a comprehensive review. *Math Probl Eng* 2019, Article ID 9658350
2. Ali N, Zafar B, Iqbal MK et al (2019) Modeling global geometric spatial information for rotation invariant classification of satellite images. *PLoS ONE* 14(7), Article ID e0219833
3. Ratyal N, Taj IA, Sajid M et al (2019) Deeply learned pose invariant image analysis with applications in 3D face recognition. *Math Probl Eng* 2019, Article ID 3547416
4. Sajid M, Iqbal Ratyal N, Ali N et al (2019) The impact of asymmetric left and asymmetric right face images on accurate age estimation. *Math Probl Eng* 2019, Article ID 8041413
5. Sajid M, Ali N, Dar SH et al (2018) Data augmentation-assisted makeup-invariant face recognition. *Math Probl Eng* 2018, Article ID 2850632
6. Mannan SK, Kennard C, Husain M (2009) The role of visual salience in directing eye movements in visual object agnosia. *Curr Biol* 19(6):R247–R248
7. Ayoub N, Gao Z, Chen D, Tobji R, Yao N (2018) Visual saliency detection based on color frequency features under Bayesian framework. *Ksii Trans Internet Inf Syst* 12:676
8. Wolfe JM, Horowitz TS (2004) What attributes guide the deployment of visual attention and how do they do it? *Nat Rev Neurosci* 5:495
9. Teuber HL (1965) *Physiological psychology*. Mcgraw-Hill, New York
10. Wang X, Zhong Y, Xu Y, Zhang L, Xu Y (2017) Saliency-based endmember detection for hyperspectral imagery. In: 2017 IEEE international geoscience and remote sensing symposium (IGARSS), p 984
11. Yamazaki T, Hasebe N, Shimizu S (2017) Considerations about saliency map from Wide Angle Fovea image. In: 2017 IEEE 26th international symposium on industrial electronics (ISIE), p 1330
12. Zhang J, Li B, Dai Y, Porikli F, He M (2018) Integrated deep and shallow networks for salient object detection. In: IEEE international conference on image processing, p 1537
13. Fu Y, Cheng J, Li Z et al (2008) Saliency cuts: an automatic approach to object segmentation. In: International conference on pattern recognition. *DBLP*, pp 1–4
14. Liu T, Sun J, Zheng NN, Tang X, Shum HY (2007) Learning to detect a salient object. In: 2007 IEEE conference on computer vision and pattern recognition, p 1

15. Cheng MM, Mitra NJ, Huang X, Torr PHS, Hu SM (2015) Global contrast based salient region detection. *IEEE Trans Pattern Anal* 37:569
16. Du B, Ma L, Zhuang Y, Chen H, Soomro NQ (2017) Moving target detection via hierarchical spatiotemporal saliency analysis. In: 2017 IEEE international geoscience and remote sensing symposium (IGARSS), p 1840
17. Jian M, Qi Q, Dong J, Sun X, Sun Y, Lam KM (2016) Saliency detection using quaternionic distance based weber descriptor and object cues. In: 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), p 1
18. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition, p 1597
19. Li S, Lee MC (2007) Fast visual tracking using motion saliency in video. In: 2007 IEEE international conference on acoustics, speech and signal processing—ICASSP'07, p 1073
20. Kulshreshtha A, Deshpande AV, Meher SK (2013) Time-frequency-tuned salient region detection and segmentation. In: 2013 3rd IEEE international advance computing conference (IACC), p 1080
21. Xue K, Wang X, Ma G, Wang H, Nam D (2015) A video saliency detection method based on spatial and motion information. In: 2015 IEEE international conference on image processing (ICIP), p 412
22. Borji A, Cheng MM, Hou Q, Jiang H, Li J (2017) Salient object detection: a survey. *Eprint Arxiv* 16:3118
23. Cheng MM, Prisacariu VA, Zheng S, Torr PHS, Rother C (2015) DenseCut: densely connected CRFs for realtime grabcut. *Comput Graph Forum* 34:193
24. Papazoglou A, Ferrari V (2013) Fast object segmentation in unconstrained video. In: 2013 IEEE international conference on computer vision, p 1777
25. Yi-Quan WU, Wen-Yi WU, Zhe P (2009) A fast iterative algorithm of the Otsu threshold based on two-dimensional histogram oblique segmentation. *J Eng Graph* 30:89
26. Delaye A, Anquetil E (2010) Learning spatial relationships in hand-drawn patterns using fuzzy mathematical morphology. In: 2010 International conference of soft computing and pattern recognition, p 162
27. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, p 511
28. Henriques JF, Caseiro R, Martins P, Batista J (2015) High-speed tracking with kernelized correlation filters. *IEEE Trans Pattern Anal* 37:583
29. Wang M, Liu Y, Huang Z (2017) Large margin object tracking with circulant feature maps. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR)
30. Tsai YH, Yang MH, Black MJ (2016) Video segmentation via object flow. In: IEEE conference on computer vision and pattern recognition. IEEE, pp 3899–3908
31. Perazzi F, Pont-Tuset J, McWilliams B, Gool LV, Gross M, Sorkine-Hornung A (2016) A benchmark dataset and evaluation methodology for video object segmentation. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), p 724
32. Wang K, Lin L, Lu J, Li C, Shi K (2015) PISA: pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence. *IEEE Trans Image Process* 24:3019
33. Goferman S, Zelnik-Manor L, Tal A (2012) Context-aware saliency detection. *IEEE Trans Pattern Anal Mach Intell* 34:1915
34. Märki N, Perazzi F, Wang O, Sorkine-Hornung A (2016) Bilateral space video segmentation. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), p 743
35. Taylor B, Karasev V, Soatto S (2015) Causal video object segmentation from persistence of occlusions. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), p 4268
36. Perazzi F, Wang O, Gross M, Sorkine-Hornung A (2015) Fully connected object proposals for video segmentation. In: 2015 IEEE international conference on computer vision (ICCV), p 3227
37. Valipour S, Siam M, Jagersand M, Ray N (2017) Recurrent fully convolutional networks for video segmentation. In: Applications of computer vision, p 29

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.