Self-propagating video segmentation using patch matching and enhanced OneCut

Shoudong Han^{1,2} · Chenfei Xia¹ · Yujun Liu¹ · Yingchun Yang¹

Received: 17 October 2019 / Accepted: 10 January 2020 / Published online: 18 January 2020
© Springer Nature Switzerland AG 2020

Abstract

In this paper, a self-propagating video segmentation approach based on patch matching and enhanced OneCut is proposed, which takes full advantage of the target's color feature, shape feature, and motion information. Firstly, an interactive key frame segmentation is performed to obtain the accurate initial contour of target. Secondly, some sampling patches are uniformly selected along the target contour of previous frame to initialize the localized classifiers. Afterwards, the patch matching is used to pass this contour to the current frame. Simultaneously, the localized classifiers are moved to current frame similarly, and their corresponding positions and parameters are also updated. Eventually, the foreground and background probability maps of current frame are calculated through the localized classifiers as well as global probability models, and then the enhanced OneCut model is constructed to obtain its segmentation result. Compared with the state-of-the-art video segmentation methods, our proposed approach performs outstandingly on the DAVIS dataset.

Keywords Video segmentation · Localized classifier · Patch matching · Enhanced OneCut

1 Introduction

Video sequences can assist in the research of tasks such as face recognition [1–3], scene segmentation, target tracking, and can also promote the development of hybrid geometric spatial image representation methods [4, 5]. Meanwhile, video segmentation is also one of the tasks which has important research value. Video segmentation aims at obtaining the binary labeling of target from videos. It is playing an important role for many higher-level machine vision fields, such as target identification [6], target tracking [7–9], video retrieval [10], video compression [11] and video analysis [12], etc. For instance, video segmentation can provide prior information for target identification, and it can improve the efficiency in image compression. However, owing to the characteristics of diversity and

complexity in different video segmentation scenes, there is no general video segmentation solution at present.

According to the different supervision ways, the existing video segmentation methods can be divided as the unsupervised video segmentation method and interactive one. Most of unsupervised video segmentation methods are based on optical flow algorithm and gradient descent method without requiring any user interaction. This kind of method is well-adapted to the video in which moving object exists. For example, as in [13], the closed contour of moving target can be obtained according to the gradient changes of optical flow field, by means of calculating the optical flow of previous frame and current frame, and then using the internal and external mapping algorithm. However, when the target is fixed or moving slowly or moving alternately, it is impossible for this kind of method to detect the possible region of target just by depending on

✉ Shoudong Han, shoudonghan@hust.edu.cn | ¹Key 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. ²Research Institute of Huazhong University of Science and Technology in Shenzhen, Shenzhen 518057, China.

motion information, thus the segmentation fails. Besides, it is also not applicable to the multi-contour target segmentation. Therefore, some unsupervised video segmentation methods based on target saliency detection and tracking are proposed. Such as the method in [14], which need pre-calculate the position of salient target in video and then passes the segmentation result frame by frame by means of target tracking.

Compared with unsupervised video segmentation methods, interactive ones can perfectly solve the problems which are mentioned above. Under some special conditions, such as during the initialization phase or in case of segmentation refining, the interactive video segmentation method usually requires to learn some prior information or corrective information from user interactions. Such as the Snapcut [15], which moves the segmentation contour of target frame by frame using localized classifiers, similar to us. However, the approach proposed in this paper is distinct from it. In the tracking part, we adopt patch matching method to track the contour of target directly instead of using affine transformation by transmitting the segmentation result of the previous frame to the current frame in the way of self-propagation, and then we additionally use optical flow to adjust the contour. In the segmentation part, our approach combines the global possibility models with the localized classifiers, and then the enhanced OneCut model is constructed to obtain final segmentation result by taking full advantage of the target's color feature, shape feature, and motion information. Enhanced OneCut model cannot only improve the corresponding segmentation performance when compared with the original OneCut model [16], it is also superior to the Grabcut model [17] which Snapcut adopted. The flow chart of our proposed video segmentation approach is shown in Fig. 1. In order to ensure effective detection of sufficient amplitude motion, the average inter-frame difference threshold is set to 0.3, and frames which inter-frame difference higher than the threshold are selected as key frames of the video.

2 Global and localized classifier modeling

The main goal of video segmentation is to extract the meaningful targets (such as human, animals or vehicles) from the background in the video. For the sake of ensuring the integrity of the target segmentation, global classifier modeling methods [18, 19] are usually employed. However, video segmentation is different from the other target extraction methods (such as target identification, target detection, or target tracking), which requires an accurate labeling for each pixel to determine whether it belongs to foreground or background. Thus, it is equally essential to guarantee the accuracy of extracted target contour, and localized classifier modeling methods [20, 21] are usually helpful for taking into account some local characteristics of target. Therefore, in this paper, the localized classifiers as well as global probability models are included simultaneously for foreground/background modeling, where Gaussian Mixture Model (GMM) is used to represent the foreground/background color distributions based on obtained segmentation result, and the position information of obtained target contour is used to represent their shape priors.

2.1 Constructing global classifier model

The to-be-segmented image sequences are arranged according to the timeline. It is supposed that there're n localized classifiers $M_1^t, \dots, M_n^t$ and a global classifier M_0^t corresponding to the target contour in t -th frame. Each localized classifier takes as input one patch on the contour of the object so there're n square sampling patches are uniformly selected along this target contour to initialize the localized classifiers. The segmentation result of the target at the t -th frame is set as S^t , and the width of each localized classifier is set as r , which is the number of pixels corresponding to the edge length of each square, reflects

Fig. 1 The flow chart of our proposed video segmentation approach

the size of the classifier. The longer the side, the richer features it contains. So the width r is conformed according to the proportion of the target in the whole image. The ratio, which is the ratio of the number of pixels belong to foreground to the number of pixels in the whole image, is set as λ . If $y_{\min}$ is the width of the image, r is described as formula (1). Because the boundary value will affect the accuracy of localized classifiers and the richness of the described features, we choose 60 as the upper bound and 20 as the lower bound in order to get more color features and get more details of the contour.

$$r = \begin{cases} 60, & \lambda \cdot y_{\min} \geq 60 \\ \lambda \cdot y_{\min}, & 20 \leq \lambda \cdot y_{\min} < 60 \\ 20, & \lambda \cdot y_{\min} < 20 \end{cases} \quad (1)$$

Supposing that $G_f^{key}(0)$ and $G_b^{key}(0)$ represent global foreground GMM and global background GMM respectively, the process of fusing the shape information is defined as follows: S^{key} is supposed as the segmentation result of the key frame, and $S^{key}(x) = 1$ represents that the pixel x is set as foreground in the key frame, $S^{key}(x) = 0$ represents that the pixel x is set as background in the key frame. Moreover, $F = \{x | S^{key}(x) = 1\}$ is regarded as the original foreground points set, and $B = \{x | S^{key}(x) = 0\}$ is regarded as background points set. Clustering is performed to the two sets respectively (kmeans++ [22] is adopted in our method), and in order to achieve as high accuracy as possible, 5 Gaussian distributions are chosen to describe the GMM of global color probability model according to experience. Supposing that $t + 1$ -th frame is current frame, $G_f^t(0)$ and $G_b^t(0)$ represent global foreground/background GMM updated by the segmentation results of the previous frame (the t -th frame), thus, the present global probability $p_{global}^1(x)$ and $p_{global}^0(x)$ of pixel x belonging to the foreground/background can be obtained. In order to satisfy the normalization of the global probability model and the localized classifier, the effective global foreground/background probabilities $p_{global}^1(x)$ and $p_{global}^0(x)$ need to be normalized by the sum of their foreground probabilities and the background probabilities.

2.2 Constructing localized classifier model

The localized classifier model constructed in this paper includes both the color probability model and the shape information of the target. Assuming that the $t + 1$ -th frame is the current frame, the color probability modeling and parameter updating process of the localized classifier are defined as follows: First, local the foreground/background GMM is initialized by the segmentation results of the previous frame (the t -th frames). $G_f^t(k)$ is set as foreground GMM

of the k -th localized classifier, and $G_b^t(k)$ represents the background GMM of the k -th localized classifier, $k \in [1, n]$, 3 Gaussian distributions are chosen empirically for the local GMM. Then, the patch matching, which will be proposed in Sect. 3, is used to obtain the initial contour of the target in the current frame and the location of the localized classifier in the current frame, and update the foreground/background GMM $G_f^{t+1}(k)$ and $G_b^{t+1}(k)$ in the current frame. So, for the k -th localized classifier in current frame, the local probability $\tilde{p}_f^{t+1}(x|k)$ that the pixel x belongs to foreground can be calculated as followed (the calculation of local probability $\tilde{p}_b^{t+1}(x|k)$ that pixel x belongs to background is similar):

$$\tilde{p}_f^{t+1}(x|k) = (1 - \xi_c) L^{t+1}(x|k) + \xi_c \tilde{p}_f^{t+1}(x|k) \quad (2)$$

In formula (2), $L^{t+1}(x|k)$ represents the estimated foreground/background label after the patch matching. And, if the pixel is estimated as foreground, the value is 1, otherwise the value is 0. $\tilde{p}_f^{t+1}(x|k)$ is a relative local foreground color probability, and it can be calculated as follows: $\tilde{p}_f^{t+1}(x|k) = p_f^{t+1}(x|k) / (p_f^{t+1}(x|k) + p_b^{t+1}(x|k))$, in which $p_f^{t+1}(x|k)$ and $p_b^{t+1}(x|k)$ are calculated by local foreground/background GMM $G_f^{t+1}(k)$ and $G_b^{t+1}(k)$ of current frame. ξ_c is the weight that the color constraint is relative to the shape constraint in the localized classifier model.

$$\xi_c = 1 - \frac{\int_{M_k} |L^{t+1}(x|k) - \tilde{p}_f^{t+1}(x|k)| \cdot w(x) dx}{\int_{M_k} w(x) dx} \quad (3)$$

In formula (3), $w(x)$ is a weight influenced by the distance, and it can be calculated as formula (4):

$$w(x) = \exp(-d^2(x)/\sigma_c^2) \quad (4)$$

In formula (4), $d(x)$ is the distance from the pixel x to the nearest contour. The closer the pixel x to the target contour is, the greater influence its color probability model has, so $w(x)$ is greater. σ_c is the empirical value, and here it is set to 0.5 of the localized classifier's size.

After the patch matching, a pixel of the current frame may exist in several localized classifiers simultaneously. So, the local probability $p_{local}^1(x)$ and $p_{local}^0(x)$ that pixel x belongs to foreground/background are obtained by superimposing all the related localized classifiers.

$$p_{local}^i(x) = \begin{cases} \frac{\sum_{k \in K} \tilde{p}_f^{t+1}(x|k) \cdot (|x - c_k| + \delta)^{-1}}{\sum_{k \in K} (|x - c_k| + \delta)^{-1}}, & i = 1 \\ \frac{\sum_{k \in K} \tilde{p}_b^{t+1}(x|k) \cdot (|x - c_k| + \delta)^{-1}}{\sum_{k \in K} (|x - c_k| + \delta)^{-1}}, & i = 0 \end{cases} \quad (5)$$

In formula (5), K represents a series of localized classifiers that cover pixel x , c_k is the center of the k -th classifier, δ is a very small constant and is set to 0.01 in our method.

3 Contour tracking based on patch matching

Because of the great sensitivity of single point sparse matching, especially for the matching of moving target contour and the background near it, the patch matching algorithm is proposed in this paper by extending the single sparse point matching. Patch matching takes full advantage of the accurate segmentation results from the previous frame, and uses the small range patch matching of each (foreground) point in the target contour to achieve the target contour tracking. Therefore, it overcomes the problem of poor robustness brought by single point sparse matching, and can greatly reduce the mismatching points and achieve significant contour tracking effect. For Snapcut [15], the transformation of contour has two steps: First, the approximate contour of the target in the current frame is obtained through affine transformation. Second, the accurate position of the sampling patch can be got through local optical flow adjustment. Affine transformation only predicts the approximate contour of the target which performs well in tracking of target with small local deformation and consistent motion. However, if the local motion information of the target is uneven or the local contour changes are inconsistent, the affine transformation is not accurate enough to cover all variations of the local contour, leading to that the initial contour is not in the neighborhood of the real contour.

We suppose t -th frame as the previous frame, and $t+1$ -th frame as the current frame. The neighborhood of to-be-matched contour point x^t is U^t , which is set as a 5×5 patch, and U_f^t is the foreground of U^t . The position $L(x^{t+1})$ of patch matching is described as follow:

$$L(x^{t+1}) = \sum_{x_i^t \in U_f^t} T(L(x_i^t)) / \sum_{x_i^t \in U_f^t} 1 \quad (6)$$

The formula (6) describes the process of target contour point tracking based on patch matching, which means to match the corresponding target contour point in the current frame. In formula (6), $L(x_i^t)$ represents the position of a pixel x_i^t in U^t , and U^t is the neighborhood of to-be-matched contour point x^t . And T is the position transformation of the pixel point in the previous frame to the most familiar pixel point in the current frame: $T: L(x_i^t) \rightarrow L(x_i^{t+1})$. For every pixel x_i^t in U^t , its position in the current frame $L(x_i^{t+1})$ can be obtained through Surf sparse matching [23].

Since the center of the sampling patch locates at the contour of the target, the transformation of contour points in formula (6) includes the position transformation of the local classifier. It is supposed that $L(x^t|k)$ represents the initial position of the k -th sampling patch in the t -th frame. Through patch matching described by formula (6), then we can further get the approximate position $L(x^t|k)$ of the k -th sampling patch in the $t+1$ -th frame, $k \in [1, n]$.

4 Enhanced Onecut segmentation model

In recent years, video segmentation method based on graph cuts framework [24–27] has become a research hotspot. The energy function used in it can fully fuse the region distinction, the edge feature in the image frame and the motion information of the target, and it also has good extendibility. In order to segment the foreground accurately and fast with less user interaction, a large number of video segmentation methods based on graph cuts have been proposed. Onecut [16], as a kind of improved method of graph cuts, improves the description of region distinction by using the distance L_1 of color histogram in the image, in order to compute the distance between the foreground and the background model, and the global optimum can be obtained by means of once graph cut. As a result, the precision and accuracy of the segmentation results can be improved.

The main idea of Onecut is to add some intermediate auxiliary nodes to the traditional graph structure [28]. These auxiliary nodes represent the statistics of each pixel's color information (color histogram) in the current frame instead of GMM in the traditional Grabcut. The main advantage of Onecut is that it avoids multiple iterations of Grabcut, so it not only reduces the time cost, but also enhances the segmentation effect. The original Onecut segmentation model is enhanced in this paper by both using the global probability model and the localized classifier model to achieve better segmentation effect, and it is successfully applied to video segmentation.

4.1 Energy function

The energy function model of enhanced Onecut in our method is described as follow:

$$E(S) = \sum_{x \in \Omega} D(x) - \tau \|\theta^S - \theta^{\bar{S}}\|_{L_1} + \eta |\partial S| \quad (7)$$

The optimization result of energy only depends on the segmentation result S . In formula (7), S represents the set of foreground pixels, and $\bar{S}$ represents the set of background

pixels, and $\Omega = S \cup \bar{S}$. θ^s and $\theta^{\bar{s}}$ represent the statistical histograms of foreground and background respectively, and τ is the weight of the distinction between foreground and background. η is the weight of the smoothness term. $|\partial S|$ is the similarity among neighboring pixels, which can be calculated by formula (8).

$$|\partial S| = \sum_{\{x,y\} \in N} \omega_{xy} |s_x - s_y| \quad (8)$$

In formula (8), s_x and s_y represent the labels of pixel x and y , and N is a set consisted of adjacent pixel pairs in the same frame. ω_{xy} is the similarity between x and y , and it can be calculated by formula (9).

$$\omega_{xy} = \frac{1}{\|x - y\|} \cdot \exp\left(\frac{-\Delta I^2}{2\sigma^2}\right) \quad (9)$$

In formula (9), $\|x - y\|$ is the Euclidean distance between pixel x and y , ΔI is the color distance between pixel x and y , and σ is set as the gradient mean of the whole image.

In addition, the $D(x)$ in formula (7) represents the enhanced Onecut data item that is added through integrating the global probability model and the localized classifier as formula (10):

$$D(x) = \begin{cases} -\log\left(\alpha \cdot p_{global}^1(x) + (1 - \alpha) \cdot p_{local}^1(x)\right) & x = 1 \\ -\log\left(\beta \cdot p_{global}^0(x) + (1 - \beta) \cdot p_{local}^0(x)\right) & x = 0 \end{cases} \quad (10)$$

In details, the global foreground probability $p_{global}^1(x)$ and the global background probability $p_{global}^0(x)$ can be calculated by the global classifier model in Sect. 2.1, and the local foreground probability $p_{local}^1(x)$ and the local background probability $p_{local}^0(x)$ can be calculated by the localized classifier model in Sect. 2.2. Since the global classifier model is of greater influence to foreground and the localized classifier model is of greater influence to background, the global GMM occupies a larger proportion when the foreground probability ($x = 1$) is calculated, while the localized GMM occupies a larger proportion when the background probability ($x = 0$) is calculated. Here, the proportion of the global classifier model is supposed to be α when the foreground probability is calculated, then the corresponding ration of localized classifier model is $1 - \alpha$, while the proportion of the global classifier model is supposed to be β when calculating the background probability, and the corresponding ration of localized classifier model is $1 - \beta$. In this paper, we set α and β as 0.6 and 0.4 respectively.

4.2 Graph Structure

In Fig. 2, it shows the connection graph of all nodes in the k -th bin [16] of the gradient histogram. S and T represent the source point and the sink point of the graph structure respectively. A_k^1 represents a virtual node that is classified into the foreground in the k -th bin, and A_k^0 represents a virtual node that is classified into the background in the k -th bin. The source point and the sink point of the original Onecut are connected only to the virtual nodes, while in our proposed enhanced Onecut, they are connected to each pixel besides the virtual nodes. V_1, V_2, V_{n_k} represent real pixel points, and the solid lines represent that the graph structure must contain the weighted sides. The dotted lines among V_1, V_2, V_{n_k} represent that if real pixels are adjacent, they have weighted sides, and if they are not adjacent, there is no weighted sides.

From Fig. 2, if the connection of each pixel to the source point S and the sink point T is removed, and the rest is the standard Onecut segmentation model. If the virtual node and the sides that are connected to it are removed, the Grabcut segmentation model is left. In conclusion, the enhanced Onecut model includes both the original Onecut model and the Grabcut model, in other word, the Onecut model and the Grabcut model are only a special case of the enhanced Onecut segmentation model in this paper.

To further simplify the scale of graph structure of the enhanced Onecut model corresponding to formula (7), firstly, all the pixels in Ω are divided into two parts: the seed points and the unknown region. Seed points include foreground seed points and background seed points, which are pixel points with obvious foreground/background labels after interaction or pre-processing. Except the seed points, the remaining pixels are set to the unknown region. In order to reduce the computational cost, the enhanced Onecut model is only applied to the pixels in the unknown region.

Fig. 2 The graph structure of enhanced Onecut segmentation model

Assuming that the segmentation result of the previous frame is accurate enough, the initial contours of the target in the current frame can be obtained by patch matching. And if the target do not move fast, the estimated initial contour is close to the true contour. The localized classifier obtained after patch matching can contain this estimated initial contour. In addition, the initial contour is adaptively extended to form a narrow band with the width calculated by formula (1) so that the true contours of the target can be covered as much as possible in the current frame. Therefore, in this paper, the pixels which are not covered by the localized classifier of current frame are set as the default foreground or background seed points according to their relationship with the contour.

In addition, for all pixel points within the coverage area of the localized classifier in the current frame, they can be selected as foreground seed points or background seed points based on the integrated foreground/background probabilities in formula (10) and the distance of the pixel to the initial predicted contour (the greater the distance is, the greater the probability of becoming a seed point is). Specifically, for a certain point in the coverage area of the localized classifier, if the integrated foreground probability of the pixel is greater than 0.9, and the pixel is within the estimated initial contour, whose distance is greater than $1/3$ of the width of the localized classifier (see formula 1), the pixel can be selected as a foreground seed point. Otherwise, if the integrated background probability of the pixel is greater than 0.9, and the pixel is outside the estimated initial contour, whose distance is greater than $1/3$ of the width of the localized classifier, the pixel can be selected as a background seed point.

5 Experiment and result analysis

In the experiment, the Visual Studio 2013 and the OpenCV image library are selected as the development tools and the experiment program is written in C++. The testing dataset used in our experiments is DAVIS (Densely Annotated Video Segmentation) [29]. It contains 50 high-quality and full HD video sequences, and are specifically divided into 30 training sequences and 20 test sequences. For training data, the total number of frames is 2079, and the average number of frames is 69.3. For the test data, the total frame number is 1376, and the average frame number is 68.8. The dataset spans multiple occurrences of common video segmentation challenges such as occlusions, inconsistent movement and appearance changes.

5.1 Comparison between global and localized classifier modeling

In this paper, global classifiers and localized partial classifiers are constructed simultaneously. In Fig. 3, it shows the probability maps of global classifier modeling and the localized classifier modeling, and the probability maps that combines both of them. From Fig. 3a, it is known that global modeling can get rich global information, but too much information in the background makes it impossible to highlight foreground information. Fig 3b shows that the Localized classifier can make the foreground information salient, but it loses many details. The results of the combination classifier are shown in Fig. 3c, it not only highlights the integrity of the target, but also considers the local features of the target.

5.2 Comparison between patch matching and other matching methods

In this paper, patch matching is used to transfer the target contour of the previous frame to the current frame. For the following two cases, patch matching is compared with both the sparse point matching strategy and the contour affine transformation of Snapcut. One is the situation in which the global motion is consistent with the local motion, and the other is that the global movement is inconsistent with the local movement of target. The comparison results are shown in Figs. 4 and 5 respectively.

From the experiment results in Figs. 4 and 5, we can see that when the global movement is highly consistent with the local movement of target, the results of target contours nearly have no obvious difference between patch matching and other matching methods. However, when the entire shifting is tiny meanwhile the local transformation is huge, the result of the affine transformation tracking is worst. Additionally, since there must be some noise interference, the result of single point sparse matching is also not good enough. Generally, the single point sparse matching usually outperforms affine transformation, while our proposed patch matching is the most accurate one when tracking contour. In details, we can find out that region-based patch matching performs better than the single sparse point matching in tracking border, and outperforms the affine transformation of Snapcut when the local transformation is huge. This will make our proposed patch matching more robust when self-propagating the contours in video.

Fig. 3 Probability maps of global and localized classifier models

Fig. 4 Testing case where global movement is consistent with the local movement of target

Fig. 5 Testing case where global movement is inconsistent with the local movement of target

5.3 Comparison between our segmentation method and other methods

Since the method we adopt belongs to interactive video segmentation method, in contrast, we select several semi-supervised segmentation methods in the DAVIS dataset for testing, such as SEA (Seamseg) [30], NLC (Non-Local Consensus voting) [31], JMP (Jumpcut) [32], HVS (Hierarchical graph-base Video Segmentation) [33], FCP (Fully Connected object Proposals) [34], BVS (Bilateral space Video Segmentation) [35]. The IoU value of every image is calculated by (11), and the average IoU value AVG is calculated by (12).

$$\overline{IoU} = \frac{1}{T} \sum_{t=1}^T IoU(t) \quad (11)$$

In (11), T is the total number of frames in the whole video, and $IoU(t)$ is the IoU value of t -th frame, $t \in [1, T]$.

$$\overline{IoU}_{all} = \frac{1}{K} \sum_{k=1}^K \overline{IoU}(k) \quad (12)$$

In (12), it is supposed that a group of videos contains K videos, and $\overline{IoU}_{all}$ represents the average value of all video IoU values in the group videos. $\overline{IoU}(k)$ is the average IoU value of k -th frame, $k \in [1, K]$.

The test videos in DAVIS dataset are divided into four categories: videos with high contrast color distributions, videos with obscured object, videos with inconsistent global and local movement, and videos with large scale change. Quantitative comparison in DAVIS dataset is made

among Our Method, BVS, FCP, HVS, JMP, NLC, and SEA as Tables 1, 2, 3 and 4. The significance of bold means the best performance.

As shown in Table 1, for videos with high contrast color distributions in foreground and background, our method has a good performance, and the average IoU value is obviously higher than other methods. Because the classifier proposed in this paper contains Gaussian Mixture Model (GMM) of the color features and the shape features with the position information of obtained target contour.

As shown in Table 2, for videos with multiple difficulty attributes such as the moving target may be temporarily and partially occluded, the IoU values of segmentation results of our method will slightly decrease, but it is still better than other methods since the global classifier is constructed in our method, so that the global feature is integrated into the enhanced Onecut segmentation model, and therefore better segmentation results can be obtained.

As shown in Table 3, the inconsistency between the whole and local motion usually occurs when the target is non-rigid, and our method can still obtain a very good performance stably. The reason is that, with the design of patch matching, the target contour tracking of our method can be achieved by using the small range region-based matching of each point on the target contour, and thus better tracking results can be obtained to guide the segmentation.

As shown in Table 4, when the scale of foreground target changes dramatically and frequently, the average IoU value of our method is also obviously higher than other compared methods, since the target contour will

Table 1 Quantitative comparison for videos with high contrast color distributions

IoU	Ours	BVS	FCP	HVS	JMP	NLC	SEA
Bear	0.943	0.954	0.906	0.938	0.929	0.806	0.912
Blackswan	0.938	0.943	0.908	0.917	0.930	0.874	0.933
Boat	0.783	0.642	0.613	0.782	0.705	0.007	0.793
Cows	0.828	0.894	0.812	0.779	0.756	0.883	0.707
Dog	0.724	0.720	0.774	0.722	0.673	0.809	0.581
Elephant	0.851	0.848	0.655	0.742	0.750	0.518	0.553
Goat	0.678	0.661	0.677	0.580	0.731	0.010	0.535
Hike	0.925	0.755	0.874	0.877	0.664	0.918	0.776
Kite-surf	0.691	0.423	0.577	0.405	0.500	0.453	0.487
Mallard-water	0.904	0.907	0.687	0.704	0.751	0.761	0.865
Paragliding	0.953	0.874	0.866	0.907	0.951	0.880	0.863
Paragliding-launch	0.641	0.637	0.571	0.537	0.589	0.628	0.577
Scooter-gray	0.226	0.503	0.483	0.433	0.123	0.586	0.241
Stroller	0.759	0.765	0.597	0.662	0.656	0.850	0.464
Train	0.874	0.871	0.841	0.846	0.873	0.729	0.854
AVG	0.781	0.760	0.723	0.722	0.705	0.647	0.676
STDEV	0.186	0.161	0.140	0.171	0.206	0.295	0.202

Table 2 Quantitative comparison for videos with obscured object

IoU	Ours	BVS	FCP	HVS	JMP	NLC	SEA
bmx-bumps	0.432	0.433	0.402	0.429	0.340	0.631	0.203
bmx-trees	0.282	0.382	0.248	0.178	0.229	0.212	0.113
Bus	0.834	0.863	0.832	0.809	0.668	0.629	0.752
Dog-agility	0.655	0.331	0.453	0.457	0.699	0.652	0.354
Kite-walk	0.840	0.870	0.682	0.765	0.509	0.813	0.498
Libby	0.509	0.768	0.316	0.553	0.295	0.635	0.226
Lucia	0.894	0.900	0.801	0.776	0.836	0.876	0.626
Motorbike	0.638	0.552	0.713	0.687	0.506	0.714	0.451
Rhino	0.844	0.779	0.794	0.812	0.716	0.682	0.736
Soccerball	0.723	0.838	0.567	0.075	0.103	0.832	0.654
Surf	0.960	0.483	0.843	0.759	0.941	0.775	0.821
Swing	0.721	0.784	0.648	0.104	0.115	0.851	0.511
AVG	0.694	0.665	0.608	0.534	0.496	0.692	0.495
STDEV	0.203	0.212	0.209	0.283	0.281	0.177	0.233

Table 3 Quantitative comparison for videos with inconsistent global and local movement

IoU	Ours	BVS	FCP	HVS	JMP	NLC	SEA
Breakdance	0.580	0.497	0.567	0.550	0.478	0.673	0.329
Breakdance-flare	0.718	0.727	0.723	0.499	0.430	0.804	0.131
Dance-jump	0.767	0.744	0.522	0.680	0.490	0.718	0.662
Dance-twirl	0.353	0.489	0.471	0.318	0.444	0.347	0.117
Flamingo	0.751	0.881	0.717	0.811	0.530	0.539	0.583
Hockey	0.853	0.828	0.647	0.698	0.677	0.810	0.714
Mallard-fly	0.712	0.597	0.541	0.436	0.536	0.617	0.557
Parkour	0.820	0.755	0.322	0.240	0.342	0.901	0.121
Rollerblade	0.680	0.584	0.449	0.461	0.726	0.814	0.138
Tennis	0.733	0.736	0.623	0.576	0.765	0.871	0.482
AVG	0.697	0.684	0.558	0.527	0.542	0.709	0.383
STDEV	0.142	0.135	0.125	0.175	0.138	0.171	0.243

Table 4 Quantitative comparison for videos with large scale change

IoU	Ours	BVS	FCP	HVS	JMP	NLC	SEA
Camel	0.853	0.666	0.734	0.876	0.640	0.768	0.649
Car-roundabout	0.825	0.848	0.717	0.777	0.726	0.509	0.708
Car-shadow	0.907	0.569	0.723	0.699	0.645	0.645	0.775
Car-turn	0.824	0.843	0.724	0.810	0.834	0.833	0.909
Drift-chicane	0.484	0.032	0.457	0.331	0.243	0.324	0.119
Drift-straight	0.494	0.394	0.668	0.295	0.618	0.473	0.513
Drift-turn	0.620	0.296	0.606	0.276	0.717	0.154	0.667
Horsejump-high	0.837	0.799	0.676	0.765	0.586	0.834	0.637
Horsejump-low	0.631	0.592	0.607	0.551	0.663	0.651	0.498
Motocross-bumps	0.559	0.396	0.306	0.534	0.761	0.614	0.470
Motocross-jump	0.388	0.335	0.511	0.099	0.583	0.251	0.386
Scooter-black	0.686	0.329	0.504	0.624	0.626	0.162	0.793
Soapbox	0.764	0.788	0.449	0.684	0.759	0.634	0.783
AVG	0.682	0.530	0.591	0.563	0.646	0.527	0.608
STDEV	0.167	0.255	0.135	0.243	0.143	0.239	0.210

Table 5 Comparison of Average IoU in DAVIS Dataset

IoU	Ours	BVS	FCP	HVS	JMP	NLC	SEA
AVG of all videos	0.718	0.662	0.628	0.597	0.607	0.639	0.557
STDEV of all videos	0.177	0.211	0.164	0.231	0.214	0.237	0.240

be adjusted adaptively as the scale of the target changes through our proposed contour self-propagating, and then our method can obtain the better segmentation results without doubt.

The comprehensive results of all videos in DAVIS dataset is shown in Table 5, specifically including the average IoU and the standard deviation of IoU. It is clearly that the average IoU value of our method is 0.718, which is

Fig. 6 The final segmentation results of several typical videos in DAVIS dataset

higher than other compared state-of-the-art methods. And it can be seen that for different categories, our method performs well in segmentation and has small standard deviation, which shows good robustness. From the experiments, we find that our proposed segmentation method is effective for most of the videos in the dataset, especially in the case of some object with strong integrity. Additionally, even in some complex scenes, our method is also robust enough. The final segmentation results of several typical videos in DAVIS dataset are shown in Fig. 6. In conclusion, testing on the DAVIS dataset, our proposed method performs more outstandingly on the whole in terms of accuracy and robustness when compared with state-of-the-art video segmentation methods.

6 Conclusion

In this paper, a self-propagating video segmentation method is proposed, which combines color features and shape features as well as moving information of target. Compared with state-of-the-art segmentation methods, the advantages of our proposed method are as follows: First, the global probability model and the localized classifier model are both adopted, and at the same time the color features, shape features and moving information are integrated. It not only guarantees the integrity of the target, but also takes into account the local contour of the target, so it has higher fault tolerance. Second, patch matching method is come up with to improve the anti-noise ability of single sparse point matching, and the motion information is made full use to increase the robustness of the contour self-propagating. Third, the traditional Onecut segmentation model is adequately enhanced in this paper by using global probability model, localized classifier model and the interframe smoothing term N -links. All these reasons make a direct contribution to the outstanding performance of our proposed self-propagating video segmentation method on the DAVIS dataset.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant No. 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.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Latif A, Rasheed A, Sajid U et al (2019) Content-based image retrieval and feature extraction: a comprehensive review. *Math Probl Eng* 21:9658350
2. 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* 2019:8041413
3. Sajid M, Ali N, Dar SH et al (2018) Data augmentation-assisted makeup-invariant face recognition. *Math Probl Eng* 2018:2850632
4. 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):e0219833
5. Ali N, Zafar B, Riaz F et al (2018) A hybrid geometric spatial image representation for scene classification. *PLoS One* 13(9):e0203339
6. Unlu E, Zenou E, Riviere N (2017) Ordered minimum distance bag-of-words approach for aerial object identification. In: *IEEE international conference on advanced video and signal based surveillance (AVSS)*, pp 1–6
7. Freitas GMD, Tozzi CL (2017) Object tracking by multiple state management and eigenbackground segmentation. *Int J Nat Comput Res* 1(4):29–36
8. Nadeem S, Rahman A, Butt AA (2017) A framework to combine multi-object video segmentation and tracking. In: *International conference on digital image computing: techniques and applications*, pp 1–7
9. Lin C, Pun CM, Huang G (2017) Highly non-rigid video object tracking using segment-based object candidates. *Multimed Tools Appl* 76(7):1–22
10. Portenier T et al (2018) Fine-grained retrieval with autoencoders. In: *International conference on computer vision theory and applications*, pp 85–95
11. Fu C, Chen D, Delp E J et al (2018) Texture segmentation based video compression using convolutional neural networks. *arXiv preprint arXiv:1802.02992*
12. Wang W et al (2018) Saliency-aware video object segmentation. *IEEE Trans Pattern Anal Mach Intell* 40(1):20–33
13. Papazoglou A, Ferrari V (2013) Fast object segmentation in unconstrained video. In: *International conference on computer vision*, pp 1777–1784
14. Li F et al. (2013) Video segmentation by tracking many figure-ground segments. In: *International conference on computer vision*, pp 2192–2199
15. Bai X et al (2009) Video SnapCut: robust video object cutout using localized classifiers. *ACM Trans Graph* 28(3):1–11
16. Tang M et al (2013) GrabCut in one cut. In: *International conference on computer vision*, pp 1769–1776
17. Rother C, Kolmogorov V, Blake A (2004) “GrabCut”: interactive foreground extraction using iterated graph cuts. *Trans Graph* 23(3):309–314
18. Sonavane R, Sonar P (2017) Classification and segmentation of brain tumor using Adaboost classifier. In: *International*

- conference on global trends in signal processing, information computing and communication, pp 396–403
19. Luo B et al (2017) Video object segmentation via global consistency aware query strategy. *IEEE Trans Multimed* PP(99):1–1
 20. Hu H et al. (2017) LabelBank: revisiting global perspectives for semantic segmentation. *arXiv preprint [arXiv:1703.09891](https://arxiv.org/abs/1703.09891)*
 21. Marcu AE, Leordeanu M (2017) Object contra context: dual local-global semantic segmentation in aerial images. *AAAI workshops. AAAI Workshops 2017*: San Francisco, CA, USA, pp 146–152. <http://www.aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15177>
 22. Arthur D, Vassilvitskii S (2015) k-means++: the advantages of careful seeding. In: *Eighteenth ACM-SIAM symposium on discrete algorithms*, pp 1027–1035
 23. Khedher MI, Yacoubi MAE, Dorizzi B (2013) Multi-shot SURF-based person re-identification via sparse representation. In: *IEEE international conference on advanced video and signal based surveillance*, pp 159–164
 24. Ponttuset J, Marques F (2016) Supervised evaluation of image segmentation and object proposal techniques. *IEEE Trans Pattern Anal Mach Intell* 38(7):1465–1478
 25. Jain SD, Grauman K (2014) Supervoxel-consistent foreground propagation in video. In: *European conference on computer vision*. Springer, pp. 656–671
 26. Badrinarayanan V, Budvytis I, Cipolla R (2014) Mixture of trees probabilistic graphical model for video segmentation. *Int J Comput Vis* 110(1):14–29
 27. Wang T, Han B, Collomosse J (2014) TouchCut: fast image and video segmentation using single-touch interaction. *Comput Vis Image Underst* 120(2):14–30
 28. Han S et al (2010) Fast image segmentation based on multilevel banded closed-form method. *Pattern Recogn Lett* 31(3):216–225
 29. Perazzi F et al. (2016) A benchmark dataset and evaluation methodology for video object segmentation. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 724–732
 30. Ramakanth SA, Babu RV (2014) SeamSeg: video object segmentation using patch seams. In: *Computer vision and pattern recognition*, pp 376–383
 31. Faktor A, Irani M (2014) Video segmentation by non-local consensus voting. In: *British machine vision conference*, p 8
 32. Fan Q et al (2015) JumpCut: non-successive mask transfer and interpolation for video cutout. *ACM Trans Graph* 34(6):195
 33. Grundmann M et al. (2010) Efficient hierarchical graph-based video segmentation. In: *Computer vision and pattern recognition*, pp 2141–2148
 34. Perazzi F et al. (2015) Fully connected object proposals for video segmentation. In: *International conference on computer vision*, pp 3227–3234
 35. Marki N et al. (2016) Bilateral space video segmentation. In: *Computer vision and pattern recognition*, pp 743–751

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