Visual Detection and Tracking Based on Fusion Saliency

Xinxing Guo, Baolong Guo* and Cheng Li and Jie Han

Institute of Intelligent Control and Image Engineering, Xidian University, School of Aerospace science and technology, No. 2 South Taibai Road, Xi’an, Shaanxi, China.
E-mail addresses: blguo@xidian.edu.cn.

Abstract. In order to improve the ability of target tracking effectively, an improved MeanShift algorithm is proposed in this paper. The algorithm combines the saliency map from the computer and from the human vision through Kinect, so that it can obtain the fusion saliency. Then an image segmentation algorithm is adopted to get the target, which is modelled for an improved MeanShift tracking framework. Experimental results show that the algorithm can significantly improve the accuracy.

1. Introduction

Tracker algorithm has been widely applied in commercial [1-2], intelligent transportation [3-4] and military field [5-7] along with the progress of computer and artificial intelligence technology. There are still a series of problems, such as the tracking accuracy under complex background conditions [8]. In contrast, human eyes can accurately locate targets in complex environments. Therefore, it can be referred by the principle of biological vision to provide an important scientific method for target information [9].

In order to solve these problems, a target tracking method based on saliency fusion is proposed in this paper. First of all, visual model and Kinect as a motion sensor are used to establish the gaze model of human vision, and extract saliency map which can be observed by human eyes. Secondly, get the sequence from the video and get the saliency map of the computer. Then, the two images are fused to a comprehensive image, and to segment target in the comprehensive image. Ultimately, the extracted target is tracked with MeanShift algorithm as Fig.1.

![Figure 1. General framework](image)
2. Saliency Detection

2.1. Human Vision Based Saliency Map
The Gauss mathematical model is used to obtain the gaze point map by Kinect, which extract the region of interest and omit the background information.

Step 1) Get the area of interest of human in the image, and use Kinect to get the plane coordinate value of the human eye's point of view.

Step 2) Model the retina structure and attentional mechanism of the human eyes and obtain its mathematical model.

Step 3) Use the mathematical model of Step 2) to model the gaze map and get the final human vision salient region, which is called the saliency map of human eyes.

The final saliency of human vision map can be obtained by merging the significant distribution map of the human eyes and the original image as Fig. 2.

![Figure 2](image1.png)

**Figure 2.** Human eyes saliency map. (a) The original image; (b) The degree of attention three dimensional distribution map; (c) Saliency map of human vision

2.2. Saliency Map of Computer
Computer saliency map is obtained as follows, firstly, ORB algorithm [10] is adopted for feature description and convex hull establishing. Then the internal and external saliency value of the convex hull is selected as the prior probability and likelihood probability respectively. Based on Bayesian theory, the acquired posterior probability can be defined as the saliency map of computer. So the calculation speed is fast, and it has relatively high accuracy.

2.3. Saliency Fusion
Based on saliency map of computer and saliency map of human vision, a reasonable fusion strategy is established through the idea of weighted value by Fig.3.

![Figure 3](image2.png)

**Figure 3.** Process of image fusion

Step 1) Calculate the influence factor of the weighted value;
Step 2) Define the centralization degree weight and the location weight value of the significant region block. Thus, the corresponding weight of average distance in saliency map between the significant region is:
\[ w_d = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[ (x_i - x_j)^2 + (y_i - y_j)^2 \right]^{1/2} \]  

(1)

where \((x_i, y_i), (x_j, y_j)\) is the heart of the salient region.

The more salient region near the centre, the greater contribution of the considered saliency map is:

\[ w_i = \frac{1}{N} \left[ (x_i - \bar{x})^2 + (y_i - \bar{y})^2 \right]^{1/2} \]  

(2)

where \((\bar{x}, \bar{y})\) is the central position of the image.

Step 3) Normalization. An image evaluation standard is structural similarity which is propose by paper [11]. This paper determines the fusion rule based on this. Get the fused image in Fig.4.

![Figure 4](image_url)

**Figure 4.** Image of fusion experiment. (a) Computer saliency map; (b) Visual saliency map; (c) Grayscale fusion result; (d) Color image fusion result

The experiment results show that the saliency map computer extraction often has some problems. But it will make the final saliency map was more focused on regional gaze position when combined with mechanism of the human vision. Fig.4 shows the advantages of fusion.

### 3. Mean Shift Tracking

#### 3.1. Target Segmentation

Image can be segment and get the target specific region after obtaining the saliency map. Grab Cut algorithm has been improved in this paper compared to the original algorithm as Fig.5. It improved without manual selection of seed points, which can improve the degree of automation of the whole calculation method.

![Figure 5](image_url)

**Figure 5.** Image segmentation

#### 3.2. Improved Mean Shift Tracking Algorithm

Equation of Mean Shift algorithm which is increased by the weights of sampling points \( w(x_i) \geq 0 \) and kernel function \( G(x) \):

\[ M_h(x) = \frac{\sum_{i=1}^{N} G_h \left( \frac{x_i - x}{h} \right) w(x_i) (x_i - x)}{\sum_{i=1}^{N} G_h \left( \frac{x_i - x}{h} \right) w(x_i)} \]  

(3)
Color range should be mapped to the feature space of pixels in the algorithm for target tracking, so this paper use RGB color space as the feature space. The vector can be calculated iteratively owing to the astringency of Mean Shift [12]. Thus, the target location will eventually converge to their true position.

4. Experiment Results and Analysis

The algorithm can effectively select object of initial frame of video sequence combined with computer and human vision. The size of square is selected according to the size of the target after the fusion results of two kinds of saliency map and target segmentation. Thus, it will not only achieve the target automatically selected, but also the size of square is reasonable. The key to capture the gaze of human eyes in each frame image to obtain the coordinates of the point value.

The Mean Shift algorithm improves the tracking results as shown in Fig.6:

![Figure 6](image_url)

**Figure 6.** Compare of tracking results. (a)(b) Tracking results of original Mean Shift algorithm for frames 96#, 169# (tracking directly); (c) (d) Tracking results of improved Mean Shift algorithm in this paper for frames 96#, 169# (tracking by detector)

It can be seen from the Figure 6 of the algorithm is improved, in the #96 and #169 of frame target calibration area are slightly offset through tracking directly. It can effectively modify the visible red square location through tracking by detector.

5. Conclusion

The algorithm in this paper achieves the combination of the human vision and computer vision, so that it can improve the tracking accuracy. But this practice is not intelligent, so the next step of the research program is to introduce learning mechanism to improve the level of intelligent system.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant number 61571346, the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University. The authors would like to thank the reviewers and the editor for their comments for improving this paper.

7. References

[1] Zhang P P, Yin-Li L U, Hong-Li Y U. Moving target tracking in video monitoring system [J]. Electronic Design Engineering, 2011.

[2] Qin X M. Research and realization on tracking of object in intelligent video surveillance [J]. Electronic Design Engineering, 2011.

[3] Xianye B, Wang K, Yang L I, et al. Multiple vehicle tracking algorithm based on Kalman filter in traffic video[J]. Applied Science & Technology, 2011.

[4] Jeong H, Yoo Y, Yi K M, et al. Two-stage online inference model for traffic pattern analysis and anomaly detection [J]. Machine Vision & Applications, 2014, 25(6):1501-1517.

[5] Dan C, Jing-Hua L I, Huang G Q, et al. Research on Recognition of Passive Acoustic Multitarget Based on Evidence Theory [J]. Journal of System Simulation, 2007.19(6):1323-1325.

[6] A Survey of Maneuvering Target Tracking-Part: Measurement Models. Li X R, Jilkov V P. Proc.Signal and Data Processing of Small Targets 2001.

[7] Xia P. Study of Accommodation of Maneuvering Target Tracking and Attacking Requirements for Submarine Equipment [J]. Ship Electronic Engineering, 2010.2010,30 (9):13-15.
[8] Xiao-Hang X U, Xiao G, Yun X, et al. Moving Object Tracking in Complex Background and Occlusion Conditions [J]. Opto-Electronic Engineering, 2013, 40(1): 23-30.

[9] Yilmaz A, Javed O, Shah M. "Object tracking": Acm computing surveys (CSUR), 38(4), p.13. 2006.

[10] Jie Han, Baolong Guo, Wei Sun. Target Tracking Method in Aerial Video Based on Saliency Fusion [C]. MEIC2015, Vol5, p724-727.

[11] Wei Sun, Objects detecting and tracking with a new particle filter; CECNet2012, Yichang, pp.3340 -3343, 2012.

[12] Li X R, Wu F C, Hu Z Y. Convergence of a mean shift algorithm [J]. Journal of Software, 2005, 16(3):365-374.