PAPER

Recognition of Moving Object in High Dynamic Scene for Visual Prosthesis

Fei GUO[†], **Yuan YANG**[†a)], **Yang XIAO**[†], **Yong GAO**[†], ***Nonmembers***, ***and*** **Ningmei YU**[†], ***Member***

SUMMARY Currently, visual perceptions generated by visual prosthesis are low resolution with unruly color and restricted grayscale. This severely restricts the ability of prosthetic implant to complete visual tasks in daily scenes. Some studies explore existing image processing techniques to improve the percepts of objects in prosthetic vision. However, most of them extract the moving objects and optimize the visual percepts in general dynamic scenes. The application of visual prosthesis in daily life scenes with high dynamic is greatly limited. Hence, in this study, a novel unsupervised moving object segmentation model is proposed to automatically extract the moving objects in high dynamic scene. In this model, foreground cues with spatiotemporal edge features and background cues with boundary-prior are exploited, the moving object proximity map are generated in dynamic scene according to the manifold ranking function. Moreover, the foreground and background cues are ranked simultaneously, and the moving objects are extracted by the two ranking maps integration. The evaluation experiment indicates that the proposed method can uniformly highlight the moving object and keep good boundaries in high dynamic scene with other methods. Based on this model, two optimization strategies are proposed to improve the perception of moving objects under simulated prosthetic vision. Experimental results demonstrate that the introduction of optimization strategies based on the moving object segmentation model can efficiently segment and enhance moving objects in high dynamic scene, and significantly improve the recognition performance of moving objects for the blind.

key words: *visual prosthesis, moving object segmentation, high dynamic scene, prosthetic vision*

1. Introduction

Potentially dry age-related macular degeneration (AMD) and ritinitis pigmentosa (PR) are the major retinal diseases causing blindness [1], [2]. At present, no effective clinical treatments are put forward to restore the blind vision. By implanting visual prosthesis at the visual pathway to generate the electrical stimulation in the vision nerve has been proven as an effective technique to restore partial vision for the blind [3]. Currently, there are three main types of retinal prosthesis: epiretinal, subretinal and suprachoroidal prosthesis according to the location where the electrodes implanted [4]–[6]. Between them, two most commercially available devices are the Argus II and Alpha IMS, respectively. The Argus II is developed by Second Sight Medical Products, which have received FDA approval in USA in February 2013 and CE marking in Europe in March 2011.While the Alpha-IMS is developed by Retinal Implant AG, which obtained CE marking in Europe in July 2013 [2]. Up to this date, there are more than 200 cases that blind patients implanted the above devices. After implanting the devices, subjects have better performance on the visual tasks [7], [8].

Great progresses have been made in visual prosthesis, the electrodes number have increased from 16 (Argus I) to 60 (Argus II). However, due to the challenges in technique and biology, visual acuity in visual prosthesis is worse than normal vision [9]. The best visual acuity in Argus II and Alpha IMS that recent clinical reported was 20/1262 and 20/546, respectively [10], [11]. This is still lower than the limit of visual acuity (20/200) for the legal blindness [12]. Although significant improvements for the activities of daily life in the implant with such visual acuity by accessing variety of daily visual tasks, it is still difficult to complete more complex visual tasks such as location and recognition in high dynamic scenes. The high density electrodes are needed in future designs. Meanwhile, prosthetic implants reported that the elicited phosphenes are unruly and have limited gray levels. Therefore, we can conclude that the visual perception elicited by the visual prosthesis can cause the poor understandings for the blind.

In visual prosthesis, image processing algorithms are introduced in the external video processing unit (VPU) to optimize the perception of objects in limited prosthetic vision. This is a variable way to improve the understanding of the visual perception for the prosthetic implants. Many studies have developed image processing strategies, and evaluating the performance by performing visual tasks in simulated prosthetic vision. Boyle et al. [13] adopted two traditional processing methods (inverse contrast and edge detection) and two image presentation techniques (distance mapping and importance mapping) to evaluate the subject perceptions under simulated prosthetic vision with different resolutions and gray scales. Van Rheede et al. [14] proposed image presentation strategies (Full-Field Presentation, Region of Interest (ROI) and Fish eye) based on retinal prosthetic vision. Results showed that the region of interest and fish eye methods increased the visual acuity of the prosthetic device user to produce favorable results during the static observation tasks. The Full-Field presentation method performs better in visual tasks that need external environmental information. Zhao et al. [15] studied the minimum information requirement of simulated prosthetic vision aimed at solving the task of object and scene recognition. Lu et al. [16] proposed the projection and nearest neighbor search methods to optimize the presentation of Chinese characters and

Manuscript received November 30, 2018.
Manuscript revised February 27, 2019.
Manuscript publicized April 17, 2019.

[†]The authors are with the Faculty of Automation and Information Engineering, Xi'an University of Technology, China.
a) E-mail: yangyuan@xaut.edu.cn

DOI: 10.1587/transinf.2018EDP7405

paragraphs. Results showed that the two optimized strategies increased the recognition of Chinese characters and the user's ability to read. Jea-Hyun Jung et al. [17] adopted a system of active confocal imaging based on the light-field technology. The system was able to help prosthetic users focus on the objects of interest interesting objects while weakening interference of background clutters. Jing Wang et al. [18] and N. Parikh [19] proposed image processing strategies based on improved itti-saliency detection method. The results demonstrated that the saliency map can provide clues for searching and performing tasks for users with visual prosthesis. Li et al. [20] proposed two image optimization processing strategies based on GBVS -saliency detection model, aims to optimize the presentation in simulated prosthetic vision. Results showed that the introduction of image processing methods can improve the performance of object recognition. Li et al. [21] proposed a real-time image processing strategy, which based on a novel saliency detection algorithm. Their results demonstrated that the effectiveness of adopting the novel saliency detection algorithm to improve the processing efficiency of strategy and the perception of objects in a scene. Guo et al. [22] proposed visual information optimization strategies, which focus on the recognition of the salient object detection in static life scenes. The optimization strategies are based on a two-stage salient object detection model and exploit the gray transform and zooming techniques to optimize the salient object presentation. For the two-stage salient object detection method, the saliency values are computed based on the ranking scores to each side of the background queries in the first stage. In the second stage, the saliency values are refined by the ranking scores to the foreground segmented from the first stage. The results showed the effectiveness of the optimization strategies and the significance for the future application of visual prosthesis.

Introduction of image processing algorithms has been proven to be beneficial for optimizing the visual presentation and improving object recognition performance. However, studies have rarely investigated the recognition of moving objects in high dynamic scenes under simulated prosthetic vision. A time-to-contact map based on depth image is proposed by McCarthy and Barnes [23], which focus on free-moving incoming object perception. Results demonstrated that the effectiveness of the proposed method for emphasizing objects posing an imminent threat of collision. Jing Wang et al. [24] proposed two image processing strategies based on an improved background-subtraction (Vibe) technique [25], aims to segment moving objects from daily scenes and optimize the presentation in simulated prosthetic vision. Results from their research showed that the adopted image-processing strategies increased the recognition and response accuracy in low resolution. But they only investigated the feasibility of perceiving a moving object in a static camera condition. However, the application of visual prosthesis is always in the moving camera and high dynamic scenes. Hence, we proposed an unsupervised moving object segmentation model that is effective and robust in high dynamic scenes, such as illumination changes and mobile camera. Using this model to segment the moving objects can make the processing strategies more suitable for the visual prosthesis application condition.

In this study, the ultimate goal is to improve the perception performance in simulated prosthetic vision. Thus, on the basis of the moving object segmentation method, Edge detection and gray transform are combined to construct two image optimization strategies. Moreover, psychological experiments are performed to evaluate the effectiveness of the optimization strategies in daily scenes. The results demonstrate that the moving object segmentation model has the superior in terms of accuracy and speed over other methods, and the proposed strategies are able to improve the perception in daily life for the recognition of moving objects under simulated prosthetic vision.

2. Material and Method

2.1 Subjects

The subjects participated in the experiment are 16 volunteers chosen from Xi'an University of technology. They (8 males and 8 females) are aged from 20 to 25 years. They are all with normal or corrected visual acuity. The experiment is performed in accordance with the Declaration of Helsinki.

2.2 Material

The materials used in the experiment were video sequences selected from our daily life. The visual field is 20° that simulates the current prosthesis device. The resolution of each frame was normalized to 320*320. In order to avoid the influence of resolution, the visual field of the main object in the image are covered the angle of $12° - 14°$.

2.3 Image Processing Strategies

In the image processing stage, input images are adjusted to the low resolution for simulating the implanted electrode array. When present the daily life scenes, the low resolution will lead to a visual features loss. Segmenting the moving object from the whole life scene and increasing the contrast between the foreground and background can optimize the moving object perception under simulated prosthetic vision. Therefore, a novel unsupervised moving object segmentation model is developed to extract the moving object in high dynamic and moving camera scenes. Furthermore, two image processing strategies are proposed to improve the perception in simulated vision, and then compared with direct lowering resolution (DLG) without any processing. Figure 1 shows the overview of the image processing strategies based on moving object segmentation model. 'SP' and 'FED' processing strategies are introduced to enhance the contrast of foreground and background. In 'SP' processing strategy, the gray-levels of background are linearly decreased to its half, the foreground is remained as the binary segmentation map.

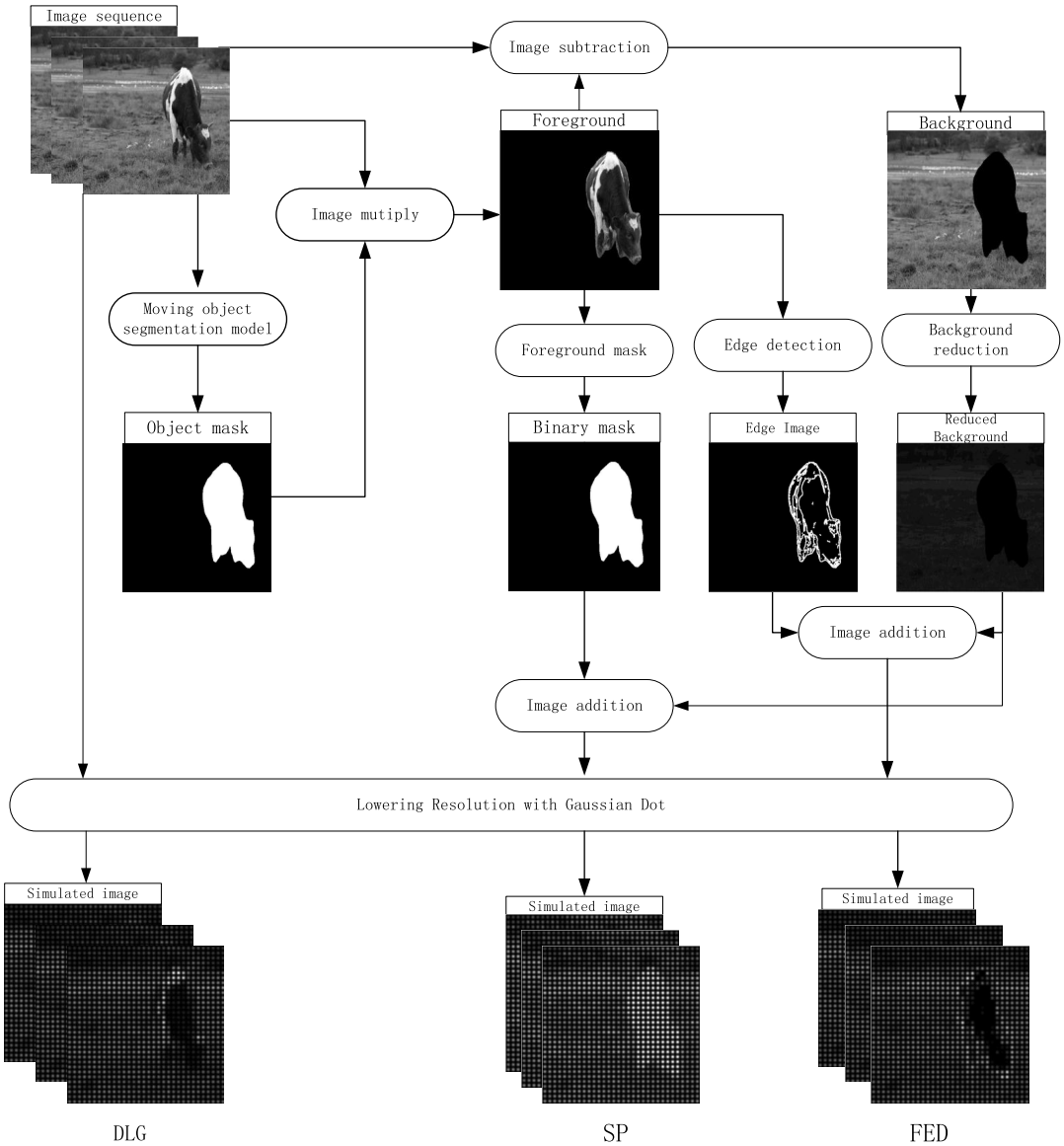

Fig. 1 Flowchart of the proposed image processing strategies based on a moving object segmentation model

For the 'FED' processing strategy, the background is transformed the same as SP strategy, edge detection is used in the foreground to extract the contour information. In the final, the processing images are processed under prosthetic vision with low resolution corresponding to the implanted electrode array.

2.4 Moving Object Segmentation Method

2.4.1 Related Work

In current, the techniques of moving object segmentation are urged widely in many applications. Supervised and unsupervised methods are the two main categories in the moving object segmentation model. For the supervised methods, the manual annotations on given frames are needed to identify the objects, and are implemented always based on deep learning, which have good performance [26], [27]. However, for the unsupervised one, they focus on automatic moving object segmentation without any annotations. Due to the lack of information prior, the unsupervised method is more difficult than the supervised one. Considering the application of visual prosthesis, the unsupervised methods are the focuses in this paper. In the unsupervised method, background subtraction (BS) is a common technique for the moving detection. In this model, the pixels that show changes from one frame to another are considered as foreground and others belonged to background. The detection performance is always relied on the background model. A Gaussian mixture model (GMM) are studied and used as the background model to deal with the dynamic background [28], [29]. Apart from GMM, Kernel density es-

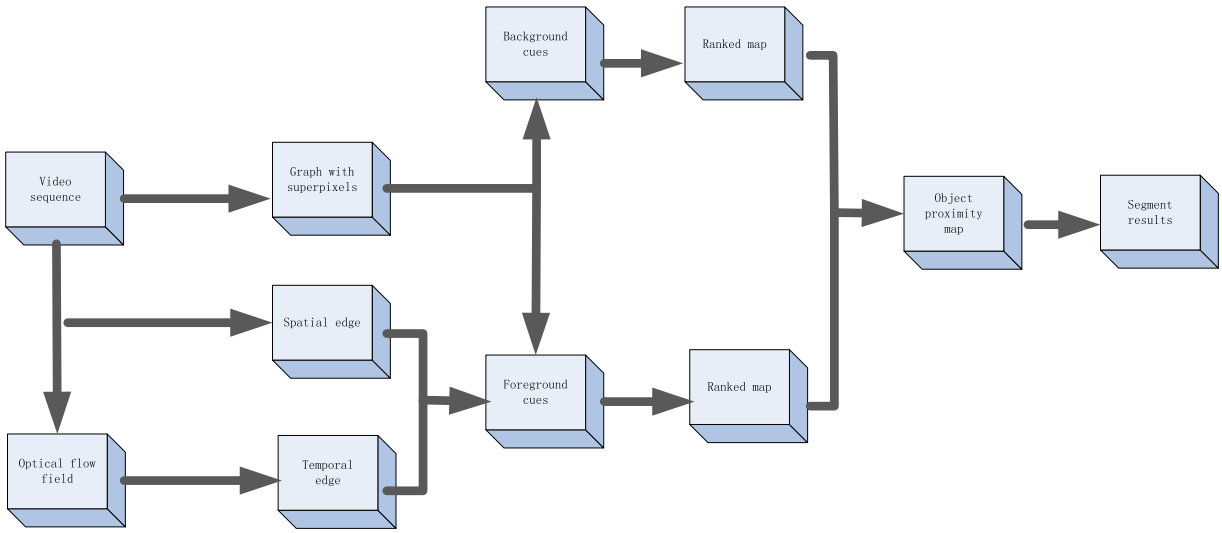

Fig. 2 Framework overview of the proposed moving object segmentation model

timation (KDE) and other non-parametric models are proposed [30]. Also, the codebook model and improvements are proposed to represent the background [31]. But so far, these background modelling methods discussed above failed to deal with the scenes with simultaneous motion of camera and moving objects. Hence, more methods based on optical flow and trajectory arises to address the challenges of moving camera. For optical flow obtained from two adjacent frames, it is widely used in motion estimation, object segmentation and motion segmentation. Kwak et al. [32]utilizes optical flow to initialize the motion field, the Bayesian filters are maintained as background model. To optimize this model, Mark Random Fileds are employed by Zamalieva et al. [33]. In addition to that, Naranyan et al. [35]increase the robustness of this model by employing the orientation of optical flow. Deqing et al. [36]captures the long range correlations in natural scenes by adopting a fully connected layered model. In [37], for the optical flow, its angle and magnitude are combined to maximize the object motion differences. In order to decrease the errors in optical flow, Tokmakov et al. [38] learned a coarse features of optical flow field by utilizing an deep learning network with end to end.

Trajectory is the combination of optical flow in time sequences. Narayanan et al. [39] employed the point trajectories to tracked the motion. Ochs et al. [40]and Brox et al. [41] uses color features to analyse the long term trajectory. To classify the trajectories, Elqursh et al. [42] and Cui et al. [43] proposed a framework with Bayesian filtering and a statistic model. Moreover, Berger et al. [44] described the background by using the linear trajectories subspace. Wu et al. [45] utilizes the motion difference to segment objects by assuming the stronger of objects motion compared with motion of camera. But due to the large computational error in the moving edge, the accuracy of object detection and segmentation is influenced. Some improvements are needed to reduce the interference of moving edges. Xu et al. [46] proposed a variational model for the accurate optical flow estimation. Liu et al. [47]proposed a SIFT-flow method based on the constraint equation of the optical flow, this method improves the moving detection effects. However, due to occlusions and large displacement, the estimated optical flow may contain significant errors. Also, most methods don't consider flow estimation and object segmentation together. Hence, optical flow and trajectory-based method can't accurately label the foreground from the moving camera.

In this study, a novel unsupervised moving object segmentation method is proposed, which aims to automatically extract the moving objects in high dynamic scenes with moving camera for the visual information optimization of visual prosthesis. The whole processing flow is illustrated by Fig. 2. The video sequences are first constructed as a graph with super-pixels. A foreground cues are generated by the integration of spatial edges map and the gradient magnitude of optical flow field. At the same time, the background cues are generated based on the boundary prior. For each super-pixel, we applied manifold ranking model to rank the foreground cues and background cues, simultaneously. The moving object maps are measured by integration of the ranking scores of foreground and background cues. Finally, the moving object segmentation maps are produced and refined by the grab-cut. In this method, a novel framework is proposed to detect the moving objects in high dynamic scene. Unlike other works, the proposed framework focuses on integrating foreground and background cues simultaneously rather than improving the optical flow accuracy. The simultaneously integration of the two cues can compensate the respective drawbacks. Meanwhile, the spatiotemporal edges are generated as a foreground cues instead of the optical flow field. This scheme can eliminate the foreground detection map errors caused by the occlusions and large displacements in the optical flow estimation. The background cues based on the boundary prior are exploited to refine the foreground ranking maps.

2.4.2 The Proposed Method

The proposed moving object segmentation method exploits the manifold ranking model with foreground cues and background cues to achieve the reliable moving object segmentation in moving camera scenes. We detailed describe the proposed method by the following sections:

Cues Extraction of Foreground and Background: motion features captured by optical flow between two consecutive frames are crucial for the moving object detection, but errors are still existed in optical flow estimation. In order to decrease the errors in detecting the target, similar to [48], [49], we explored a spatiotemporal foreground features to locate the moving object, which integrate the spatial edges and motion edge to compensate the errors of single motion features. For a video frame $F(x)$, a spatial edge map $E_s(x)$ of the frame is computed first using 'sobel' operator. The optical flow field V of the pairs of the consecutive frame is estimated by [47]. Then, the temporal edge $E_t(x)$ is generated by Eq. (1).

$$E_t(x) = \|\nabla V(x)\| \tag{1}$$

where $\|\nabla V(x)\|$ is the gradient magnitude of optical flow field $V(x)$.

Then, we combine the spatial and temporal edges as the spatiotemporal edge features using Eq. (2).

$$E_k(x) = E_s(x) * E_t(x) \tag{2}$$

For the spatiotemporal edge map, distinct motion patterns and spatial gradient indicate the location of moving object. The high values of pixels within the spatiotemporal edge map are as the foreground cues to the subsequent processing.

To further decrease the influences of the errors in optical flow estimation, in this paper, the background cues are also be generated to estimate the moving object. For the background cues, we defined the pixels in the frame boundary to be the background cues based on the boundary prior. The boundary priors inspired by [50] indicated that humans tend to gaze at the image centre. This theory is used widely in saliency detection, image segmentation and related researches [51]–[53]. In this paper, the pixels in the boundaries of a given frame are as the background cues to be ranked to estimate the moving object.

Manifold Ranking: in the proposed moving object segmentation method, we model the segmentation as a manifold ranking problem with background and foreground cues, simultaneously. Manifold ranking labels the graph by employing the intrinsic manifold structure of data (e.g. images). Given a node as query, the other nodes are ranked with the relevance to the given query. In the manifold ranking, the relevance between the given queries and the other nodes is defined by a ranking function needed to be learned. For example, in a dataset $X = \{x_1, \ldots, x_i, \ldots, x_n\} \in R^{m\times n}$, some data are set as queries and the others need to be ranked based on relevance with the queries.We define a ranking function $f : x \to R^n$, it can be treat as a vector $f = [f_1, \ldots, f_n]^T$ that assigns a ranking score f_i to each node X_i. Meanwhile, we define a indicator vector$y = [y_1, \ldots, y_n]$, in which $y_i = 1$ if the X_i is query node, else $y_i = 0$. Next, a graph model $G(V, E)$ are constructed, in which V are the dataset X and E are the edges. An affinity matrix $W = \left[W_{ij}\right]_{m\times n}$ is defined to weight the edges E. Based on G, the degree matrix $D = diag\{d_{11}, \ldots, d_{nn}\}$ is obtained, where $d = \sum_j w_{ij}$. Thus, the optimal ranking function f_{ran} of queries are captured by solving the following optimization [52]:

$$f_{ran} = argmin\frac{1}{2}\Big(\sum_{ij}^{n} w_{ij}\left\|\frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}}\right\|^2 + \mu\sum_{i=1}^{n}\|f_i - y_i\|^2\Big) \tag{3}$$

Where μ controls the balance between the smoothness constraint (the first term) and the fitting constraint (the second term).i and j indexes the super-pixels on on the graph. By setting the derivative to be zero, the ranking function can be written as:

$$f_{ran} = (D - \alpha W)^{-1} y \tag{4}$$

Where $\alpha = (1 + \mu)^{-1}$, W is the unnormalized Laplacian matrix.

Graph Construction:Given the input video frame, superpixel segmentation is applied first by using SLIC [54]. Then, a graph model $G(V, E)$ is constructed, in which nodes V are the set of super-pixels, E are the links of the adjacent superpixels. Based on the manifold ranking theory mentioned above, the weight w between two nodes is described as

$$w_{ij} = e^{-\frac{\|c_i - c_j\|}{\sigma^2}} \tag{5}$$

Where i and j are the indexes of super-pixel nodes, c_i and c_j are the mean value of two super-pixels, respectively. σ is the constant to control the weight strength. Moreover, the affinity matrix W and degree matrix D are obtained.

Ranking with Foreground Cues:For the foreground cues, we utilizes the nodes in the spatiotemporal edges proximity map as the queries, the other nodes are as the unlabelled data. The query nodes are the super-pixels in the spatiotemporal proximity map with high value, which are defined as

$$y(i) = \begin{cases} 0 & \text{if } E_k(i) < T \\ 1 & \text{if } E_k(i) \geq T \end{cases} \tag{6}$$

where $E_k(i)$ is value of i-th nodes in the spatiotemporal probability map. T is the adaptive threshold generated by otsu [55]. If the i-th node is query, the indicator vector $y(i) = 1$, else $y(i) = 0$. Hence, the indictor vector y is given. The object detection probability map P_{fo} based on

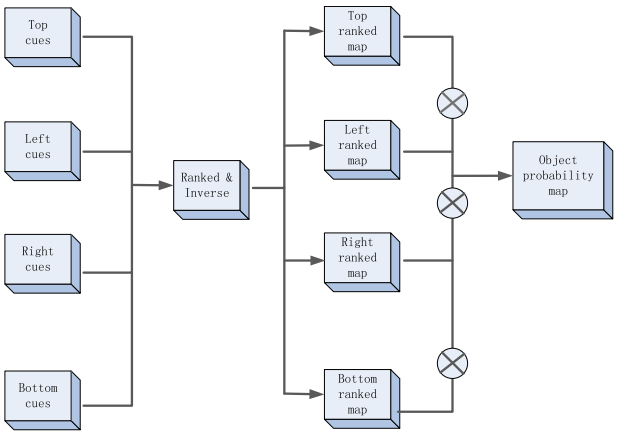

Fig. 3 Scheme of the ranking with background cues (top, bottom, left and right side boundaries)

foreground cues is ranked by ranking functions, which is described as

$$P_{fo} = \bar{f}_{ran}(i) \tag{7}$$

where $\bar{f}_{ran}$ is the normalized version of f_{ran}, and i indexes the nodes on the graph.

Ranking with Background Cues: For the background cues, given the dissimilar of different side super-pixels in the image boundary, the nodes in the boundary are divided into four sides: the bottom, the top, the right and the left part to be ranked separately instead of the all the boundary nodes, simultaneously. The scheme is illustrated in Fig. 3. This can decrease the effects of imprecise queries and improve the ranking effect. We selected the right boundary prior in details to describe the ranking scheme.

For the right boundary, the right-side nodes are utilized as queries, the other nodes are as the unlabelled data. Hence, the indictor vector y is obtained, all the other nodes are ranked according to the ranking function. We normalized the ranking value, and the probability map based on top boundary is written as:

$$P_r = 1 - \bar{f}_{ran}(i) \tag{8}$$

Where i indexes the super-pixel node on the graph, $\bar{f}_{ran}$ is the normalized version of f_{ran}.

Similarly, we compute the other three maps P_t, P_l and P_b by using top, left and bottom boundaries as queries, respectively. The object detection probability map P_{ba} is obtained by integrating the four probability maps by the following process:

$$P_{ba} = P_t \times P_b \times P_l \times P_r \tag{9}$$

The final moving object detection map P is generated by combining the background and foreground cues, which are written as:

$$P = P_{ba} \times P_{fo} \tag{10}$$

Obtained the final moving object map, we employed grab-cut method [56] to segment the detection result. We segment the moving probability map as a binary mask by setting an adaptive threshold, the binary mask is as the accurate region to the grab-cut. We can effectively segment the accurate moving objects. Due to the accurate location of the moving object instead of the manual rectangle region, the grab-cut method can get the relatively good results with efficient.

2.5 Prosthetic Vision Model

For the simulated prosthetic vision, a Gaussian distribution-based phosphene model is introduced [57]. The video images must be down-sampled to 24*24 and 32*32 resolution to match the number of electrodes in the visual prosthesis. In details, the images are divided into regions with fixed size, and the pixels in the regions are combined. The mean gray value of the pixels in the region is used as the central luminance value of the Gaussian points. The luminance distribution of simulated prosthetic vision is as the Gaussian curve. This model is described as:

$$I(x, y) = A(u_x, u_x) \bullet G(x, y) \tag{11}$$

where$A(u_x, u_x)$is the gray value of the stimulated pixels and $G(x, y)$ represents the Gaussian distribution function, which is shown as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-u_x)^2+(x-u_y)^2}{2\sigma^2}} \tag{12}$$

The images using the prosthetic vision model correspond to the electrode arrays. This process is called 'Lowering Resolution with Gaussian dots,(LRG)'.

2.6 Experimental Setup

Subjects are seated 60 cm in front of a 21 inches LCD monitor (Lenovo INC, BeiJing, 1280*1024 resolutions, 26° visual fields). The videos display on the centre of the monitor, randomly. The experimental process is controlled by the psychological toolbox software 'PsychToolbox-3'. Before the start of the experiment, the subjects are provided with a list of the experimental objects. This helps the subjects familiarize themselves with the upcoming objects and the experimental protocol. During the experiment, the videos are divided into three groups ('DLG', 'SP' and 'FED') and randomly presented to the participants.

2.7 Data Analysis

The recognition score (RS) are used to quantify the recognition results. If the subjects are able to correctly recognize the objects and give the right name, RS is set to 2. If the subjects can not correctly name the object, they can describe the shape or specific features of objects, RS is set to 1. Otherwise, the RS is set to 0. The values of RS are normalized to the recognition accuracy (RA) under different processing strategies, as shown in Eq. (13). The software of Statistical

Fig. 4 Qualitative comparison of our method with SAVOS [49], calMoSeg [58] and SCUBU [59] models on the representative subset of the FBMS dataset

Product and Service Solutions (SPSS) for Windows (SPSS Inc.) is adopted to perform statistical analysis. The results are expressed in the form of $mean \pm SEM$ (standard error of mean). A two-factor analysis of variance (ANOVA) is adopted as the metric to evaluate the effect of statistical significance of the resolution and processing strategies.

$$RA = \frac{RS}{2} \times 100\% \tag{13}$$

3. Results and Discussion

3.1 Results of Moving Object Segmentation

In order to illustrate the advantages of the moving object segmentation model, our method is compared with three state-of-art methods, which are SAVOS [49], calMoSeg [58] and SCUBU [59] based on the Freiburg-Berkeley Motion Segmentation dataset (FBMS) [40]. The video sequences provided by the FBMS dataset (including 59 videoes) contains complex scenes with the camera motion such as translation, rotation and scaling transformations, which lead to a more challenge in moving object segmentation. The comparison of segmentation effects using the FBMS dataset is comprehensive. Before the evaluation, all the optical flows of different methods are estimated in davance by [47]. First, the qualitative comparisons are performed with other state-of art methods. Given the length of the paper, Fig. 4 shows the binary masks of the moving object segmented by different models, which including 12 representive videos selected from the FBMS-testing dataset. From the analysis of the images, we can conclude that the results processed by our method have advantage over others. It ensures the clarity and completeness of the object contour. Moreover, it is robust to the changes in scenes and completes the segmentation task well in multi-objects and moving camera scenes.

Furthermore, the quantitative comparisons are executed using the metric of $F-measure$. In the quantitative evaluation, *Precision* reflects the proportion of correct salient pixels with the salient pixels, while *recall* corresponds to the fraction of correctly assigned salient pixels with the ground truth. However, *Precision* and *recall* ignored the true negative assignments, so we utilized the $F-measure$ metric that is the combination of precision and recall, which is defined as:

$$F-measure = \frac{(1+\beta^2) \times Precision \times recall}{\beta^2 \times Precision + recall} \tag{14}$$

In the perception evaluation, the *Precision* is more important than *recall*. As suggested by many moving object detection works [49], [58], [59], β^2 is often set to 0.3 to raise the importance to the *Precision* value.

Table 1 Quantitative comparison of our method with three state-of-art methods on the representative subset of FBMS dataset using $F-measure$ metric, and the average $F-measure$ on FBMS dataset. The best results are boldfaced

Videos	ours	SAVOS [49]	calMoSeg [58]	SCUBU [59]
Camel01	**0.823**	0.812	0.226	0.173
Cars1	**0.856**	0.766	0.542	0.324
Cars4	**0.944**	0.786	0.874	0.532
Cats01	0.743	**0.847**	0.706	0.466
Dogs01	**0.914**	0.821	0.546	0.265
Farm01	**0.846**	0.816	0.723	0.473
Goats01	0.893	**0.924**	0.864	0.305
Horses04	**0.836**	0.634	0.346	0.105
People1	**0.927**	0.656	0.842	0.776
People2	**0.893**	0.825	0.802	0.763
Tennis	**0.885**	0.871	0.773	0.726
Lion01	0.876	0.805	**0.905**	0.793
Average	**0.673**	0.616	0.524	0.364

In Table 1, it shows the comparison of $F-measure$ scores on the representative subset of FBMS dataset and the average $F-measure$ scores on the whole FBMS dataset. The proposed method obtained a higher average score with other models, which increased 9% with SAVOS, 28% with calMoSeg and 85% with SCUBU, respectively. It demonstrated that the proposed method have more robustness in different moving camera scenes.

In Table 2, the average time taken by each method are

Table 2 Average time taken in FBMS dataset. The best and second results are boldfaced and unerlined, accordingly

Mothod	ours	SAVOS [49]	calMoSeg [58]	SCUBU [59]
Time(s)	<u>0.106</u>	0.178	0.625	**0.005**
Code Type	Matlab	Matlab	Matlab	C++

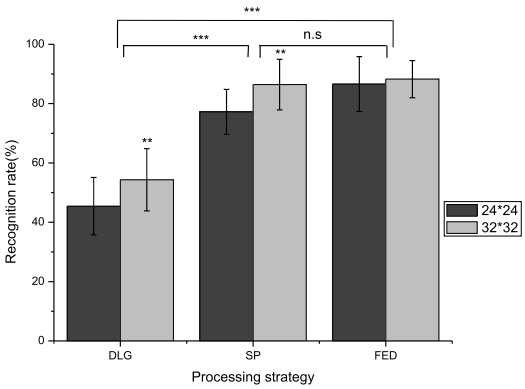

Fig. 5 Recognition accuracy of three processing strategies at two resolutions (* p<0.5, **p<0.01, *** P<0.001, n=16)

evaluated on an Intel Core I7 machine with 16GB RAM. The test video sequences are the resolution of 640*480. It showed that the proposed method have taken the sub-lowest average time. In our method, the learnt optimal ranking affinity matrix is computed only once, the only change is the indictor vector, which is a binary vector. The main time consumptions are included in the super-pixel segmentation and optical flow estimation. These results indicated that our method can detect and segment moving objects in moving camera scenes with superior efficiency.

3.2 Results of Moving Object Recognition

The recognition rate results of moving objects outdoor at two resolutions, 24*24 and 32*32 using the three different image processing strategy, DLG, SP and FED are shown in Fig. 5. The RA score using DLG strategy at resolution of 24*24 is 44.06 ± 10.04%, which is the lowest RA score. However, the RA scores significantly ($p < 0.001$) improve to 75.68 ± 8.17% and 85.31 ± 8.65% under SP and FED strategies, respectively. The RA scores under DLG strategy at 32*32 resolution is 55.94 ± 12.92%, while, under SP and FED strategies, the average RA scores significantly improve to 85.34 ± 8.65% and 86.56 ± 7.69% ($p < 0.001$), respectively. The highest RA score is obtained in FED at the resolution of 32*32. The statistical analysis indicates that the processing strategies have significant effects ($F_{strategy} = 119.277$, $p < 0.001$) on the RA score. Meanwhile, resolution has significant impact on RA scores ($F_{strategy} = 54.004$, $p < 0.01$). For the two processing strategies, there is no significant interaction between them.

Compared with SP and FED, the performance of object recognition is the worst in DLG. Without any optimization processing, it is hard for the prosthetic implants to recognize objects in dynamic scene, especially in mobile camera and insufficient contrast of luminance. According to the behavioural studies, larger, brighter and fast moving objects are biased by human attention [61]. Therefore, the moving segmentation model is introduced to extract the moving object from the whole scenes. Gray levels and edge information have an influence on the recognition of object and face when implanting visual prosthesis [62]–[64]. Based on the segmentation results, we attempt to reduce the gray levels to weaken the background and remain the binary mask and edge information to enhance the foreground. Thus, they can effectively increase the contrast between the foreground and background. The results demonstrated that SP and FED were significantly advantageous to DLG in guiding subjects to perceive moving objects.

3.3 Limitations

In the proposed method, occlusions and large displacement may occur in optical flow estimation when the moving objects are in high dynamic scene [47]. Although the spatiotemporal edges are proposed as the foreground cues instead of the optical flow field to eliminate the errors in optical flow estimation, the occlusions and large displacements can still cause the inaccuracy results in some extent. Also, the current optical flow estimation models have high computational cost, which can decrease the efficiency of moving object detection. Furthermore, in order to eliminate the detection errors of foreground ranking map, the background cues are exploited. The background cues rely on the boundary prior theory, which assumed that the pixels in the image boundaries tend to be the background pixels and the detection objects are always in the image center. However, the assumption sometimes may fail, especially when the moving objects touch the image border, which will lead to the accuracy results.

4. Conclusion

In this paper, an unsupervised moving object segmentation method is proposed that exploits the manifold ranking model fused background and foreground cues, simutaneously. Based on the moving segmentation results, two image optimization strategies are proposed to improve the perception of moving objects in simulated prosthetic vision. The experimental results demonstrated that the moving object segmentation method outperforms the existing methods. Furthermore, psychological experiments indicated that 'SP' and 'FED' strategies are benificial to optimize the perceptions of moving objects. It is hoped that the proposed moving object segmentation-based image processing strategies may make a great contribution to the further development of visual prosthesis, which assists the implants to obtain independent mobility in real-life.

Acknowledgments

We are thankful to the volunteers from Xi'an Univer-

sity of technology. This work is supported in part by projects of the Natural Science Foundation of China (Grant no.51477138), the Key Research and Development Program of Shaanxi Province (Grant no.2017ZDXM-GY-130), Xi'an City Science and Technology Project (Grant no.2017080CG/RC043XALG009), and Industrial research project of Science and Technology Department of Shaanxi Province (Grant no.2017GY-083).

References

[1] A.C. Weitz and J.D. Weiland, "Visual Prostheses," Neural Computation, Neural Devices, and Neural Prosthesis, Springer New York, pp.157–188, 2014.
[2] Y.H.-L. Luo and L. da Cruz, "A review and update on the current status of retinal prostheses (bionic eye)," Brit. Med. Bull., vol.109, no.1, pp.31–44, 2014.
[3] L. Yue, J.D. Weiland, B. Roska, and M.S. Humayun, "Retinal stimulation strategies to restore vision: Fundamentals and systems," Progress in Retinal & Eye Research, vol.53, pp.21–47, 2016.
[4] E. Zrenner, "Will retinal implants restore vision?," Science, vol.295, no.5557, pp.1022–1025, 2002.
[5] R.A. Fernandes, B. Diniz, R. Ribeiro, and M. Humayun, "Artificial vision through neuronal stimulation," Neuroscience Letters, vol.519, no.2, pp.122–128, 2012.
[6] R.K. Shepherd, M.N. Shivdasani, D.A.X. Nayagam, C.E. Williams, and P.J. Blamey, "Visual prostheses for the blind," Trends in Biotechnology, vol.31, no.10, pp.562–71, 2013.
[7] V.C. Coffey, "Vision Accomplished: The Bionic Eye," Optics & Photonics News, vol.28, no.4, pp.24–31, 2017.
[8] E. Zrenner, K.U. Bartz-Schmidt, D. Besch, F. Gekeler, A. Koitschev, and H.G. Sachs, "The Subretinal Implant ALPHA: Implantation and Functional Results," Artificial Vision, Springer International Publishing, pp.65–83, 2017.
[9] C.D. Eiber, N.H. Lovell, and G.J. Suaning, "Attaining higher resolution visual prosthetics: a review of the factors and limitations," Journal of Neural Engineering, vol.10, no.1, pp.011002, 2013.
[10] M.S. Humayun, J.D. Dorn, L. da Cruz, G. Dagnelie, J.-A. Sahel, P.E. Stanga, A.V. Cideciyan, J.L. Duncan, D. Eliott, E. Filley, A.C. Ho, A. Santos, A.B. Safran, A. Arditi, L.V.D. Priore, and R.J. Greenberg, "Interim Results from the International Trial of Second Sight's Visual Prosthesis," Ophthalmology, vol.119, no.4, pp.779–788, 2012.
[11] K. Stingl, K.U. Bartz-Schmidt, D. Besch, C.K. Chee, C.L. Cottriall, F. Gekeler, M. Groppe, T.L. Jackson, R.E. MacLaren, A. Koitschev, A. Kusnyerik, J. Neffendorf, J. Nemeth, M.A.N. Naeem, T. Peters, J.D. Ramsden, H. Sachs, A. Simpson, M.S. Singh, B. Wilhelm, D. Wong, and E. Zrenner, "Subretinal visual Implant Alpha IMS-Clinical trial interim report," Vision Research, vol.11, Part B, pp.149–160, 2015.
[12] L. Dacruz, F. Merlini, and M. Arsiero, "Subjects blinded by outer retinal dystrophies are able to recognize outlined shapes using the Argus(R) II retinal prosthesis system: A comparison with the full shapes recognition task," Investigative Ophthalmology & Visual Science, vol.53, no.14, 5507, 2012.
[13] J.R. Boyle, A.J. Maeder, and W.W. Boles, "Region-of-interest processing for electronic visual prostheses," Journal of Electronic Imaging, vol.17, no.1, pp.142–154, 2008.
[14] J.J. van Rheede, C. Kennard, and S.L. Hicks, "Simulating prosthetic vision: Optimizing the information content of a limited visual display," Journal of Vision, vol.10, no.14, pp.71–76, 2010.
[15] Y. Zhao, Y. Lu, Y. Tian, L. Li, Q. Ren, and X. Chai, "Image processing based recognition of images with a limited number of pixels using simulated prosthetic vision," Information Sciences, vol.180, no.16, pp.2915–2924, 2010.
[16] Y. Lu, H. Kan, J. Liu, J. Wang, C. Tao, Y. Chen, Q. Ren, J. Hu, and X. Chai, "Optimizing Chinese character displays improves recognition and reading performance of simulated irregular phosphene maps," Investigative Ophthalmology & Visual Science, vol.54, no.4, pp.2918–2926, 2013.
[17] J.-H. Jung, D. Aloni, Y. Yitzhaky, and E. Peli, "Active confocal imaging for visual prostheses," Vision Research, vol.111, pp.182–196, 2015.
[18] J. Wang, H. Li, W. Fu, Y. Chen, L. Li, Q. Lyu, T. Han, and X. Chai, "Image Processing Strategies Based on a Visual Saliency Model for Object Recognition under Simulated Prosthetic Vision," Artificial Organs, vol.40, no.1, pp.94–100, 2016.
[19] N. Parikh, L. Itti, and J. Weiland, "Saliency-based image processing for retinal prostheses," Journal of Neural Engineering, vol.7, no.1, pp.835–841, 2010.
[20] H. Li, X. Su, J. Wang, H. Kan, T. Han, Y. Zeng, and X. Chai, "Image processing strategies based on saliency segmentation for object recognition under simulated prosthetic vision," Artificial Intelligence in Medicine, vol.84, pp.64–78, 2017.
[21] H. Li, T. Han, J. Wang, Z. Lu, X. Cao, Y. Chen, L. Li, C. Zhou, and X. Chai, "A real-time image optimization strategy based on global saliency detection for artificial retinal prostheses," Information Sciences, vol.415, pp.1–18, 2017.
[22] F. Guo, Y. Yang, and Y. Gao, "Optimization of Visual Information Presentation for Visual Prosthesis," Intl. J. Biomedical Imaging, vol.2018, pp.12, 2018.
[23] C. McCarthy and N. Barnes, "Time-to-contact maps for navigation with a low resolution visual prosthesis," Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), pp.2780–2783, 2012.
[24] J. Wang, Y. Lu, L. Gu, C. Zhou, and X. Chai, "Moving object recognition under simulated prosthetic vision using background-subtraction-based image processing strategies," Information Sciences, vol.277, no.2, pp.512–524, 2014.
[25] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," IEEE Transaction on Image Processing, vol.20, no.6, pp.1709–1724, 2011.
[26] J.C. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," IEEE International Conference on Computer Vision (ICCV), pp.686–695, 2017.
[27] C. Sergi, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L.V. Gool, "One-Shot Video Object Segmentation," CVPR, pp.5320–5329, 2017.
[28] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," Pattern Recognition Letters, vol.27, no.7, pp.773–780, 2006
[29] H.-H. Lin, J.-H. Chuang, and T.-L. Liu, "Regularized background adaptation: a novel learning rate control scheme for Gaussian mixture modeling," IEEE Transactions on Image Processing, vol.20, no.3, pp.822–836, 2011.
[30] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," CVPR, pp.302–309, 2004.
[31] T. Badal, N. Nain, M. Ahmed, and V. Sharma, "An adaptive codebook model for change detection with dynamic background," 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp.110–116, 2015.
[32] S. Kwak, T. Lim, W. Nam, B. Han, and J.H. Han, "Generalized background subtraction based on hybrid inference by belief propagation and bayesian filtering," IEEE Int. Conf. Comput. Vis. (ICCV), pp.2174–2181, Nov. 2011.
[33] D. Zamalieva, A. Yilmaz, and J.W. Davis, "A multi-transformational model for background subtraction with moving cameras," Eur. Conf. Comput. Vis. (ECCV), Springer, vol.8689, pp.803–817, 2014.
[34] Y. Zhu and A. Elgammal, "A multilayer-based framework for online background subtraction with freely moving cameras," The IEEE Int. Conf. Comput. Vis. (ICCV), pp.5142–5151, Oct. 2017.

[35] M. Narayana, A. Hanson, and E. Learned-Miller, "Coherent motion segmentation in moving camera videos using optical flow orientations," IEEE Int. Conf. Comput. Vis. (ICCV), pp.1577–1584, Dec. 2013.
[36] D. Sun, J. Wulff, E. Sudderth, H. Pfister, and M. Black, "A fully-connected layered model of foreground and background flow," IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR), pp.2451–2458, June 2013.
[37] P. Bideau and E. Learned-Miller, "It's moving! a probabilistic model for causal motion segmentation in moving camera videos," Eur. Conf. Comput. Vis. (ECCV), vol.9912, pp.433–449, 2016.
[38] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.531–539, July 2017.
[39] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," Eur. Conf. Comput. Vis. (ECCV), Springer, vol.6311, pp.438–451, 2010.
[40] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," IEEE Trans. Pattern Anal. and Mach. Intell., vol.36, no.6, pp.1187–1200, June 2014.
[41] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," Eur. Conf. Comput. Vis. (ECCV), vol.6315, pp.282–295, 2010.
[42] A. Elqursh and A. Elgammal, "Online moving camera background subtraction," Eur. Conf. Comput. Vis. (ECCV), Springer, vol.7577, pp.228–241, 2012.
[43] X. Cui, J. Huang, S. Zhang, and D.N. Metaxas, "Background subtraction using low rank and group sparsity constraints," Eur. Conf. Comput. Vis. (ECCV), vol.7572, pp.612–625, 2012.
[44] M. Berger and L. Seversky, "Subspace tracking under dynamic dimensionality for online background subtraction," IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR), pp.1274–1281, June 2014.
[45] Y. Wu, X. He, and T.Q. Nguyen, "Moving object detection with a freely moving camera via background motion subtraction," IEEE Trans. Circuits Syst. Video Technol., vol.27, no.2, pp.236–248, Feb. 2017.
[46] L. Xu, J. Chen, and J. Jia, "A segmentation based variational model for accurate optical flow estimation," European Conference on Computer Vision, Springer, Berlin, Heidelberg, vol.5302, pp.671–684, 2008.
[47] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.33, no.5, pp.978–994, 2011.
[48] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," Computer Vision & Pattern Recognition IEEE, pp.3395–3402 2015.
[49] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware Video Object Segmentation," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.40, no.1, pp.20–33, 2018.
[50] B.W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, Journal of Vision," Journal of vision, vol.7, no.14, p.4, 2007.
[51] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," ICCV, pp.277–284, Sept. 2009.
[52] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," ECCV, Springer, Berlin, Heidelberg, vol.7574, pp.29–42, Oct. 2012.
[53] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency Detection via Graph-Based Manifold Ranking," IEEE Conference on Computer Vision and Pattern Recognition, pp.3166–3173, 2013.
[54] R. Achanta, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels," Technical report, no.149300, EPFL, June 2010.
[55] N. Ostu, "A threshold selection method from gray-histogram," IEEE Trans. Syst. Man. & Cybern, vol.9, no.1, pp.62–66, Doi: 10.1109/TSMC.1979.4310076, 1979.
[56] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: interactive foreground extraction using iterated graph cuts," ACM transactions on graphics, vol.23, no.3, pp.309–314, 2004.
[57] J.S. Hayes, V.T. Yin, D. Piyathaisere, J.D. Weiland, M.S. Humayun, and G. Dagnelie, "Visually guided performance of simple tasks using simulated prosthetic vision," Artificial Organs, vol.27, no.11, pp.1016–1028, 2003.
[58] P. Bideau and E. Learned-Miller, "It's moving! A probabilistic model for causal motion segmentation in moving camera videos," European Conference on Computer Vision, Springer, Cham, vol.9912, pp.433–449, Oct. 2016.
[59] K. Yun, J. Lim, and J.Y. Choi, "Scene conditional background update for moving object detection in a moving camera," Pattern Recognition Letters vol.88, pp.57–63, 2017.
[60] P. Ochs, J. Malik, and T. Brox, "Segmentation of Moving Objects by Long Term Video Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.36, no.6, pp.1187–1200, 2014.
[61] A. Treisman and S. Gormican, "Feature analysis in early vision: evidence from search asymmetries," Psychological review, vol.95, no.1, pp.15–48, 1988.
[62] J. Boyle, W. Boles, and A. Maeder, "Challenges in digital imaging for artificial human vision," Proceeding of SPIE, Human Vision and Electronic Imaging VI, vol.4299, pp.533–544, Doi: 10.1117/12.429525, 2001.
[63] R.W. Thompson, G.D. Barnett, M.S. Humayun, and G. Dagnelie, "Facial recognition using simulated prosthetic pixelized vision," Investigative ophthalmology & visual science, vol.44, no.11, pp.5035–5042, 2003.
[64] J. Boyle, A. Maeder, and W. Boles, "Inherent visual information for low quality image presentation," Proceeding of the Aprs Workshop on Digital Image Computing, pp.51–56, 2003.

Fei Guo received his B.S. degree and M.S. degree in electronics engineering from Xi'an University of Technology, Xi'an, China, in 2011 and 2014, respectively. Now he is a Ph.D. candidate in Xi'an University of Technology. His main research interests include image processing, computer vision and VLSI Design.

Yuan Yang received the B.S. degree in power electronics from Xi'an University of Technology, Xi'an, China, in 1997, and the M.S. and Ph.D. degrees in electronics engineering from Xi'an University of Technology, Xi'an, China, in 2000 and 2004, respectively. Since 2000, she has been a member of the faculty of School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China, where she is currently a professor. From March to August 2004, she was in the center of VLSI, Kyushu University, Fukuka, Japan, as a visiting scholar. She then came back to Xi'an University of Technology, and engaged in teaching and researches in electronics engineering. She is the author of three books, more than 80 articles, and more than 10 inventions. Her main research interests include VLSI design, circuit and system design. Prof. Yang's awards and honors include the first prize in science and technology of Shaanxi Province, the first prize in science and technology of Shaanxi colleges and Universities, the first prize in science and technology of Xi'an, excellent doctoral dissertations in Shaanxi.

Yang Xiao received the B.S. degree in electronics engineering from Xi'an University of Technology, Xi'an, China, in 2016. Now he is a M.S. candidate in Xi'an University of Technology. His main research interests include deep learning and VLSI Design.

Yong Gao received his B.S. degree in applied physics from Xi'an University of Technology in 1982. He received the M.S. and Ph.D. degrees in Microelectronics and solid-state electronics from Xi'an Jiaotong University, Xi'an, China, in 1988 and 1995,respectively. From 1996 to 2009, he has been a Professor with the Electronics Department, Xi'an University of Technology, China. Since 2010, he has been a professor in the department of electronics in Xi'an university of technology and the department of electrical engineering at Xi'an Polytechnic University in China. His research mainly include power electronic devices, computer vision and very large scale integration design.

Ningmei Yu received the B.S. degree in electronics engineering from Xi'an University of Technology, Xi'an, China, in 1986. She received the M.S. and Ph.D. degrees from TOHOKU University, Sendai, Japan, in 1996 and 1999, respectively. Since 2001, she has been a member of the faculty of School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China, where she is currently a professor. Her main reaserch mainly include wireless communication, image compression chip and digital-analog hybrid integrated circuit technology.