SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2018, Vol. 61 012205:1–012205:12
doi: 10.1007/s11432-016-9150-x

Automatic salient object sequence rebuilding for video segment analysis

Tie LIU[1*], Haibin DUAN[1], Yuanyuan SHANG[2], Zejian YUAN[3] & Nanning ZHENG[3]

[1]*School of Automation Science and Electrical Engineering, Beihang University, Beijing* 100191*, China;*
[2]*Information Engineering College, Capital Normal University, Beijing* 100048*, China;*
[3]*The School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049*, China*

Received 3 May 2017/Accepted 30 June 2017/Published online 29 September 2017

Abstract Detection of salient object sequences from video data is challenging when the salient object changes between consecutive frames. In this study, we addressed the salient object sequence rebuilding problem with video segment analysis. We reformulated the problem as a binary labeling problem, analyzed the potential salient object sequences in the video using a clustering method, and separated the salient object sequence from the background by applying an energy optimization method. Our proposed approach determines whether temporal consecutive pixels belong to the same salient object sequence. The conditional random field is then learned to effectively integrate the salient features and the sequence consecutive constraints. A dynamic programming algorithm was developed to resolve the energy minimization problem efficiently. Experimental results confirmed the ability of our approach to address the salient object rebuilding problem in automatic visual attention applications and video content analysis.

Keywords salient object, video attention, sequence segment analysis, conditional random model

Citation Liu T, Duan H B, Shang Y Y, et al. Automatic salient object sequence rebuilding for video segment analysis. Sci China Inf Sci, 2018, 61(1): 012205, doi: 10.1007/s11432-016-9150-x

1 Introduction

The human visual attention attracts many researchers from physiology, psychology, neural systems, and computer vision. A range of visual attention models have been proposed to represent human vision [1], and these have found applications in unmanned police patrols and autonomous mobile robots [2]. In these models, a single object is usually designated as the target to be recognized or detected. Salient objects are those that are of most interest within the image, and their study has been an important area of visual attention research [3–6]. They have been used in automatic image cropping [7], adaptive image displays for small devices [8], and image collection browsing [9]. A range of methods have been proposed for saliency detection [1, 10–13], and were reviewed extensively by Itti in a recent paper [1]. A similar approach to video images has been attracting growing interest, having applications in areas such as video surveillance, augmented reality, and digital video editing. A recent example applied salient object detection to sequences of images [4]. In most applications, a challenge arises from the dynamic and complex scenes recorded in a video, in which the salient object changes from frame to frame due to the movement of the object itself, or of the recording device. Figure 1 shows an example of this. In this study, we investigated automatic salient object sequence rebuilding, a common problem in salient object tracking and video content analysis.

* Corresponding author (email: liutiel@163.com)

Figure 1 (Colore online) An example in which the salient object changes across consecutive images (frame #10, #13, #15, #22).

This is a special case of the salient object detection and tracking problem that is encountered in all practical visual systems. The approach used in [4] supposed that only one salient object is present in the video, and that a global appearance model can be built and integrated in a conditional random field framework. In contrast with this approach, we assumed that multiple salient object sequences will be present, making the single global appearance model invalid. We compared the rebuilding problem with the salient object tracking problem, which is in some respects different from traditional object tracking. The most fundamental difference is that the salient objects in a video are not specified in advance, so that the visual features cannot be pre-determined, whereas most object tracking approaches assume priors for the object being tracked or user actions [14, 15], and build on these in a recursive way. In salient object tracking, the constraints are defined consecutively, and make no presuppositions about the category, size, or visual features of the object. This is similar to salient object detection, but differs in that temporal coherence information is used in object tracking [16].

However, the assumption that the same salient object will appear in consecutive frames is not always valid, particularly when the salient object changes from one frame to another in the video. In this case, the salient object must be rebuilt to capture the target, which poses a challenge for the design of the tracking algorithm. Most previous work on object tracking algorithms presupposes a stable object [4, 17–19]. In [17], a constant velocity motion model was applied to the object, and the visual appearance model was assumed to remain stable throughout the video. In [4], a video segment with a single salient object sequence was considered. In contrast, our study addressed the problem of salient object rebuilding when the object changes between consecutive frames, and when multiple objects are present.

Previous studies have addressed salient object tracking without the need for prior information [16]. In [16], the particle filter algorithm [20] was applied to saliency based object tracking and a saliency map was computed using Monte Carlo importance sampling. This allowed detection to switch between salient regions as the salient object changed, while a sampling method limited the global optimization of the salient object sequence. The use of global information has been shown to be useful in salient object sequence tracking, requiring a global optimization problem to be resolved when addressing the salient object rebuilding problem.

The study makes two main contributions. First, we reformulated the automatic salient object sequence rebuild problem as an energy minimization problem in a conditional random field framework. We then considered whether the salient object should be rebuilt, and used a dynamic programming algorithm to resolve the energy minimization problem. Second, we proposed a novel approach to salient object sequence segment analysis (SSA), in which the potential salient object sequences were clustered. To extract the segments, we applied the computed salient features. The sequences were then clustered, and each potential pixel from the image was assigned to a segment. The variables were computed to decide whether pixels from two consecutive frames belonged to the same salient object sequence. Finally, the SSA results were integrated with the energy optimization problem, and a dynamic programming algorithm was developed to resolve the energy minimization problem efficiently. Experimental results demonstrated the effectiveness of the algorithm.

Figure 2 (Color online) An example in which multiple salient objects appear, while the sequence index is defined to distinguish between the different salient object sequences. Previous approaches [3,4] assume a single salient object sequence, and output one rectangle for all the salient objects.

2 Problem formulation

We first formulate the salient object sequence rebuild problem as a binary labeling problem using a conditional random field framework, following [21], but add consideration of the salient object sequence index. From Figure 2, it can be seen that, when multiple salient objects appear, the sequence index can be defined so that the different salient object sequences are distinguished. The video segment is represented as a sequence of images $I_{1,\dots,T}$. Given an image I_t at time t, the salient object is represented as a binary mask $A_t = \{a_x^t\}$, and the sequence index $S_t = \{s_x^t\}$ is increased to distinguish between the different salient object sequences. Each pixel x is given an index s_x^t to indicate which salient object sequence this pixel belongs, and s_x^t is also written as $S(x_t)$ in the following. For each pixel x, $a_x^t \in \{1, 0\}$ is a binary label to indicate whether the pixel x belongs to the salient object. For sequential images $\{I_t\}$, $t \in \{1, \dots, N\}$, the probability of the sequential binary maps, $\{A_t\}$, $t \in \{1, \dots, N\}$, and the salient object sequence index S_t, can be modeled as a conditional distribution

$$P(A_{1,\dots,N}, S_{1,\dots,N}|I_{1,\dots,N}) = \frac{1}{Z}\exp(-E(A_{1,\dots,N}, S_{1,\dots,N}|I_{1,\dots,N})), \tag{1}$$

where $E(A_{1,\dots,N}, S_{1,\dots,N}|I_{1,\dots,N})$ is the energy function and Z is the partition function.

The energy function $E(A_{1,\dots,N}|I_{1,\dots,N})$ describes the saliency in a single image and the temporal constraint in a sequence. To address the problem of salient object rebuilding from consecutive images, we propose a salient SSA approach in which the sequence analysis is applied to the full video. The analysis is still modeled on the energy function. We next consider the salient feature, and model the salient object A_t and the consecutive salient objects A_t, A_{t-1} in the energy function. The energy function $E(A_{1,\dots,N}, S_{1,\dots,N}|I_{1,\dots,N})$ can then be decomposed as follows:

$$E(A_{1,\dots,N}, S_{1,\dots,N}|I_{1,\dots,N}) = \sum_{t=1}^{N}(E(A_t|I_t) + E(A_t, A_{t-1}, S_t, S_{t-1}|I_{1,\dots,N})). \tag{2}$$

In this formulation, the first item $E(A_t|I_t)$ models the salient object constraint from the current image, while the second item $E(A_t, A_{t-1}, S_t, S_{t-1}|I_{1,\dots,N})$ models the consecutive constraints from the salient object sequence. We will discuss these separately. In the first image, when $t = 1$, only the energy item $E(A_t|I_t)$ is counted. If $a_x^1 = 1$, then the salient object sequence $S(x_1) = 1$. Otherwise $S(x_1) = 0$. The salient object sequence $A_{1,\dots,N}$ can then be resolved by minimizing the energy

$$A^*_{1,\dots,N}, S^*_{1,\dots,N} = \arg\min_{A_{1,\dots,N}} \sum_t (E(A_t|I_t) + E(A_t, A_{t-1}, S_t, S_{t-1}|I_{1,\dots,N})). \tag{3}$$

To achieve this, the state space is modeled as a large 3D graph constructed from spatial-temporal space. To resolve such a 3D graph efficiently, following [3], we represent A_t as a rectangle R_t, greatly decreasing the state space. The pixels are assigned as $a_x^t = 1$ if $x_t \in R_t$, and $a_x^t = 0$ otherwise. For convenience, we also represent a_x^t as a_x as a default.

Figure 3 Salient object features from Figure 1.

2.1 Salient object discovery

To discover the salient object, the energy $E(A_t|I_t)$ can be defined using static or/and dynamic salient features. To simplify the calculation, we define $E(A_t|I_t)$ as the linear combination of salient feature constraint $F_k(a_x^t, \cdot)$ and the spatial coherence constraint $M(a_x^t, a_{x'}^t, I_t)$, following [3]. This yields the following:

$$E(A_t|I_t) = \sum_x \sum_{k=1}^{K} \lambda_k F_k(a_x^t, \cdot) + \sum_{x,x'} \beta M(a_x^t, a_{x'}^t, I_t), \tag{4}$$

where λ_k is the parameter for multiple salient feature constraint $F_k(a_x^t, \cdot)$, and can be trained by the learning algorithm together with the parameter β.

The salient feature $F_k(a_x^t, \cdot)$ indicates the important regions in the image, and, following [3], is computed using the local, regional, and global salient features in image I_t with learned parameters. $F_k(a_x^t, \cdot)$ is formulated from a normalized feature map $f_k(x, \cdot) \in [0, 1]$ for every pixel, and is written as follows:

$$F_k(a_x^t, \cdot) = \begin{cases} f_k(x, \cdot), & a_x = 0, \\ 1 - f_k(x, \cdot), & a_x = 1. \end{cases} \tag{5}$$

The salient feature map $f_k(x, \cdot)$ represents the possibility that the pixels belong to the salient object, and the salient feature constraints define the penalty term for incorrect assignment of the salient object labels. The combined salient features from Figure 2 are shown as Figure 3, where it can be seen that, when two objects appear within an image, the salient features cannot be reliably figured out a salient object.

The spatial coherence constraint $M(a_x^t, a_{x'}^t, I_t)$ exploits the relationship between two adjacent pixels. Following the contrast-sensitive potential function used in interactive image segmentation [22], we define $M(a_x^t, a_{x'}^t, I_t) = |a_x - a_{x'}| \exp(-\lambda d_{x,x'})$, where $d_{x,x'} = ||I_x - I_{x'}||^2$ is the $L2$ norm of the color difference. λ is a robust parameter that weights the color contrast and, following [23], can be set as $\lambda = (2\langle||I_x - I_{x'}||^2\rangle)^{-1}$, with $\langle\cdot\rangle$ being the expectation operator. This feature function can be viewed as a penalty term when adjacent pixels are assigned different labels. The parameter β is learned from an image data set, following [3].

2.2 SSA

The energy $E(A_t, A_{t-1}|I_{1,\ldots,N})$ models the temporal coherent constraint between two consecutive images when the salient objects are from the same salient object sequence. To address the salient object rebuild problem, we must judge whether the neighboring pixels from consecutive frames belong to the same salient object sequence. The energy $E(A_t, A_{t-1}|I_{1,\ldots,N})$ is defined as follows:

$$E(A_t, A_{t-1}, S_t, S_{t-1}|I_{1,\ldots,N}) = \sum_{x,x'} C(a_x^t, a_{x'}^{t-1}, s_x^t, s_{x'}^{t-1}, I_{1,\ldots,N}). \tag{6}$$

We simplify $C(a_x^t, a_{x'}^{t-1}, s_x^t, s_{x'}^{t-1}, I_{1,\ldots,N})$ as the temporal coherence constraint $C(a_x^t, a_{x'}^{t-1}, \cdot)$ which models the temporal similarity between salient objects from two consecutive frames.

To decompose the problem, we define $\delta(x_t, x'_{t-1})$ to indicate whether two neighboring pixels from consecutive frames belong to the same sequence. The temporal coherence constraint is effective only

when applied to pixels from the same salient object sequence. We can compute $\delta(x_t, x'_{t-1})$ by analyzing the salient object sequence segments, which are introduced in the next section. The temporal coherence constraint $C(a_x^t, a_{x'}^{t-1}, \cdot)$ can then be written as follows:

$$C(a_x^t, a_{x'}^{t-1}, \cdot) = \delta(x_t, x'_{t-1}) D(a_x^t, a_{x'}^{t-1}, \cdot) + (1 - \delta(x_t, x'_{t-1})) D_s, \tag{7}$$

where $D(a_x^t, a_{x'}^{t-1}, \cdot)$ is defined to describe the similarity between two neighboring pixels from the same salient object sequence, and D_s is a small penalty from the switch between two salient object sequences.

Following [4], we define the similarity $D(a_x^t, a_{x'}^{t-1}, \cdot)$ as the similarity of two salient objects, with centers on x_t, x'_{t-1}. To be efficiently, only the similarity in shape of two salient objects is considered, and we then set

$$D(a_x^t, a_{x'}^{t-1}, \cdot) = ||x_t - x'_{t-1}||^2 + \gamma ||s_t - s_{t-1}||^2, \tag{8}$$

where s_t is the scale of the rectangle enclosing a salient object, and γ is a weighting between location difference and scale difference, following [24]. This places a smoothness constraint on the same salient object sequence.

3 Our approach

As defined by the above formulation, the temporal coherence constraint should be effective only when the neighboring pixels belong to the same salient object sequence. To decompose this problem, $\delta(x_t, x'_{t-1})$ defines whether two neighboring pixels from consecutive frames belong to the same sequence. In this section, we introduce the SSA algorithm used to calculate $\delta(x_t, x'_{t-1})$. This has two steps: the potential salient object position is extracted independently from each frame, then salient object sequence clustering is applied to these potential positions in the whole video to get the salient object sequences.

3.1 Potential salient object position extraction

In the case of video, SSA of all pixels is impractical. Instead, we extract the potentially salient object positions to reduce the whole state space, then apply SSA to these potentially salient pixels. This removes the need for precise salient object labelling when calculating of $\delta(x_t, x'_{t-1})$, as only the potentially salient object pixels are calculated. To speed up the calculation, the algorithm is run in three steps, as follows.

First, we compute the salient map for each frame in the same way as the salient feature constraint, and define $F(a_x, \cdot) = \sum_{k=1}^{K} \lambda_k F_k(a_x^t, \cdot)$ and $f(x, \cdot) = \sum_{k=1}^{K} \lambda_k f_k(x, \cdot)$. These are formulated as in (5), allowing the calculated salience map for salient object discovery to be leveraged directly.

Next, we define $T(x^*, \cdot)$ to measure the potential salient object positions

$$T(x^*, \cdot) = \min_{R_{x^*}} \frac{\sum_{a_x=1:x\in R; a_x=0; x\notin R} F(a_x, \cdot)}{\sum_{x\in R} f(x, \cdot)} + \frac{\sum_{a_x=1:x\in R; a_x=0; x\notin R} F(a_x, \cdot)}{\sum_{x\notin R} f(x, \cdot)}, \tag{9}$$

where R_{x^*} is the rectangle surrounding x^* with the potential size of $[0.1, 0.7] \times \min(w, h)$, with aspect ratios of $\{0.5, 0.75, 1.0, 1.5, 2.0\}$. Here, w and h are the width and height of the image.

Finally, we extract the potential salient object positions by applying a threshold, as follows:

$$X^* = \{x^* : T(x^*, \cdot) < T_0\}, \tag{10}$$

where T_0 is computed from the statistical value of $T(x, \cdot)$ for those positions labeled salient object. We use a Gaussian function to fit the distribution of $T(x, \cdot)$, and compute T_0 as one standard deviation left of the mean of the function. We apply the mean-shift algorithm to combine the positions X^* with overlapped rectangles, and represent the final potential positions as $X = \{X_n\}$.

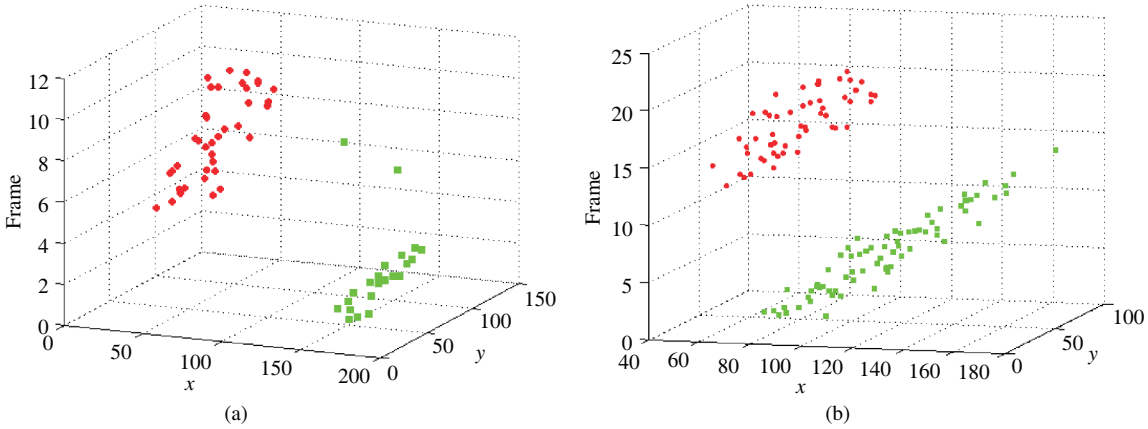

Figure 4 (Color online) Clustering result of SSA with different segments marked by red and green points. (a) SSA for Figure 1; (b) SSA for Figure 10.

3.2 Salient object sequence segment clustering

Figure 4 shows the potential salient object positions from a spatial-temporal viewpoint. Each spatial position corresponds to a 3D point $m_n = [X_n, t_n]$ in the video volume, where t_n is the temporal location (frame number). Spectral cluster methods are used to extract the salient object sequences, following [24]. Given a set of points $\{m_n\}$ in R^3, spectral clustering builds an affinity matrix A then clusters the data points based on the eigenvector analysis of the Laplacian matrix of D_{ij}:

$$D_{ij} = \exp(-||x_i - x_j||^2/2\delta_x^2 - ||x_i - x_j||^2/2\delta_t^2), \tag{11}$$

where the scaling parameters δ_x and δ_t are computed following [24].

The output of the spectral clustering is a set of sequence segments $\{S_k\}$ in which each sequence segment S_k contains a set of points m_n, as shown in Figure 4. We define the sequence segments related to frame I_t as follows:

$$S^t = \{S_k : t_n = t,\ [X_n, t_n] \in S_k\}, \tag{12}$$

where $[X_n, t_n]$ is the 3D point. The key characteristic of salient object tracking is that the most salient object in the frame is output. However, overlaps between sequence segments may appear, as it takes time to rebuild the salient object. In this case, S_t may contain more than one salient object sequence.

Each pixels x_t in image I_t is assigned a salient object sequence with the minimal distance as follows:

$$S(x_t) = \arg\min_{S \in S^t} \min_{[X_n, t_n] \in S} ||X_t - X_n||_2, \tag{13}$$

where $||X_t - X_n||_2$ is the L_2 norm of the spatial distance. We can then compute $\delta(x_t, x'_{t-1})$ to decide whether two consecutive pixels belong to the same salient object sequence

$$\delta(x_t, x'_{t-1}) = \begin{cases} 1, & \text{if} \quad S(x_t) = S(x'_{t-1}), \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

From the definition of $\delta(x_t, x'_{t-1})$, the temporal coherence constraint will be effective only when neighboring pixels from consecutive frames belong to the same salient object sequence. This represents the salient object rebuild problem in energy minimization.

4 Optimization and the algorithm

To resolve this energy minimization problem, we developed the dynamic programming algorithm. First, the rectangle R_t enclosing the salient object is defined as the state variable (with the position and the

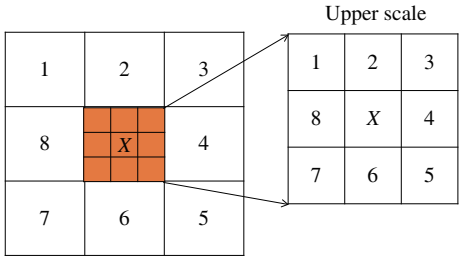

Figure 5 (Color online) A coarse-to-fine algorithm speeds up the dynamic programming of a large 3D graph.

size of the rectangle). Following [4], we then define U_t as the possible strategy between two consecutive frames, so that the optimal value function can be written as

$$O_t(R_t) = \arg\min_{U_t} O_{t-1}(R_{t-1}) + E(A_t|I_t) + E(A_t, A_{t-1}, S_t, S_{t-1}|I_{1,\ldots,N}), \tag{15}$$

where the initial object function is $O_0(\cdot) = 0$. A forward algorithm is used to resolve the salient object sequences $R_{1,\ldots,T}$. However, resolving this energy minimization problem using a dynamic programming algorithm is challenging. We therefore designed the algorithm acceleration introduced in the following section.

4.1 Algorithm acceleration

To accelerate the resolution of the 3D graph and reduce the computing cost associated with the huge state space, a coarse-to-fine algorithm was introduced. In the algorithm design, all salience maps are down-sampled into pyramids. At the coarsest scale, the whole state space is searched, whereas at the upper scale, only neighboring state space is searched. The process is shown in Figure 5.

However, when the salient object switches between frames, the coarse-to-fine algorithm becomes invalid, because of the limitation of the search space. To address this, when multiple sequence segments exist in a frame, the whole state space is searched. Otherwise the original algorithm is used. The coarse-to-fine algorithm accelerates the dynamic programming and makes energy minimization on the 3D graph possible.

4.2 Flow chart of proposed algorithm

As noted above, the proposed salient object detection algorithm include salient object discovery and salient object sequence segmentation analysis. Here we extend this to three steps.

(1) The salient features and the pairwise features of each frame are computed. We compute the local, regional, and global salient features and combine them using the learned parameters. This follows the approach in [3].

(2) Salient object SSA. The 2D potential salient object positions are extracted using (9) and (10). The salient object sequences are then clustered using the spectral clustering algorithm, and each pixel is assigned to a salient object sequence using (12) and (13). Finally, $\delta_t(a_x^t, a_{x'}^{t-1})$ is computed using (14).

(3) The dynamic programming algorithm is run to locate the salient object sequences, by deciding whether two consecutive pixels are from the same salient object sequence.

A flow chart of the proposed algorithm is shown as Figure 6.

5 Experiments

As our interest is in video content analysis, we collected from the internet videos containing at least two different segments of salient object sequences. In frames where two different salient object sequences overlap, the salient object rebuild problem suggests that confusion will arise about their locations. Videos in which two salient objects appear are very challenging to address using the detection algorithm from [3].

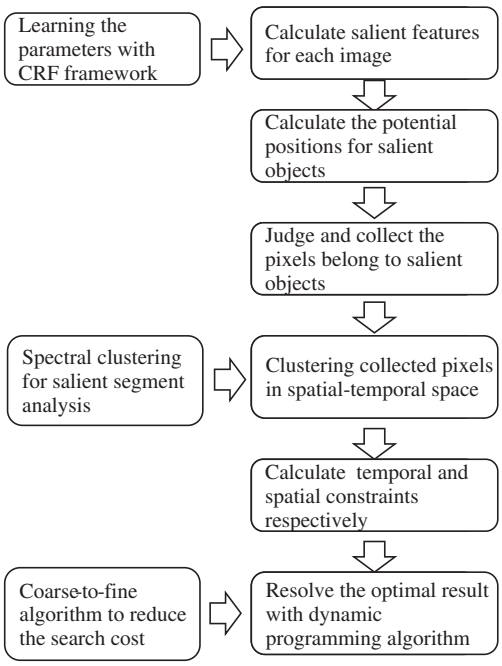

Figure 6 Flow chart of the proposed algorithm.

To test the effectiveness of our proposed algorithm, we conducted three experiments. First, we investigated the effectiveness of the SSA, by comparing the results from our proposed algorithm with those from an algorithm that did not apply such analysis. Second, we compared the performance of the proposed algorithm with that of a salient object tracking algorithm using the particle filter algorithm. Third, we applied our proposed algorithm to the vision system of an unmanned aerial vehicle.

5.1 Effectiveness of SSA

As shown in Figure 7, most of the collected videos used in our experiments contain two salient object sequences. As shown in Figure 7(a), the salient object detection algorithm from [3] was found to perform poorly when a second salient object appeared. In these cases, the algorithm detected a small salient region as the results, and the precision/recall/F-measure for these frames were very poor. To investigate the effectiveness of applying SSA to the salient object rebuild problem, we compared our approach with one in which no SSA was used, and the temporal coherence constraint was defined on all neighboring pixels. As shown in Figure 7(b), errors appeared in frames introducing the new salient object which required the dynamic programming algorithm to search in a large state space. In contrast, our approach analyzed the salient object sequence segment, and introduced the temporal coherence constraint only when two pixels belong to the same salient object sequence. As can be seen from Figure 7(c), our approach was able to deal effectively with the salient object rebuilding problem.

We collected 20+ videos containing multiple salient object sequences. The SSA was able to distinguish between the salient objects in 85 percent of the frames in which multiple objects appeared. This represented an improvement in precision of 155% with comparable recall rate. Figure 8 shows further examples in which multiple salient object sequences appear (a) a car sequence showing multiple thatched cottages; (b) a scene in which two different people appear successively, causing the salient object sequence to be switched; (c) a person walking past a static car, in which attention is focused on the moving person; (d) a person walking in front of a sculpture, replacing the sculpture as the salient object. These are typical scenes in which the multiple salient object sequences appear, or the salient object switches. We applied the sequence salient analysis to test the effectiveness of the proposed approach. The results are shown in Figure 9. The clustering results of the multiple salient object sequences demonstrated the ability of the proposed approach to distinguish the different salient object sequences. The sequence index can then be

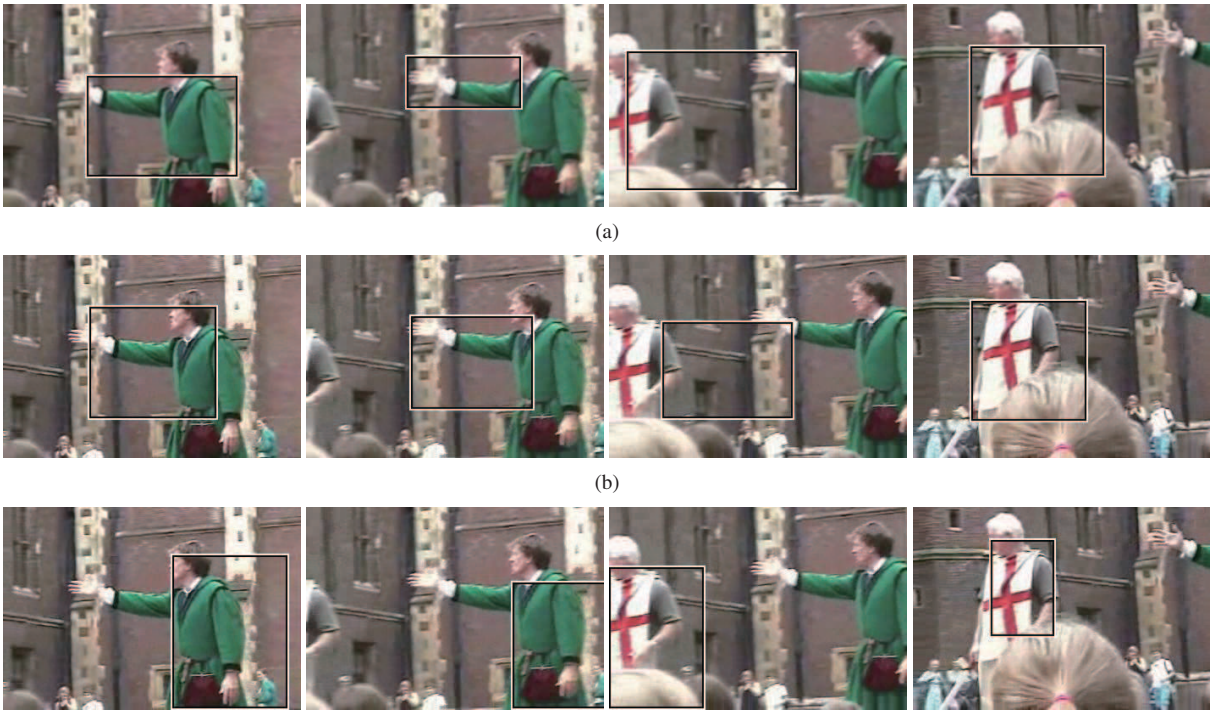

Figure 7 (Color online) Effectiveness of SSA. (a) Salient object detection algorithm from [3] with a single image; (b) salient object tracking without SSA; (c) our approach.

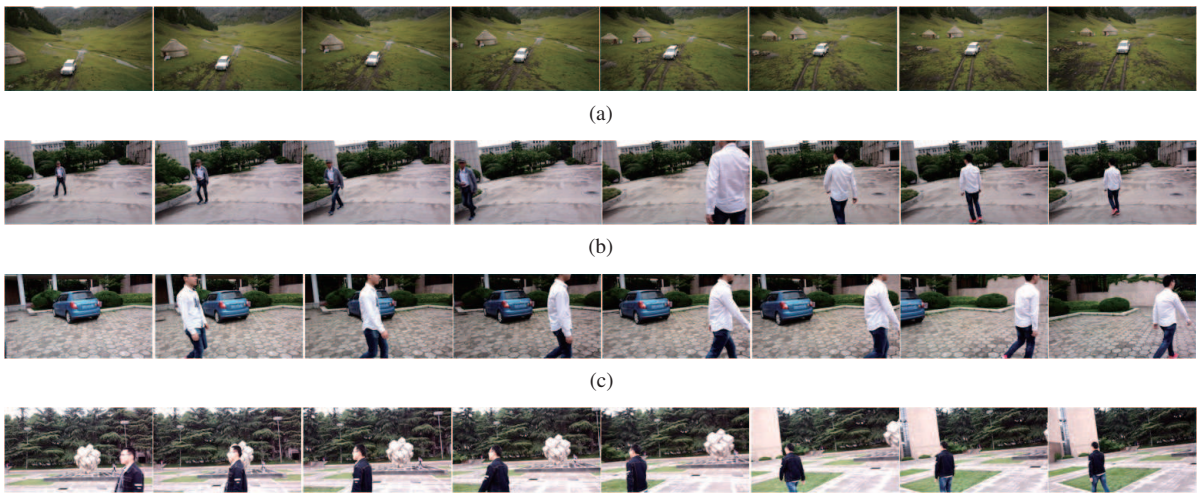

Figure 8 (Color online) Examples used to compare the effectiveness of our approach with SSA. (a) Car sequence with cottages; (b) two different people appearing successively; (c) a person walks past a car; (d) a person walks in front of a sculpture.

integrated with the global optimization framework to detect multiple salient object sequences.

5.2 Comparison with salient object tracking algorithm

We next compared our approach with the salient object tracking algorithm from [16], in which the posterior probability distribution is designed using the salient maps and the particle filter algorithm is leveraged.

For videos in which a new salient object appears, the posterior probability distribution from the

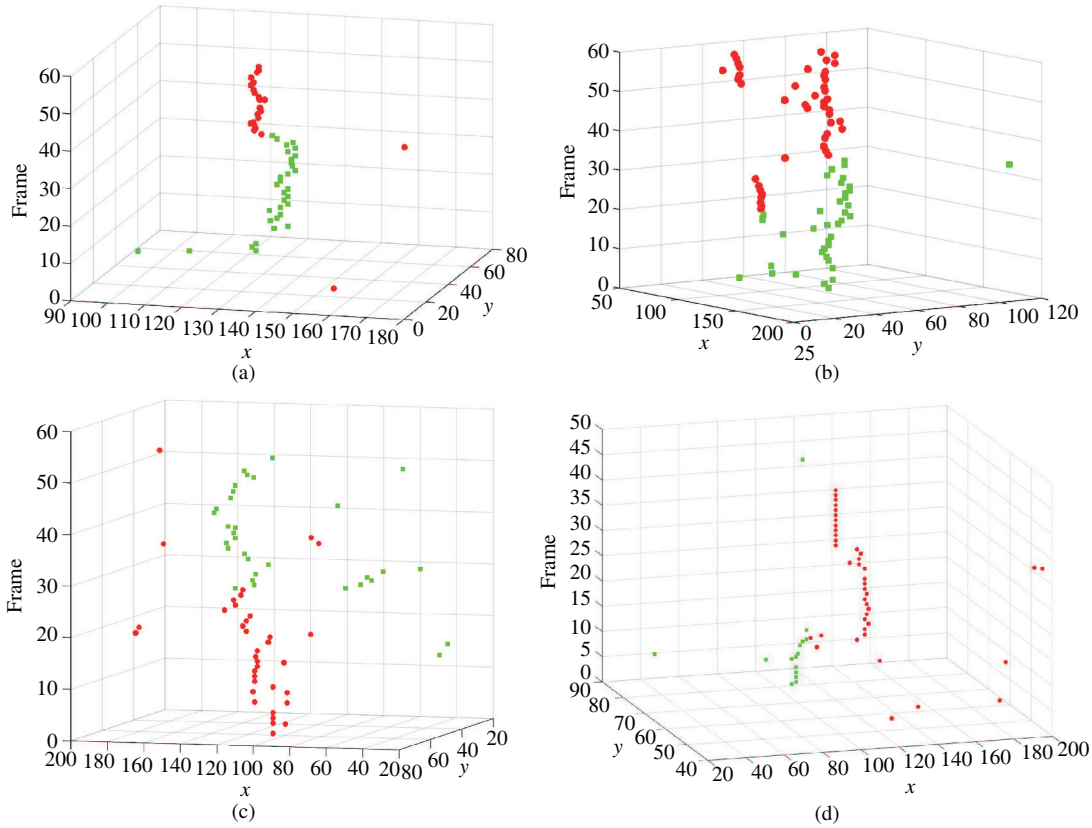

Figure 9 (Color online) Clustering result of SSA with different segments marked by red and green points, from the examples in Figure 8. (a) SSA for Figure 8(a); (b) SSA for Figure 8(b); (c) SSA for Figure 8(c); (d) SSA for Figure 8(d).

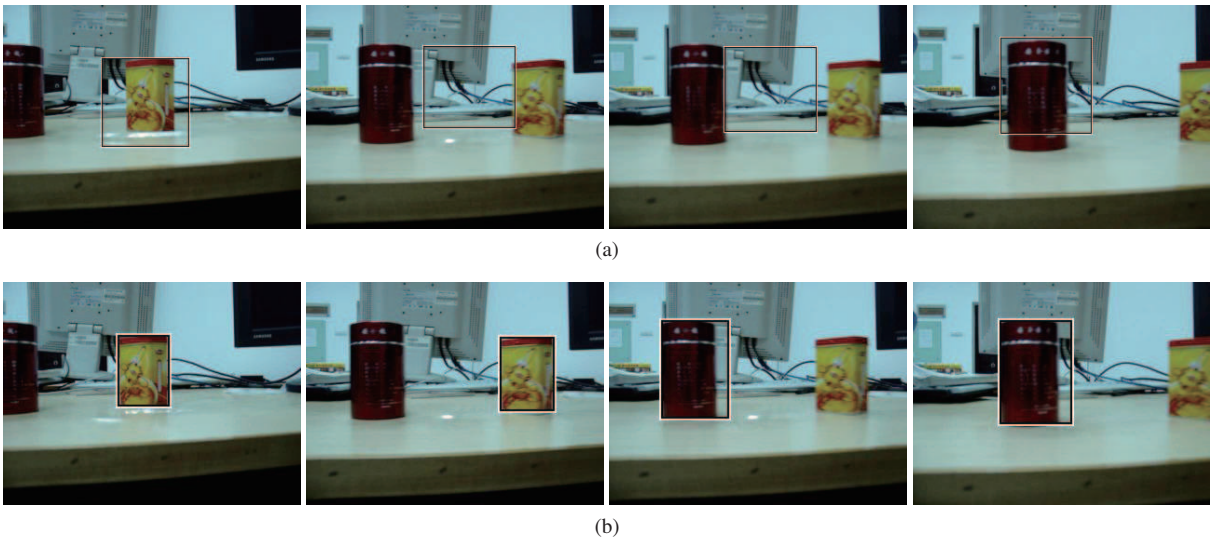

Figure 10 (Color online) Comparison of algorithms. From left to right: frame #5, #12, #13, #17. (a) Zhang's approach [16]; (b) our approach.

salient maps is usually inconsistent, and this may introduce confusion to the sampling function. Further, when the salient object shifts, the sampling function is unable to capture the new object because of the limitation of the sampling space. Figure 10(a) shows an example in which salient object tracking is lost in frames #12 and #13. In contrast Figure 10(b) shows that our approach was able to address the salient object rebuild problem, switching the calculated salient objects between frames #12 and #13.

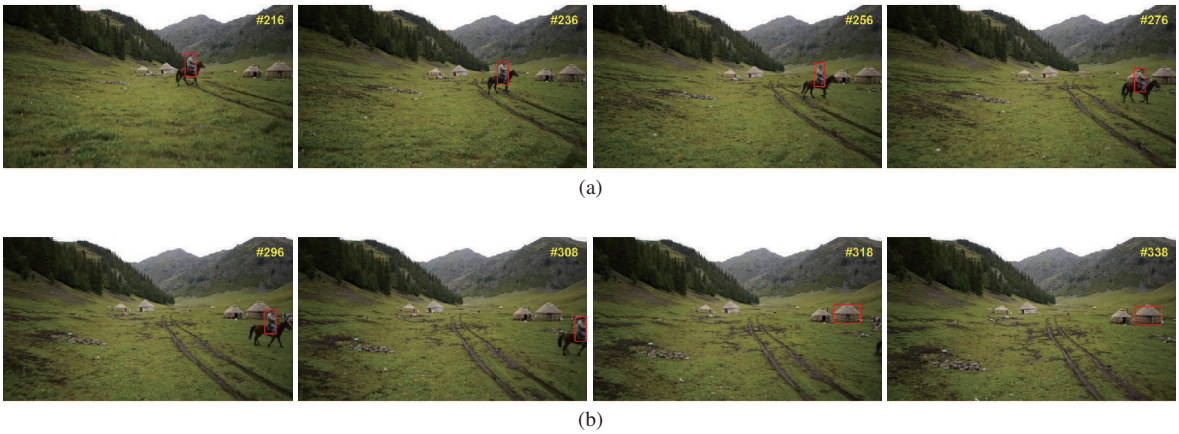

Figure 11 (Color online) Salient object tracking with UAV vision system. The salient object is rebuilt well in frames #318, #338. (a) From left to right: frame #216, #236, #256, #276; (b) from left to right: frame #296, #308, #318, #338.

5.3 Application to an unmanned aerial vehicle (UAV) vision system

A typical application of salient object detection and tracking is the vision system of a UAV used for land-sea search and surveillance operations [25]. Automatic extraction of salient objects is used to augment human object tracking. For example, in patrol applications, the automatic object discovery and tracking module is used to confirm a target. A challenge arises from the salient object rebuild problem when the target changes. We applied our proposed approach to an off-line video representing this scenario, and resolved the global optimization when the salient object sequence changed. Figure 11 shows the detection results when using the salient object discovery and rebuild algorithm with the UAV vision system. It can be seen that the salient object transferred satisfactorily in frame #318 and frame #338.

We next designed an online algorithm for the vision system of the UAV. A selected number of frames were stored in a stack as $\{A_{t-N}, \ldots, A_t\}$, which was updated with each new frame A_t. The salient feature of A_t was calculated and the potential salient object positions were extracted. The SSA was done on $\{A_{t-N}, \ldots, A_t\}$, and the new salient object sequence was identified through spectral clustering on the candidate positions. The candidate salient object positions were limited and the spectral clustering was shown to be highly efficient. We continued the dynamic updated 3D graph, which could be computed efficiently because the previous state variables in $\{A_{t-N}, \ldots, A_{t-1}\}$ had been calculated and stored in the previous steps. Using this method, we were able to update the salient object sequence on the UAV vision system and realize salient object detection and tracking.

While automatic salient object automatic detection and tracking is already used in police patrol and robot vision systems, a large amount of video data are collected in which the salient objects can be labeled by algorithm and adjusted by hand. These labeled data can then be used for training of salient features selection and parameters, as well as for evaluation of the algorithms. We are currently collecting data, and will produce more quantitative evaluation results in a future study.

6 Conclusion

We have presented a novel approach to automatic rebuilding of salient object sequences from video datas. The rebuilding process is performed using an efficient global optimization framework, and the salient object rebuild problem is addressed via sequence segment analysis. Our approach is able to address the salient object rebuild problem in video sequences in which the salient object changes. Several important issues require further investigation. First, multiple salient object sequences still pose challenges to the salient object sequence rebuilding. In future research we will investigate the multiple salient object sequence discovery and tracking problem further. Next, current approaches do not consider occlusions, which may interrupt the salient object sequence. Finally, the application of salient object detection to

real-time vision systems poses challenges in algorithm design and optimization. These issues are left for future work.

Acknowledgements This work was supported by National Key RD Program of China (Grant No. 2016YFB100 1001), National Natural Science Foundation of China (Grant No. 61603022), and China Postdoctoral Science Foundation and Aeronautical Science Foundation of China (Grant No. 20135851042).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Borji A, Itti L. State-of-the-art in visual attention modeling. IEEE Trans Pattern Anal Mach Intel, 2011, 35: 185–207
2 Ma L, Chen L, Zhang X J, et al. A waterborne salient ship detection method on SAR imagery. Sci China Inf Sci, 2015, 58: 089301
3 Liu T, Yuan Z J, Sun J, et al. Learning to detect a salient object. IEEE Trans Pattern Anal Mach Intel, 2011, 33: 353–367
4 Liu T, Zheng N N, Yuan Z J, et al. Video attention: learning to detect a salient object sequence. In: Proceedings of the International Conference on Pattern Recognition, Tampa, 2008. 1–4
5 Feng J, Wei Y, Tao L, et al. Salient object detection by composition. In: Proceedings of the IEEE International Conference on Computer Vision, Barcelona, 2011. 1028–1035
6 Yang C, Zhang L, Lu H, et al. Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013. 3166–3173
7 Santella A, Agrawala M, Decarlo D, et al. Gaze-based interaction for semi-automatic photo cropping. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, 2006. 771–780
8 Chen L Q, Xie X, Fan X, et al. A visual attention mode for adapting images on small displays. Multim Syst, 2003, 9: 353–364
9 Rother C, Bordeaux L, Hamadi Y, et al. Autocollage. In: Proceedings of the International Conference and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH), Boston, 2006. 847–852
10 Jiang H, Wang J, Yuan Z, et al. Automatic salient object segmentation based on context and shape prior. In: Proceedings of the British Machine Vision Conference. Durham: BMVA Press, 2011
11 Jiang H, Wang J, Yuan Z, et al. Salient object detection: a discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013. 2083–2090
12 Zhao Y, Lu S J, Qian H L, et al. Robust mesh deformation with salient features preservation. Sci China Inf Sci, 2016, 59: 052106
13 Wu X M, Du M N, Chen W H, et al. Salient object detection via region contrast and graph regularization. Sci China Inf Sci, 2016, 59: 032104
14 Comaniciu D, Ramesh V, Meer P. Real-time tracking of nonrigid objects using mean shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, 2000. 142–149
15 Wei Y, Sun J, Tang X, et al. Interactive offline tracking for color objects. In: Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, 2007
16 Zhang G, Yuan Z, Zheng N N, et al. Visual saliency based object tracking. In: Proceedings of the Asian Conference on Computer Vision. Berlin: Springer, 2009. 193–203
17 Liu D, Chen T. A topic-motion model for unsupervised video object discovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007
18 Yun X, Jing Z L, Xiao G, et al. A compressive tracking based on time-space Kalman fusion model. Sci China Inf Sci, 2016, 59: 012106
19 Yang Y X, Yang J, Zhang Z X, et al. High-speed visual target tracking with mixed rotation invariant description and skipping searching. Sci China Inf Sci, 2017, 60: 062401
20 Doucet A, Freitas N, Gordon N. Sequential Monte Carlo Methods in Practice. Berlin: Springer, 2001
21 Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001. 282–289
22 Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intel, 2000, 22: 888–905
23 Blake A, Rother C, Brown M, et al. Interactive image segmentation using an adaptive GMMRF model. In: Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2004. 428–441
24 Sun J, Zhang W, Tang X, et al. Bi-directional tracking using trajectory segment analysis. In: Proceedings of the IEEE International Conference on Computer Vision, Beijing, 2005
25 Zhang Y, Su A, Zhu X, et al. Salient object detection approach in UAV video. In: Proceedings of the 8th International Symposium on Multispectral Image Processing and Pattern Recognition. Bellingham: SPIE Proceedings, 2013. 8224