IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2020, accepted March 5, 2020, date of publication March 16, 2020, date of current version March 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981140

Robust Video Object Segmentation via Propagating Seams and Matching Superpixels

YUN LIANG, YUQING ZHANG, YIHAN WU, SHUQIN TU, AND CAIXING LIU
College of Mathematic and Informatics, South China Agricultural University, Guangzhou 510642, China

Corresponding author: Caixing Liu (liu@scau.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61772209, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2019A050510034 and Grant 2019B020219001, in part by the Production Project of Ministry Education China under Grant 201901240030, and in part by the College Students Innovations Special Project of China under Grant 201910564037.

ABSTRACT Video object segmentation aims at separating foreground object from background, and it is far from well solved for different challenges such as deformation, occlusion and motion blurs. This paper proposes a robust video object segmentation method by propagating patch seams and matching superpixels. First, we predict the initial object contour based on pixel-level target labels calculated by patch seam propagation and rough sets. By a patch seam, we map a current patch to its most similar patch from last frame and obtain its labels based on the labels of mapped patch. Second, we utilize superpixels as middle level cues to optimize predicted object contour. The bidirectional distance based on three brightness channels is provided to match superpixels between adjacent frames. Using the boundaries of matched results and initialized object contour, many candidates of object contours are constructed. Third, we define an energy function based on multi-features to measure contour candidates, and the contour with minimum energy is the final segmented result of current frame. Finally, by propagating patch seams and matching superpixels, we compute video object segmentation results frame by frame. Fourteen videos of SegTrack-v2 data are used to evaluate our method. The quantitative and qualitative evaluations show that our method performs better than most present methods especially in dealing with occlusion, deformation and motion blurs.

INDEX TERMS Video object segmentation, seam propagation, superpixel match, energy minimization.

I. INTRODUCTION

Video object segmentation is a hot topic in computer vision and machine learning, which provides the basic semantic unit for high level video processes such as target detection, action recognition, video retrieval, video editing and so on. It focuses on how to accurately segment the moving object from a video frame by frame based on certain criteria. Due to the complex changes of object and its surrounding backgrounds, there are many challenges such as occlusion, deformation, illumination changes, motion blur, fast motion and so on. Recently, a robust video object segmentation is still far from well solved.

Video object segmentation is achieved by predicting object contours frame by frame in recent years. The key point is how to accurately separate the moving object and its surrounding background according to the similarity or inheritance of object contour. Unfortunately, most present methods often produce results with errors or drifts in dealing with the various kinds of challenges. At the same time, the errors or drifts about object contours are gradually accumulated and propagated frame by frame, and often bring final video object segmentation failure.

According to whether using learning scheme, video object segmentation methods are classified into two kinds: learning based method and no-learning based method. The learning based method [1]–[5] often constructs a neural network and loss energy function, and trains the network to optimally classify video frames into foreground object and background. This kind of method produces accurate results sometimes, however, it greatly depends on the training data [7], [8]. If the training data is not robust or not big enough, the segmenting errors emerge. Furthermore, it requires special hardware devices (GPU) and high implementing time for training

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

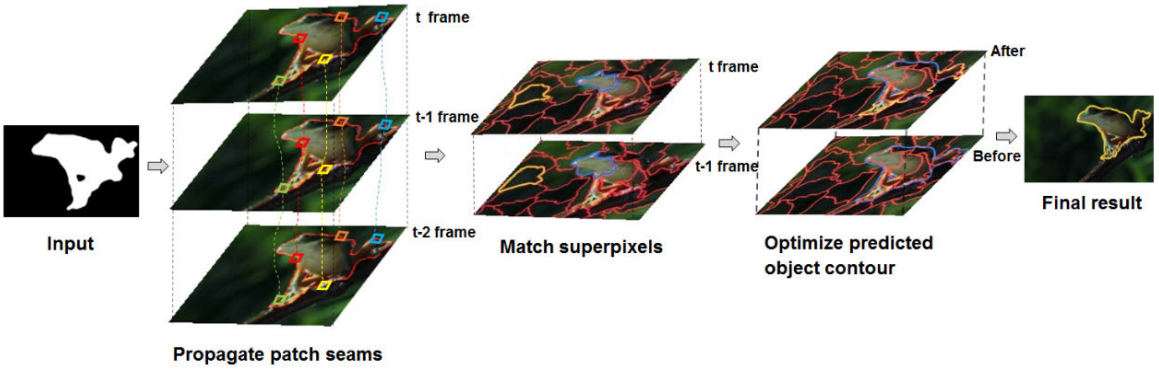


FIGURE 1. The flowchart of our method. After inputting the ground truth on the first frame, we first initialize object contour by propagating patch seams (column 2), and second match superpixels (column 3), then optimize the predicted result based on matched result (column 4), and finally get the segment result (column 5).

and learning. The no-learning based method [11]–[15] often uses the inherited cues from appearance and motion of foreground object to detect target and further segment it. This kind of method performs well in separating object from background by exploiting their various kinds of features such as texture, brightness, structure, motion and so on. Therefore, this paper focuses on how to design robust algorithm to efficient use the cues of object, especially the cues from different levels.

The SeamSeg method [19] firstly proposes a video object segmentation by introducing seams. It propagates the object contour based on seam cues between adjacent frames. This method succeeds in dealing with slowly changes of target, but without out considering high level cues it fails on great occlusion and deformation. Therefore, the seamSeg method provides a good way to predict general object contour.

This paper defines a new robust video object segmentation method by combining the pixel-level cues from object contour and the superpixel-level cues from object local regions. Figure 1 shows the flowchart of our method. First, we define a patch seam propagation method to initialize the object contour as the second column in Figure 1. It predicts the object contour on pixel level by minimizing the energy describing patch seams. Then, we use superpixels to match local regions of adjacent frames. The superpixels with blue (yellow) boundaries in the third column of Figure 1 are the matched ones about object (background). This match uses the boundaries of superpixels to produce many contour candidates. Finally, we optimize the predicted object contour by defining an energy function based on multi-features evaluation to measure contour candidates. The candidate with minimum energy decides the segmented result. The fourth column in Figure 1 shows the result before and after optimization process. In a word, our method employs seam propagation to use the low-level cues to predict object contour, and utilizes superpixel match to use the middle level cues to optimize predicted object contour. The combination of two-level cues successfully employs both temporal and spatial information to deal with the contour drifts or errors and finally improve the segmenting accuracy.

We organize this paper as follow. Section II describes the related word. Section III describes the proposed method in details. In Section IV, many experiments are done to verify the proposed method. In Section V, we give a conclusion.

II. RELATED WORK

Video object segmentation aims at segmenting the moving target object from a video sequence. As it can produce target for people to identify and analyze video content, video object segmentation is widely used in artificial intelligence (AI), visual perception, and plays an important role in intelligent monitoring, semantic analysis, visual navigation and so on. If the target is automatically detected by method, this kind of segmentation belongs to the unsupervised method. Otherwise, it belongs to the supervised method. In this paper, we focus on the supervised method, in which the target object is specified by user on the first frame.

Due to the constant changes of object and its surrounding background, there are various kinds of challenges for video object segmentation such as illumination change, motion blur, occlusion, deformation, fast motion, cluttered background, and so on. These challenges often coexist at same time and result in contour drift which finally leads to segmenting errors. Therefore, many people focus on providing an accurate and robust video segmentation method recently.

Recently, various methods for supervised video object segmentation have been proposed. They are roughly divided into two categories: one is based on deep learning, and the other is based on non-deep learning. The methods based on deep learning methods often firstly construct a neural network architecture to represent object and background such as the Deep Neural Network (DNN) [1] or Convolutional Neural Networks (CNN) [2]. Then, they use the network to train a large number of foreground and background dataset to obtain the decision function, and utilize it to classify pixels into foreground and background to achieve object segmentation. This kind of method focuses on learning the representation of appearance [3] and motion cues [4], [5]. For example, using the co-attention of siamese networks in

propagating mask and detecting object, Oh *et al.* [1] define a fast and reference guided video segmentation method. At the same time, Oh *et al.* [2] combine two learning models to improve segmented result by many guiding scribbles. Caelles *et al.* [3] define a new one-shot video object segmentation method based on a fully-convolutional neural network to learn the appearance of object by transferring the generic semantic information learned from the public ImageNet [6]. Tokmakov *et al.* [4] define a recurrent neural network to segment target by learning its motion cue. Xu *et al.* [5] define a spatiotemporal CNN model to pretrain and learn the dynamic appearance and motion cues of video sequence to guide object segmentation. Voigtlaender *et al.* [7] propose a fast end-to-end embedding learning method to use a semantic pixel-wise embedding with a global and local matching mechanism to transfer target information from the first and previous frame to the current frame. Zhang *et al.* [8] propose a segmentation framework based on dual-stream DNN, which uses the robust pixel-level features of all video frames to generate foreground images. Zheng *et al.* [9] propose a new method by using a ResNet encoder to model spatial information, and a Conv LSTMs-based decoder to model temporal information. Li *et al.* [10] propose an attention-guided network to adaptively strengthen inter-frame and intra-frame features to improve the accuracy of video object segmentation. By training and learning, the video object segmentation methods based on deep learning sometimes produce accurate results [28]–[30]. However, they often need special hardware which limit their public usage and need high implement time. At the same time, they are not very robust and sometimes very sensitive to video contents as their performances are greatly depending on the training data.

The video object segmentation methods based on non-deep learning use the appearance features of object and background, and the motion cues from video to separate foreground object from background. Generally, these methods are classified into three kinds. The first kind is defined based on optical flow. For example, Nagaraja *et al.* [11] use time series information of optical flow to enhance the consistency of color distribution in successive frames. Tsai *et al.* [12] define a video object segmentation method by the optical stream and spatiotemporal cues, which constructs the segmentation model based on a priori tag data, and corrects results by re-estimating the optical flow at segmented edge. Sokeh *et al.* [13] propose to segment video by the optical flow to obtain the intensity of change between successive frames. The second kind of non-deep learning method is proposed via graph cut. For example, Huarong *et al.* [14] segment video object by graph cut and the support vector machine (SVM) trained by previous segmented results. Zhang *et al.* [15] segment video object by layering a directed and acyclic graph, which predicts foreground by evaluating the motion, appearance and shape of object represented by the graph. The third kind of non-deep learning method is achieved by transferring object contours between adjacent frames. For example, Lu *et al.* [16] propose a parameter model of object contour based on Bezier curve, which optimize object contour propagation by the spatiotemporal confidence and a small amount of manual target labels. Wang *et al.* [17] first compare the local and global saliency of object based on the boundary and motion cues between adjacent frames, and then define a joint spatiotemporal energy to propagate object contour to segment foreground. Similarly, Wang *et al.* [33] combine the close spatiotemporal relationships with the consistent motion patterns and similar appearances to define a super-trajectory to segmented target. Compared with learning based method, the non-deep learning method of video object segmentation do not need high time cost and special hardware. However, as shown in [31], [32], they struggle on how to successfully using multi-level object features to deal with segmenting challenges.

This paper proposes a robust video object segmentation method by using low-level features and middle-level features of object and its surrounding background. With the low-level features, we define an object contour propagating algorithm based on patch seams to compute pixel-level object contour. With the middle-level features, we optimize the propagated results by matching superpixels between adjacent frames to use local structure information to reduce errors or drifts from pixel-level contour propagating. Combining the above two-level features, we utilize the inherited cues about brightness and gradient of contour pixels from last frames to separate an object with its surroundings, and effectively use structure cues to deal with challenges from occlusion and deformation. Without training/learning, just by propagating and optimizing, the proposed video object segmentation method produces more favorable results on the videos of SegTrack-v2 data from than most of the present methods.

III. THE PROPOSED VIDEO OBJECT SEGMENTATION METHOD

A new robust video object segmentation is defined by propagating patch seams and matching superpixels. It first predicts the initial object contour by propagating patch seams. Second, it divides the interest region including predicted target object into many superpixels. Then, using the locality advantage of superpixel in presenting object structure, edges and semantic information, we optimize the predicted contour by an energy function. This energy function is defined to evaluate the contour candidates formed by fusing the boundaries of matched superpixels and the predicted object contour. The final segmented result is the contour candidate of object with minimum energy. The details of our proposed method are as follows.

A. INITIALIZE OBJECT CONTOUR VIA PROPAGATING PATCH SEAMS

Our method first constructs patch seams to propagate video object contours between adjacent frames. As demonstrated in [19], [20], a seam is a connected path with pixel-level width, which is formed by minimizing an energy function. However, the traditional seam is defined by the differences

of brightness and color between adjacent two pixels, which is very sensitive to object deformation and similar background in separating object. Therefore, this paper employs patch seams as in [19] which define a seam based on the total differences between a group of pixels from different patches rather than only on the differences between two pixels. Following this, our method performs more robust and favorably than most present video object segmentation methods especially in dealing with the challenges of occlusion and deformation.

Our method initializes the object contour by four steps. First, we define an interest region including object and its adjacent surroundings on the last frame. Each pixel in this region is marked by object label and background label according to the object contour of last frame. Second, we construct patch seams by minimizing an energy function to form seam relationship between the patch from current frame with the patch from last frame. Third, we propagate the labels from the patch of last frame to the current patch to get patch-wise label based on patch seams. Finally, using the rough set scheme, we get pixel-wise label according to patch-wise labels and use it to segment out the foreground object on current frame.

1) PROPAGATE PATCH-WISE LABLES BASED ON PATCH SEAMS

According to the consistency of motion and appearance, the foreground object of current frame is very similar to the one of last frame. With this prior, we use patch seams to propagate patch-wise labels from last frame to current frame to predict object. As shown in [19], we propose an energy function based on measuring the brightness and position differences between patches from last frame and current frame, and measuring the consistence of adjacent patches from current frame. Using this energy, patches connected by a seam own similar appearance and near position between adjacent frames. For the differences of color and center between two patches describe the changes from motion and appearance, we define the energy function for patch seam by Equation 1. Assigned a patch with $p * p$ pixels, the energy between patch $I_{t-1}(x, y)_p$ centered at (x, y) on $(t-1)$ frame and patch $I_t(i, j)_p$ centered at (i, j) on t frame is defined by:

$$\begin{aligned} E_{i,j,t}(x, y) = {} & \sigma_1 * \left\| I_t(i, j)_p - I_{t-1}(x, y)_p \right\|_2 \\ & + \sigma_2 * \|(i, j) - (x, y)\|_2 \\ & + \sigma_3 * \sum\nolimits_{\delta,\varepsilon} \left\| I_t(i, j)_p - I_t(i+\delta, j+\varepsilon)_p \right\|_2 \end{aligned} \quad (1)$$

where, $||*||_2$ is the Euclidean distance of two vectors. The first item measures the total differences of two patches from three-channel colors in RGB and the gradient values of x, y directions. For this item, smaller value means that the patches are more similar. The second item computes the position distance between two patches, which is introduced to penalize the connected patch on current frame far from that of last frame. It ensures that the patches connected by seam are adjacent and cannot drift too far. The third item calculates the differences of patch $I_t(i, j)_p$ and patch $I_t(i+\delta, j+\varepsilon)_p$. The $(i+\delta, j+\varepsilon)$ is the center near to (i, j) which describes the appearance and motion consistence of the adjacent patches on current frame. This item is proposed to reduce the incoherence of approximate neighborhood and capture object motion more accurately. The σ_1, σ_2,σ_3 represent the weight coefficients for the three items, and are all set to 1 in our experiments.

By minimizing the above energy function, we connect the current patch with the last patch from the previous frame by their patch seam. This means that the last patch moves to the current patch. Then, we propagate patch-wise label by this patch seam. In details, we assign the label for each pixel of the current patch to be the value with its related pixel of the last patch. However, each pixel is included into p ∗ p patches, and it is assigned for $p * p$ times. Therefore, we employ the rough set scheme [22] to compute the final label of each pixel to obtain the predicted object contour.

2) OBTAIN THE PREDICTED OBJECT CONTOUR VIA ROUGH SET SCHEME

After propagating patch-wise labels, we obtain $p*p$ labels for each pixel. Then, we use the rough set scheme to compute the unique final label for each pixel. According to the final label, we employ the morphology algorithm [35] to produce the predicted object contour.

The labels from the last frame have two kinds, object label and the background label. All the p ∗ p labels of each pixel either belong to object or background. With these labels, the rough set is used to compute the final label of each pixel. If $L_{i,j}$ is the rough set about the labels of pixel (i, j), we define:

$$I_t(i, j) \in R(X) \leftrightarrow \left| L_{i,j} \right|_X \geq \alpha * p^2 \quad (2)$$

$$I_t(i, j) \in U - R(X) \leftrightarrow \left| L_{i,j} \right|_{X'} \geq \alpha * p^2 \quad (3)$$

where Equation 2 defines the pixel with final label as object, $R(X)$ describes the set of such pixels. $\left|L_{i,j}\right|_X$ computes the number of object labels in the $p * p$ labels of pixel (i, j). Equation 3 defines the pixel with final label as background, $U-R(X)$ describes the set of such pixels, and U describes the related pixels in current frame to compute object contour. $\left|L_{i,j}\right|_{X'}$ describes the number of background label for pixel (i, j). The α in Equation 2 and Equation 3 separately constraints the ratio of being object label and background label, and are set to 0.8 in our experiments.

With Equation 2, the final label of a pixel is set to object label when its object label ratio reaches to 0.8. With Equation 3, the final label of a pixel is set to background label when its background label ratio reaches to 0.8. The other pixels are set to boundary label if they cannot satisfy Equation 2 or Equation 3. Following this, we utilize the rough set scheme to identify the object contour. However, this contour is not continuous and smooth. We improve it by morphology algorithm [19] such as corrosion, expansion, and open/close process, and finally obtain the predicted object contour. Figure 2 describes the process of predicting object

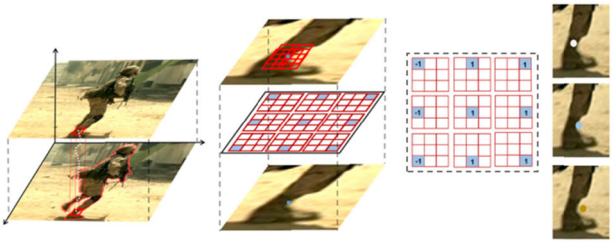


FIGURE 2. Predict object contour by patch-wise labels. For the white pixel in the first column on the right foot of the solider, from left to right: propagate a seam, form patch-wise labels, show all labels (1 for object, -1 for background), decide final label (the top one with white pixel).

contour based on patch-wise labels. First, we compute the patch-wise labels by propagating seam between adjacent frames as the first column. Then, we obtain $p * p$ labels for each pixel as the second column. The third column is an example of the patch-wise labels of a pixel by $3 * 3$ patch. Finally, we use the rough set to decide the final label for a pixel as the last column where the white, blue and orange points separately describe a pixel belonging to object, boundary and background. More details about the achievement of rough set can be reviewed from paper [17].

B. MATCH SUPERPIXELS BETWEEN ADJACENT FRAMES

The predicted object contour based on patch seams is not accurate especially in dealing with segmenting challenges. That's because it is produced only by propagating pixel-level label and cannot consider the structure information of video object. Although the patch-wise seam performs better than the pixel-wise seam, it is still specified by a rigid matrix without considering any semantics.

Therefore, this paper proposes a superpixel match algorithm by greatly utilizing the structure and semantic information of object local region to improve the accuracy of predicted object contour. A superpixel is a local region with irregular boundary, and each pixel in it has similar texture, brightness, color, and structure. Each superpixel describes a part of an object such as the hand or leg of a person, and provides useful middle-level visual cue. We use the good performance of superpixels in representing object parts with specific semantic to improve video segmented result by superpixel match. Our superpixel match algorithm is defined by the following three steps.

(1) First, compute superpixels. we use the SLIC algorithm [21] to do superpixel segmentation on the interest regions from both previous frame and current frame. The interest region is the local region including both target object and its surroundings.

(2) Second, match superpixels between adjacent frames. We match the superpixels on current frame to the ones from previous frame by the distance measuring the similarity between two superpixels. We define the distance by Equation 5 based on the Hausdorff Distance [23]. Supposed $S_{t,i}$ is a superpixel on t frame and $S_{t-1,j}$ is a superpixel from $(t-1)$ frame, the superpixel $S_{t-1,k}$ matched with $S_{t,i}$ is computed by:

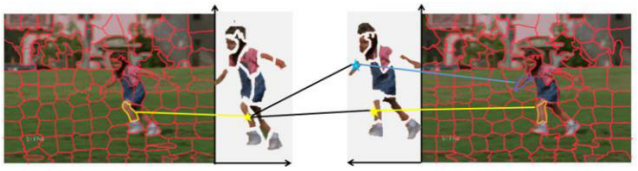


FIGURE 3. MatchSuperpixels between frames. The left yellow superpixel is from current frame. It is matched to the right yellow one by Equation 4 for owning smaller distance than others such as the blue one.

$$min_{j\in U_{t-1}}(D(S_{t,i}, S_{t-1,j})) \tag{4}$$

where U_{t-1} is the set of superpixels from $(t-1)$ frame. The distance D is defined by Equation 5, where H is defined by the two-direction distances $h(S_{t,i}, S_{t-1,j})$ and $h(S_{t-1,i}, S_{t,j})$. σ is set to 10 as the weight.

$$D(S_{t,i}, S_{t-1,j}) = \exp(-\frac{H(S_{t,i}, S_{t-1,j})^2}{2\sigma}) \tag{5}$$

$$H\left(S_{t,i}, S_{t-1,j}\right) = \max(h(S_{t,i}, S_{t-1,j}), h(S_{t-1,i}, S_{t,j})) \tag{6}$$

We define h by equation 7. $N_{t,i}$ describes the pixel number of $S_{t,i}$ while a is a pixel belonging to it. Similarly, b is a pixel of $S_{t-1,j}$ who has the minimum color distance with a. The color distance $\|C_a - C_b\|_2$ between pixel a and b is defined via the Euclidean distance on RGB.

$$h(S_{t,i}, S_{t-1,j}) = \frac{1}{|N_{t,i}|}\sum\nolimits_{a\in S_{t,i}}(\min_{b\in S_{t-1,j}}\|C_a - C_b\|_2) \tag{7}$$

By Equation 4, we compute the matched superpixel for each superpixel from current frame as Figure 3. Then, we calculate the labels for current superpixel based on it.

(3) Third, calculate the labels of superpixels. We obtain the pixel labels for each superpixel on current frame based on its matched superpixel. For any pixel a in $S_{t,i}$, it is related to pixel b from its matched superpixel $S_{t-1,k}$ by $min_{b\in S_{t-1,j}}\|C_a - C_b\|_2$. Therefore, we mark a with the label of b. If half of the pixel labels for a superpixel is the object label, then the superpixel is marked as matched to an object superpixel. Otherwise, it is marked as matched to a background superpixel.

C. OPTIMIZE OBJECT CONTOUR BASED ON MATCHED SUPERPIXELS

The predicted object contour often has errors for the drifts from patch seams. For example, the blue line in Figure 4 describes the predicted object contour. The segmented errors are shown in the dotted rectangles. Using the local structure information, superpixel can produce accurate local contour such as the red boundary of superpixel on left foot. With these local boundaries, the errors of predicted object contour can be greatly reduced. Therefore, we use the boundaries of matched superpixels to optimize the predicted object contour. Three steps are defined to achieve this optimization.

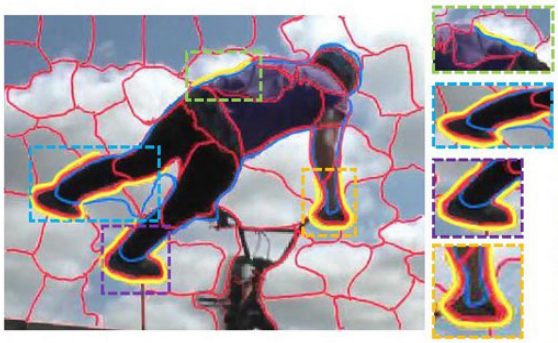


FIGURE 4. Improve the predicted object contour by the boundaries of matched superpixels. Blue line is the predicted object contour, red lines are the boundaries of superpixels, yellow lines are the improved ones.

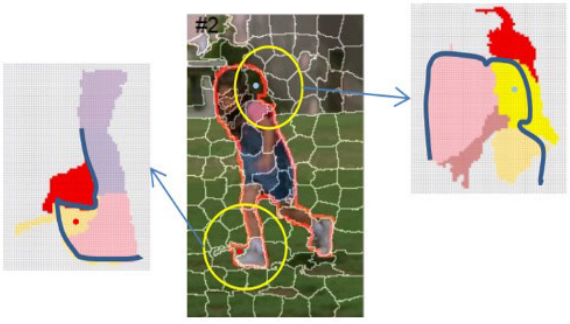


FIGURE 5. The relationships between the matched superpixels with the predicted contour (red line). Each region with white line is a superpixel. The blue lines in the zoomed region is the local predicted object contour.

First, we construct many contour candidates according to the matched superpixels and predicted object contour. We select three kinds of superpixels on current frame to construct contour candidates. The first kind is the superpixel included in the predicted contour but matched to a background superpixel. This kind of superpixel usually drifts the object contour to background such as the superpixel recorded by the red point in Figure 5. The second kind is the superpixel outside the predicted contour but matched to an object superpixel. This kind of superpixel means that part of object contour maybe lost in contour propagation. The third kind is the superpixel which is crossed by predicted contour such as the one recorded by the light blue point in Figure 5. This kind of superpixel usually provides more accurate local contour for video object. With the above three kinds of superpixels and the predicted contour, we construct many contour candidates.

Second, this paper defines an energy function to evaluate each candidate based on multi-feature measures. If R_{t-1} is the segmented result of $(t-1)$ frame and CR_t is a candidate of t frame, the energy function is:

$$E\left(CR_t, R_{t-1}\right) = w_1 * \|C_t - C_{t-1}\|_2 + w_2 * \|G_t - G_{t-1}\|_2 + w_3 * S + w_4 * |L_t - L_{t-1}| \quad (8)$$

FIGURE 6. Optimize object contour by the boundaries of matched superpixels. The predicted object contour (green line in the first column) is optimized by two matched superpixels (yellow circles in the second column) to produce better segmented result (red line in the third column).

where $||*||_2$ is the Euclidean distance of two vectors. C_t, C_{t-1} are color values of object, G_t and G_{t-1} are their centers. L_t and L_{t-1} represent the lengths of object contours, and computes the absolute value of the two contours. S describes the match degree of object shape computed by the Hu moment invariants [34] M_t^i and M_{t-1}^i as defined by Equation 9. We normalize the invariant moments for the Hu moment into seven classifications, then we measure shape match by:

$$S = \sum_{i=1}^{7} \left| \frac{1}{M_t^i} - \frac{1}{M_{t-1}^i} \right| \quad (9)$$

Third, the final optimized object contour is calculated by minimizing the energy function. Then, the final segmented result is the candidate contour with the minimum energy. Both Equation 8 and Equation 1 are energy function, they are achieved by dynamic programming. However, Equation 1 is used for initializing the object contour, Equation 8 is used for computing the final target contour. They are separately used in two steps.

Figure 6 describes how to optimize object contour by the boundaries of matched superpixels. The green line in the first column is the predicted object contour. It included part background near the head. By the energy function of Equation 8, it is greatly improved by two matched superpixels denoted by yellow circles in the second column. Finally, we get the optimized segmented result as the red line in the third column.

IV. EXPERIMENTS

The proposed method is experimented on Intel Core i5-3340@3.10GHz quad-core CPU, 8G, with the Windows10 64-bit PC, MATLAB R2014a, OpenCV 3.3.1 as the execution environment. Our test data includes fourteen videos from the SegTrack-v2 [11], and each video includes an average of 1,066 pixel-level annotations per frame. They cover different video segmentation challenges including object occlusion and deformation, poor lighting conditions, motion blurs, cluttered background and so on. These videos often are used to evaluate the performances of video segmenting methods.

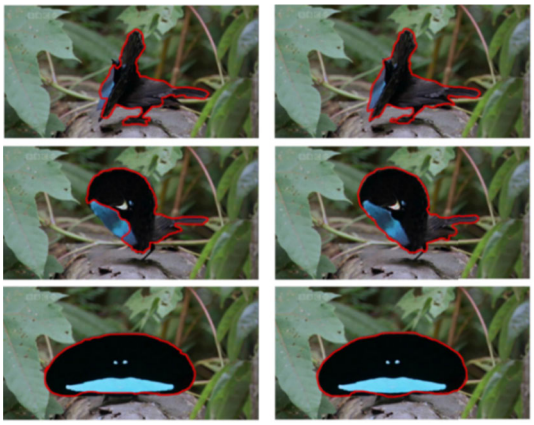


FIGURE 7. Comparisons with SeamSeg by the bird of paradise video. Left column are the SeamSeg results [19], right column are our results.

FIGURE 8. Comparisons with SeamSeg by monkey video. Left column shows the SeamSeg results [19], right column shows our results.

According to the experimental performances of different parameter values, we set our parameters as follows. The patch size in patch seam propagation is $3 * 3$. The SLIC algorithm has an iteration number of 10 and a spatial weight of 25. The parameters of the energy function of the image feature are $\alpha_1 = 2.0, \alpha_2 = 0.7, \alpha_3 = 1.5$, and $\alpha_4 = 0.2$ respectively. Normally, our method costs 5-6 seconds when producing a result for an image with size 300*400. Compared with the learning based segmentation methods and the interactive segmentation methods, the proposed method performs better for costing less time.

A. QUALITATIVE EVALUATION

Our qualitative evaluation includes the comparisons with SeamSeg [19] defined by seam propagation and some other well-known present video object segmentation methods [12], [24]. Comparing with the Seamseg, the improvement by our superpixel match process is demonstrated. Comparing with other method, the better performance by combining patch seam propagation and superpixel match is demonstrated.

1) COMPARISON WITH THE SEAMSEG METHOD

The SeamSeg method [19] firstly achieves the video object segmentation by propagating seams between adjacent frames. It successfully copes with gradually and slowly changes from object and background in video segmentation. However, without considering the middle-level local structure cues, it introduces contour drifts in object occlusion (see Figure 7) and deformation (see Figure 8). Our method successfully deals with these challenges by shrinking the contour parts which have drifted to background or extending the contour parts which have discarded object local regions.

Figure 7 shows the comparisons on video bird of paradise. The object (bird) has a great chang. In the second row, the light blue part of its body reoccluded. By computing pixel-level object contour via seam propagation, the SeamSeg method discards the part of bird wing in the first row, and the reoccluded part of body in the second row. Our method deals with the discards by matching their superpixels into object and successfully extends the predicted object contour to the ideal one as shown in the right column of Figure 7.

Figure 8 describes the comparisons on monkey video. As in left column, the SeamSeg method introduces serious contour drift on the left hand of monkey. This contour drift is propagated frame by frame and introduces more errors from the second row to the third row. Our method matches the superpixels of left hand to monkey body and produces more accurate results by adding these superpixels into target region by our energy function.

2) COMPARISONS WITH OTHER VIDEO SEGMENTATION METHODS

We select two present video segmentation methods proposed in [12], [24] which are well-known for their good performance. The two methods utilize the similar scheme with use to design their algorithms and achieve favorable segmentation results. Wen *et al.* [24] also use superpixels not pixels to construct the features of foreground object. As we using seam propagation to inherit temporal object cues and using superpixel to exploit spatial object cues, Tsai *et al.* [12] employ the spatiotemporal relationship to propagate the target contours in segmenting video object. From Figure 9 to Figure 11, we demonstrate the comparisons in dealing with challenges from fast motion, motion blur, great deformation, illumination changes.

Figure 9 shows the comparisons about drift video. The car undergoes fast motion and motion blur. Using both spatial cues and temporal cues, our method and Tsai's produce more favorable results than Wen's. In addition, by combining pixel-level and superpixel-level cues, our method propagates object contour more accurate than Tsai's on the car tail.

Figure 10 shows the comparisons about parachute video. The parachute contains great illumination changes and fast motion. Our method produces more accurate object contour especially on the left side boundary of parachute than the other two methods. The main reason is our seam prop-

FIGURE 9. Comparisons by drift video.

FIGURE 10. Comparisons by parachute video.

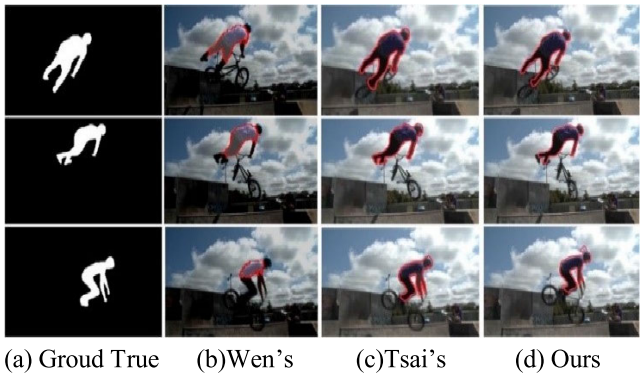


FIGURE 11. Comparisons by BMX video.

agation process can separate object with background on pixel level.

Figure 11 demonstrates the comparisons for BMX video. The person has various deformation and similar background influence in this video. By predicting object contour with seam propagation and optimizing it with superpixel match, our method produces better result than Wen's. It shows our robust in dealing with deformation as demonstrated in Figure 8. Our contour drift besides the head in the third row is introduced by a little superpixel which is more similar to person body than white cloud.

3) OUR RESULTS ON MORE VIDEOS

Figure 12 are ours results from many videos. As shown in the first row and forth row, our method performs good in dealing with various appearance deformation. The third row shows that our method robust in dealing with similar background. According to the fifth row and seventh row, our method successfully deals with the fast motion and occlusion.

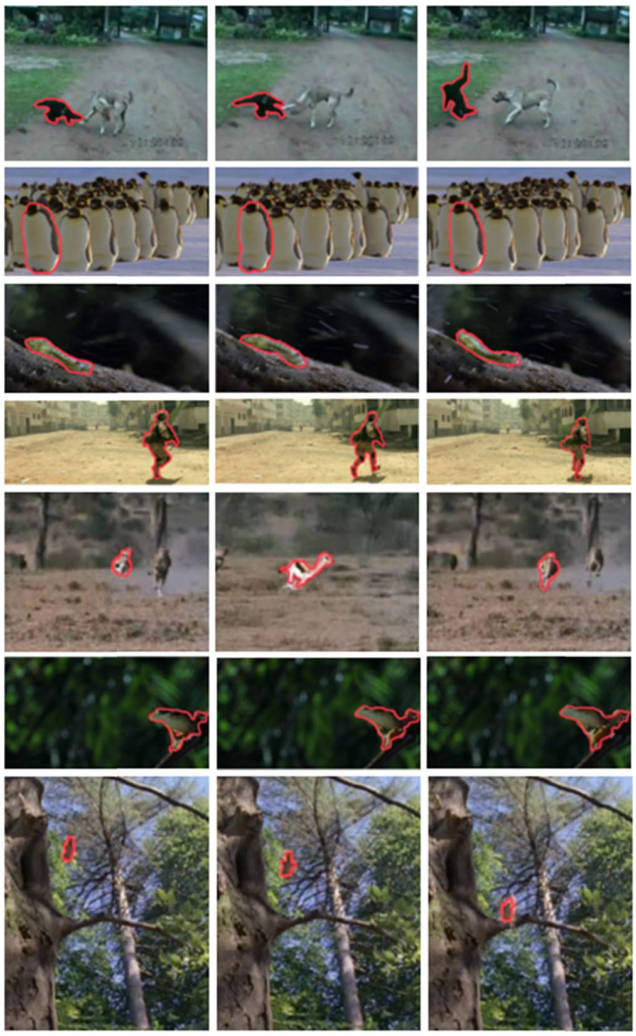


FIGURE 12. Our results on more videos.

B. QUANTITATIVE COMPARATIVE ANALTSIS OF OBJECT TRACKING RESULTS

This paper uses the error measure [19] shown in table 1 and the intersection-over-union (overlap) ratio [12] in table 2 to show our quantitative evaluation. The error measure describes the average number of pixels mis-labelled per frame for each video. The overlap ratio describes the average ratio between the interaction region and union region of the segmented result and ground truth. Sometimes the error measure is sensitive to object size [25], the overlap ration is a good demonstration.

All the videos provided in SegTrack v2 [25] are utilized to do our quantitative evaluation. This public data consists of 14 videos with 24 objects and 947 annotated frames.

TABLE 1. The average error measure of different methods.

Videos	Wen's	Ramakanth's	Ours
Girl	2406	**1361**	**1521**
Birdfall	**160**	**229**	257
Parachute	291	**283**	**211**
Cheetah	1321	**545**	**697**
Monkeydog	**276**	**420**	613
Penguin	571	**371**	**298**
Drift	5618	**3109**	**2378**
Hummingbird	**2808**	**6346**	8212
BMX	3538	**2422**	**1964**
Frog	**4002**	9171	**6097**
Worm	**1083**	2025	**1572**
Soldier	**1165**	1991	**1501**
Monkey	**718**	**845**	1460
Paradise	**2983**	4527	**3498**
Average	**1924**	2403	**2162**

TABLE 2. The overlap rate of different methods.

Videos	Li's	Cai's	Wen's	Tsai's	Wang's	Ours
Girl	**89.2**	62.0	79.7	**87.9**	14	**82.9**
Birdfall	62.5	36.4	**78.1**	57.4	32.5	**68.2**
Parachute	**93.4**	59.3	92.6	**94.5**	69.9	**94.5**
Cheetah	37.3	38.7	**55.1**	33.8	33.1	**65.4**
Monkeydog	**71.3**	25.7	**81.3**	54.4	22.1	**61.5**
Penguin	51.5	40.1	91.4	**93.9**	20.8	**95.1**
Drift	74.8	57.2	63.7	**84.3**	43.5	**88.5**
Hummingbird	**54.4**	25.1	52.9	**69.0**	28.8	30.5
BMX	**85.4**	36.0	68.9	**88.0**	27.9	**85.2**
Frog	**72.3**	38.8	61.8	**81.4**	45.2	34.1
Worm	**82.8**	44.3	77.4	**89.6**	27.4	68.7
Soldier	83.8	54.2	**84.9**	**86.4**	43.0	77.6
Monkey	84.8	58.7	**86.1**	**88.6**	61.7	75.6
Paradise	**94.0**	46.5	91.6	**95.2**	44.3	86.5
Average	74.1	44.5	**76.1**	**78.8**	39.5	**72.5**

We select six well-known segmentation methods including Wen's [24], Ramakanth's [19], Tsai's [12], Li's [25], Cai's [26] and Wang's [27] to do the quantitative comparisons. We choose them for their favorable performances and available data or code. Table 1 focuses on comparisons with the seam-based method [19], [24], while Table 2 is for other well-known methods.

Table 1 shows the average error measure comparisons, in which smaller value means better results. The best result is colored red and the second is blue. Our method produces the best results for videos Parachute, Penguin, Drift and BMX, and seven second results. Compared with Wen's method, our method is more robust by producing the best and second results for 10 videos. Compared with Ramakant's method, our method performs better for owning smaller average error.

Table 2 describes the overlap rates comparisons, in which bigger value means better result. We show the best, second and third result by red, blue and green. Our method performs best in video Parachute, Cheetah, Penguin Drift. Among 14 videos, eight videos of our results are ranked as the best, second or third. The average overlap rate of our method reaches 72.5%. According to both Table 1 and Table 2, it is clear that our method performs robust and produces acceptable results, especially in video Parachute, Penguin, Drift, BMX and Cheetah. That mainly due to our method can effectively deal with the contour drift by some challenges.

V. CONCLUSION

This paper proposes a new robust video segmentation method based on patch seam propagation and superpixel match. By propagating patch seams, the pixel-level labels for object and background are transformed from previous frame to current frame. It greatly uses the inherited and temporal cues to produce the initialized object contour. To reduce contour drift introduced by pixel-level label propagation, we define superpixel match algorithm to utilize the spatial cues to optimize the initialized contour. By matching superpixels between adjacent frames, the boundaries of superpixels formed by semantic division are employed to construct many contour candidates. Then, the final segmented result is the object contour candidate with the minimum value of our proposed energy function. For combing both pixel-level and superpixel-level cue, our method produces more favorable results than most of the optical flow based methods.

By using the temporal cues based on seam propagation and the spatial cues based on superpixel match, our method produces favorable results in dealing with some video object segmentation challenges. Many experiments have done on the SegTrack-v2 data based on all its fourteen videos. Both the

qualitative evaluation and quantitative evaluation shows that our method produces more accurate results and performs more robust than many present methods. Sometimes, our segmented result drifts away from the ideal target position. In the future, we will introduce learning scheme to set more efficient parameters to deal with these problems.

REFERENCES

[1] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7376–7385.

[2] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Fast user-guided video object segmentation by interaction-and-propagation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5247–5256.

[3] S. Caelles, K. . .-K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. V. Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 221–230.

[4] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4481–4490.

[5] K. Xu, L. Wen, G. Li, L. Bo, and Q. Huang, "Spatiotemporal CNN for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1379–1388.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[7] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen, "FEELVOS: Fast End-To-End embedding learning for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9481–9490.

[8] L. Zhang, Y. Lu, L. Lu, and T. Zhou, "Refined video segmentation through global appearance regression," *Neurocomputing*, vol. 334, pp. 59–67, Mar. 2019.

[9] J. Zheng, W. Luo, and Z. Piao, "Cascaded ConvLSTMs using semantically-coherent data synthesis for video object segmentation," *IEEE Access*, vol. 7, pp. 132120–132129, 2019.

[10] J. Li, Y. Zhao, J. Fu, J. Wu, and J. Liu, "Attention-guided network for semantic video segmentation," *IEEE Access*, vol. 7, pp. 140680–140689, 2019.

[11] N. S. Nagaraja, F. R. Schmidt, and T. Brox, "Video segmentation with just a few strokes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3235–3243.

[12] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3899–3908.

[13] H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Superframes, a temporal video segmentation," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 566–571.

[14] C. Huarong, Q. Kanglai, and W. Bin, "Temporal Coherent Video Segmentation with Support Vector Machine and Graph Cut," *J. Comput.-Aided Des. Comput. Graph.*, no. 8, p. 1, Aug. 2017.

[15] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 628–635.

[16] Y. Lu, X. Bai, L. Shapiro, and J. Wang, "Coherent parametric contours for interactive video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 642–650.

[17] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[18] L. Fan, W.-C. Wang, F. Zha, and J. Yan, "Exploring new backbone and attention module for semantic segmentation in street scenes," *IEEE Access*, vol. 6, pp. 71566–71580, 2018.

[19] S. A. Ramakanth and R. V. Babu, "SeamSeg: Video object segmentation using patch seams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 376–383.

[20] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–9, Aug. 2008.

[21] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to State-of-the-Art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[22] Z. Pawlak, "Rough set approach to knowledge-based decision support," *Eur. J. Oper. Res.*, vol. 99, no. 1, pp. 48–57, May 1997.

[23] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2010.

[24] L. Wen, D. Du, Z. Lei, S. Z. Li, and M.-H. Yang, "JOTS: Joint online tracking and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2226–2234.

[25] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2192–2199.

[26] Z. Cai, L. Wen, Z. Lei, N. Vasconcelos, and S. Z. Li, "Robust deformable and occluded object tracking with dynamic graph," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5497–5509, Dec. 2014.

[27] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1323–1330.

[28] Z. Wang, J. Xu, L. Liu, F. Zhu, and L. Shao, "RANet: Ranking attention network for fast video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3978–3987.

[29] J. Johnander, M. Danelljan, E. Brissman, F. S. Khan, and M. Felsberg, "A generative appearance model for End-To-End video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8953–8962.

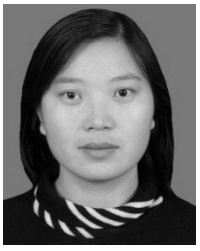
[30] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.

[31] S. Xu, D. Liu, L. Bao, W. Liu, and P. Zhou, "MHP-VOS: Multiple hypotheses propagation for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 314–323.

[32] Y. T. Hu, J. B. Huang, and A. G. Schwing, "Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation," in *Proc. IEEE Eur. Conf. Comput. Vis. (ECCV)*, pp. 786–802, Sep. 2018.

[33] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 985–998, Apr. 2019.

[34] A. Sit and D. Kihara, "Comparison of image patches using local moment invariants," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2369–2379, May 2014.



YUN LIANG was born in Linyi, Shandong, China, in 1981. She received the M.Sc. and Ph.D. degrees from the School of Information Science and Technology, Sun Yat-sen University, in 2005 and 2011, respectively.

From 2005 to 2012, she was an Assistant Professor with the Informatics College, South China Agriculture University, Guangzhou, China, where she has been an Associate Professor with the College of Mathematics and Informatics, since 2013. From 2016 to 2017, she was with Simon Fraser University. Her research interests include computer vision, image computation, machine learning, and so on.

YUQING ZHANG was born in Tangshan, Hebei, China, in 1999. She is currently pursuing the degree in computer science and technology with the College of Mathematics and Informatics, South China Agriculture University, Guangzhou, China. Her research interests include image processing, video object segmentation, and video editing.

YIHAN WU was born in Shanwei, Guangdong, China, in 1999. He is currently pursuing the degree in software engineering with the College of Mathematics and Informatics, South China Agriculture University, Guangzhou, China. His research interests include image segmentation, video object segmentation, and video tracking.

SHUQIN TU was born in Jiangxi, China, in 1977. She received the M.Sc. degree from the School of Computer Science, South China University of Technology, Guangzhou, China, in 2004, and the Ph.D. degree from the College of Engineering, South China Agricultural University, in 2014. Since 2004, she has been an Assistant Professor with the College of Mathematics and Informatics, South China Agriculture University. Her research interests include image processing, smart agriculture, machine learning, and so on.

CAIXING LIU was born in Shaoguan, Guangdong, China, in 1962. He received the bachelor's degree from the Computer Science College, Nanjing University, in 1984. He is currently a Professor with the College of Mathematics and Informatics, South China Agriculture University, Guangzhou, China. His research interests include artificial intelligent, smart agriculture, computer vision, and so on.