Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Multimodal Failure Matching Point Based Motion Object Saliency Detection for Unconstrained Videos

Jiang Qian, Jingkang Wei, Hui Chen & Gongping Chen

To cite this article: Jiang Qian, Jingkang Wei, Hui Chen & Gongping Chen (2022) Multimodal Failure Matching Point Based Motion Object Saliency Detection for Unconstrained Videos, Applied Artificial Intelligence, 36:1, 2110695, DOI: 10.1080/08839514.2022.2110695

To link to this article: https://doi.org/10.1080/08839514.2022.2110695

Published online: 22 Aug 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2110695 (3021 pages)
https://doi.org/10.1080/08839514.2022.2110695

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Multimodal Failure Matching Point Based Motion Object Saliency Detection for Unconstrained Videos

Jiang Qian[a], Jingkang Wei[b], Hui Chen[a], and Gongping Chen[c]

[a]Information Management Center, Anhui Business College, Wuhu, Anhui, China; [b]Information Services Department, Tianfu International Airport Branch Company, Chengdu, Sichuan, China; [c]College of Artificial Intelligence, Nankai University, Tianjing, China

ABSTRACT

Inspired by classical feature descriptors in motion matching, this paper proposes a multimodal failure matching point collection method, which is defined as FMP. FMP is, in fact, a collection of unstable features with a low matching degree in the conventional matching task. Based on FMP, a novel model for the saliency detection of motion object is developed. Models are evaluated on the DAVIS and SegTrackv2 datasets and compared with recently advanced object detection algorithms. The comparison results demonstrate the availability and effectiveness of FMP in the detection of motion object saliency.

ARTICLE HISTORY
Received 22 April 2022
Revised 30 July 2022
Accepted 3 August 2022

Introduction

In unconstrained videos containing complex scenes and movements, the human visual system pays accurate attention to motion object without training. For computers to achieve this visual capability, years of extensive research experiments have been utilized to explore and analyze visual significance. Research in related fields includes a static significance detection model built by multi-feature fusion learning (Jian et al., 2021), and a biology-based human eye gaze heuristic model for significance analysis by assessing eye gaze data (Parkhurst, Law, and Niebur 2002; Ramanathan et al. 2010). In the past few decades, many different schemes and meaningful models have been proposed and widely applied in various fields, including object saliency (Lin et al. 2019; Luo et al. 2011; Shi et al. 2012), image segmentation (Cheng et al. 2014), video compression (Guo and Zhang 2010), target tracking, behavior detection, video quality evaluation, object detection in low-contrast environments (Jian et al. 2019), etc.

Focusing on motion object saliency detection in video, Wei et al. (2012) exploited two common priors about backgrounds in natural images, namely boundary and connectivity priors. Using appropriate prior exploitation is helpful for ill-posed saliency detection. Yang et al. (2013) ranked the similarity of the image elements (pixels or regions) with foreground cues or background cues via

CONTACT Jingkang Wei jingkangw@163.com Information Services Department, Tianfu International Airport Branch Company, Chengdu, China

graph-based manifold ranking. Their Saliency detection is carried out in a two-stage scheme to detect background regions and foreground salient objects efficiently. Zhu et al. (2014) proposed a robust background measure, called boundary connectivity. It characterized the spatial layout of the image regions relative to image boundaries. they also proposed a principled optimization framework to integrate multiple low-level cues, including their background measure, to obtain clean and uniform saliency maps. Top-down models (Liu et al. 2007; Yang and Yang 2012) analyzed task-driven visual attention, which often entails supervised learning with class labels from a large set of training examples. Muwei et al. (2021) proposed an efficient video saliency-detection model, based on integrating object-proposal with attention networks. It combines static visual object features with spatial information to refine the final result through a neural network to effectively detect video significance. Jian et al. (2022) proposed a framework that uses local geometric structure information to estimate the center of mass of significant objects, further calculate the foreground and background robust seeds, and establish a reliable significance detection model.

Although we proposed many motion object saliency detection in video as described above. We observed that the existing models were not sufficient to effectively highlight the intact salient objects with suppressed background regions or were not successfully detecting the whole motion object with the interference of obvious background objects. An interesting point is that motion object significance detection is an innate ability of biological vision systems starting from the biological vision mechanism, unlike supervised learning tasks relying on backpropagation algorithms, and unsupervised motion object significance detection models are more consistent with this view. This is also a major motivator for this work.

As shown in Figure 1, in this paper, by analyzing the multimodal failure matching point (FMP) generated in the classical feature descriptor matching algorithm, a model of motion object saliency detection based on FMP is proposed. The model can accurately detect motion object saliency of unconstrained videos containing complex scenes and movements as well as suppress background regions.

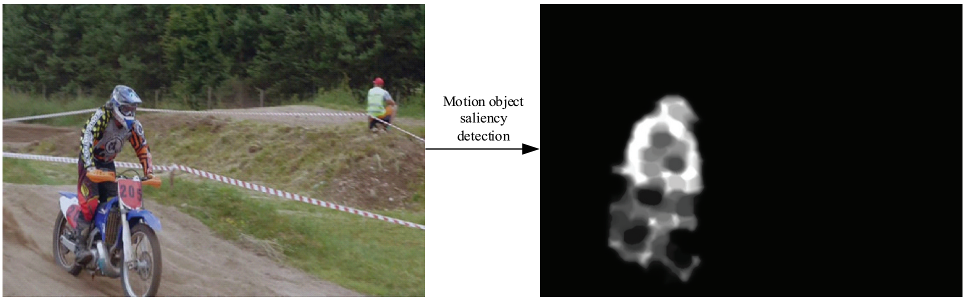

Figure 1. Motion object saliency detection.

First, we analyzed the matching results of the classical feature descriptors in the video and institute the Failure Matching Point (FMP) determination rules. Usually, these feature descriptors vary significantly between frames, with resulting from local distortions of the relative motion between the motion object and the background (Qian, Luo, and Xue 2021). Second, the multimodal feature descriptors were screened through a joint decision framework to obtain the FMP focused on the motion object. Finally, the amount of FMP was further enriched using convex bump-based concave detection and a Gaussian beam was used to represent the relative intensity of the motion object saliency. We evaluated and compared six advanced algorithms on DAVIS (Pont-Tuset et al. (2017)) and SegTrackv2 (Li et al. 2013) datasets. The main contributions of this paper can be summarized as follows:

- We found the basic law of the relative motion intensity and local feature distortion between the motion object and the background from the image feature and matching task, simulating the natural sensitivity of biological vision to the targets with high motion intensity in the scene without learning.
- We designed a joint decision framework to screen the multimodal feature descriptors, whose different modes are complementary to the differential features, enhancing the sensitivity of FMP to the significance detection of motor targets.
- We newly design an unsupervised motion saliency detection model, and demonstrate its effectiveness, while approximating the current more advanced supervised-like motion saliency detection model in accuracy.

The rest of this paper is arranged as follows. In Section 2, we first analyzed the relationship between motion object saliency and matching degree, and then described the collection method of FMP proposed in this paper. The proposed motion object saliency detection model is discussed in detail in Section 3. In Section 4, we provided the experimental results. Finally, the conclusions and future work are drawn in Section 5.

Definition of FMP

In this section, we first analyzed the relationship between motion object saliency and matching degree, and then introduced the definition and collection method of FMP.

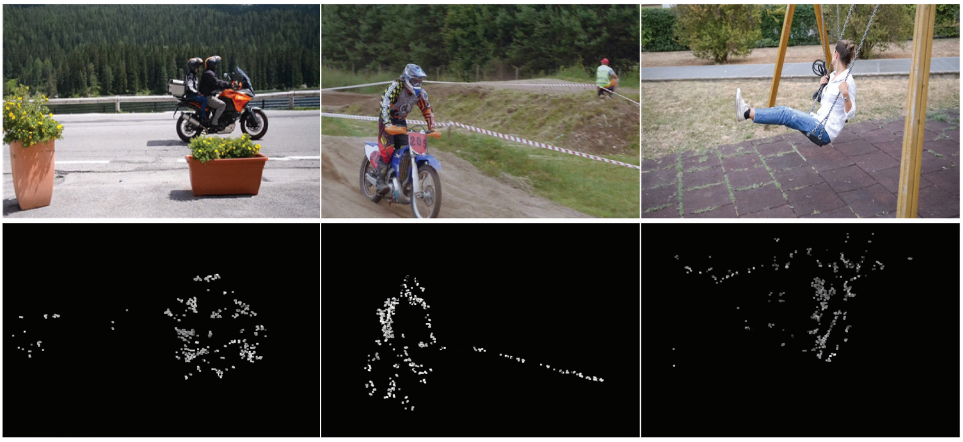

Figure 2. Schematic diagram of matching degree.

Matching Degree and Motion Object Saliency

Figure 2 shows the matching results of the ORB (Oriented FAST and Rotated BRIEF) feature descriptors in three public datasets, where the matching strength and gray values are inversely proportional. It can be observed that points with a low matching degree are concentrated in or near the edge region of the moving target. These low matching points are invalid in the regular matching task.

As shown in Figure 3, the first row represents the background region, and the second row represents the motion object area. As can be seen from the window in Figure 3, although the camera is in motion, the information between the cameras in the background is constant, and the two windows between the adjacent frames of the first line are highly similar. In the windows of the second row, the same part of the motion target shows displacement for

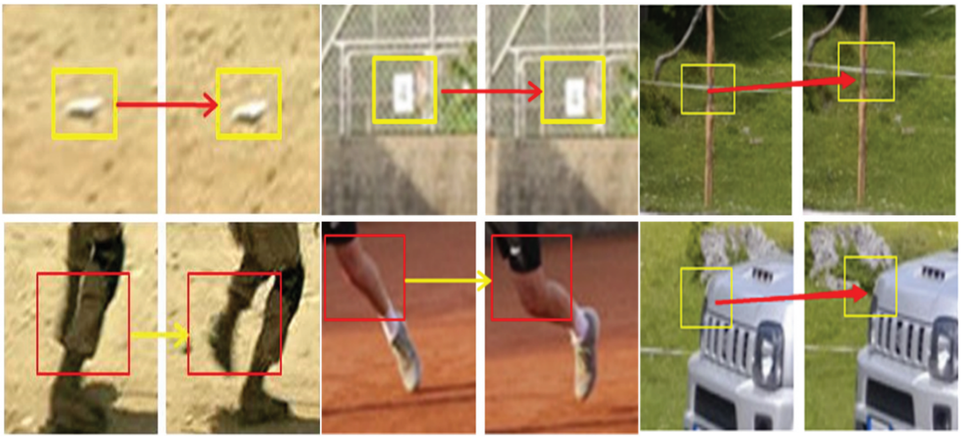

Figure 3. Foreground and background motion analysis.

the background owing to motion, and the motion target itself may undergo deformation, leading to a low similarity in local information of the same target and the close surrounding between two adjacent windows (frames) in the second row, which accounts for the observation that in Figure 2, the points with a low matching degree are concentrated at the motion target and in its close surrounding. We believe that motion object saliency refers to the local information difference caused by motion, and the greater the difference indicates the greater the relative intensity of motion object saliency. Therefore, this paper uses local matching methods to detect failure matches with a low matching degree, aiming to use these points to characterize the motion object saliency.

Formal Definition and Collection Method of FMP

In this paper, local failure matching point with low matching degree is used to represent the saliency of motion object, characterized by a low local matching degree. The local optimal matching of the feature descriptor will be used to describe its matching degree. Define the input video sequence $F = \{F^1, F^2 \ldots\}$,$P^k = \{p_1, p_2 \ldots\}$ representing the set of feature descriptor keys of frame K.

And define a weight matrix W^k of $\left|P^k\right| \times \left|P^{k+1}\right|$,where $\left|P^k\right|$ is the number of feature points of P^k. W^k is defined as:

$$W^k_{i,j} = \left\|M\left(p^k_i\right) - M\left(p^{k+1}_j\right)\right\|, p^k_i \in P^k, p^{k+1}_i \in P^{k+1} \tag{1}$$

where M represents the feature descriptor of feature point P. We use the local window W to find the locally optimal matching descriptor BM^k for the k-th frame

$$BM^k_i = minW^k_{i,j}, p^k_i \in P^k, p^{k+1}_i \in P^{k+1}, BM^k_i \in BM^k \tag{2}$$

In this way, a pair of matching vectors in the local window We are obtained. Select candidate descriptors characterized by descriptors with significant differences in the local vector field. Compare the current frame with the previous frame and the next frame respectively. If a candidate descriptor appears at the same location in all three frames, the descriptor is considered to be a stable feature descriptor with low local matching, which is FMP in this paper.

Figure 4 illustrates the main detection process of FMP. On the one hand, compare the current frame t with the previous frame $t - \Delta T$ and the next frame $t + \Delta T$ respectively, and obtain two sets of local matching points. On the other hand, using the random sample consensus (RANSAC) algorithm on the above two sets, subsets of points are extracted, and the points in each subset have the same matching vector

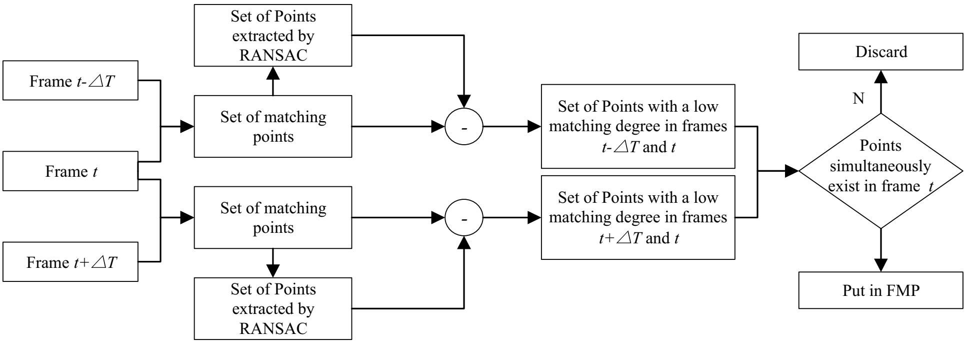
Frame *t*-△*T*
Frame *t*
Frame *t*+△*T*
Set of Points extracted by RANSAC
Set of matching points
Set of matching points
Set of Points extracted by RANSAC
-
-
Set of Points with a low matching degree in frames *t*-△*T* and *t*
Set of Points with a low matching degree in frames *t*+△*T* and *t*
Discard
N
Points simultaneously exist in frame *t*
Put in FMP

Figure 4. FMP definition flowchart.

field. $DK^t(p_i^t)$ represents the inter-frame offset of the best matching descriptor p_i in the *t-th* frame. PB represents the background descriptor with the same motion law as the camera:

$$DK^t(p_i^t) = \left\|dx(p_i^t) + dy(p_i^t)\right\|, p_i^t \in P^t \tag{3}$$

$$PB^t = \left\{p_i^t | u(DK^t) > DK^t(p_i^t)\right\}, p_i^t \in P^t \tag{4}$$

where $dx(p_i^t), dy(p_i^t)$ represents the offset of p_i^t in the x and y directions of two adjacent frames. $u(DK^t)$ is the mean of DK^t. The optimal matching subset consists of *PB*. Two distinct descriptor subsets are obtained by subtracting the optimal matching subset from the corresponding matching set. One weakly matching descriptor subset is obtained from frame t and $t - \Delta T$, and the other is obtained from frame t and $t + \Delta T$ If descriptors in two weakly matching descriptor subsets appear at the same position in frame t, such descriptors are finally defined as FMP.

The dynamic changes of motion object on time scales are different. When the foreground object moves slowly, the optimal motion object region cannot be extracted efficiently simply by matching 3 adjacent frames. In this paper, we increased the number of frames matched in the experiment and find the optimal motion object region by matching 5 adjacent frames.

Table 1 shows that 5-frame analysis can detect more feature descriptors than 3-frame analysis. Although increasing the number of FMP in the background, the percentage increase was much smaller than that on the motion object. The reason is that FMP is mainly focused on the marginal regions with staggered motion object and background. The 5-frame analysis increases the retrieval range of the FMP, obtaining more FMP to describe the motion object and the degree of relative movement.

Table 1. Three-frame analysis and five-frame analysis data comparison.

	FMP from three-frame analysis	FMP from five-frame analysis	Percentage of FMP in Ground Truth by three-frame method	Percentage of FMP in Ground Truth by five-frame method
Motor Cross	271	423	93.8%	94.6%
Parkour	110	237	91.3%	89.4%
Tennis	138	241	92.2%	94.2%
Soldier	230	546	95.5%	97.8%
Horse Jump	130	226	89.3%	93.7%
swing	353	642	87.9%	91.1%
bmx-bumps	122	243	91.7%	92.3%
kite-surf	242	465	84.6%	88.4%
motorbike	145	304	93.1%	94.5%
paragliding-launch	253	497	88.9%	90.7%
girl	94	158	97.3%	96.3%

In some experiments, the three-frame analysis method detected a higher percentage of FMP, and it was the difference in camera motion rate that resulted in more FMP in the background. The comprehensive experimental results show that the five-frame analysis method is more effective in the detection of FMP in moving target regions.

Motion Object Saliency Detection Based on FMP

In this section, we will discuss motion object saliency detection model based on multimodal FMP. The detection model can be seen as a reverse application of the neglected failure matching point in the routine matching tasks of multiple classical feature descriptors, mainly due to the local distortion caused by the relative motion object and the background.

Relevant Work

As shown in Figure 5, any 3 feature descriptors are used to detect the failure matching point of low matching degree, providing 3 sets of FMP set. For each algorithm, allowing loose constraints on the number of feature descriptors, and a set of loose FMP set with large numbers of FMP have been generated, while the other two algorithms use strict constraints to generate 2 sets of strict FMP set. The 2 sets of strict FMP set were used to determine removing the FMP present on the background in the loose FMP set, while increasing the number of FMP in the motion object region. Next, the FMP is superimposed to form a new set, using a convex packet-based concave detection algorithm to further clarify the FMP within the motion object region, and then expresses the saliency of the motion object using a Gaussian beam. Finally, the motion object were filled using a morphological closed manipulation.

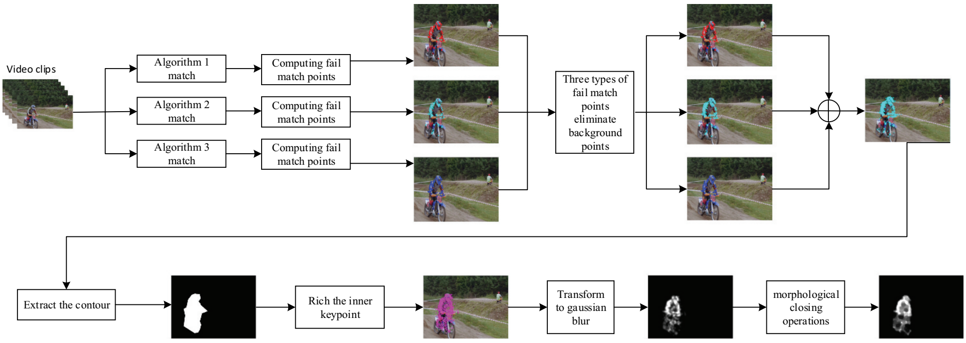
Video clips
Algorithm 1 match
Algorithm 2 match
Algorithm 3 match
Computing fail match points
Computing fail match points
Computing fail match points
Three types of fail match points eliminate background points
Extract the contour
Rich the inner keypoint
Transform to gaussian blur
morphological closing operations

Figure 5. General flowchart of experiment.

Motion Object Saliency Detection Model Based on Multimodal FMP

In this paper, three classical algorithms of SIFT (Lowe 1999), ORB (Rublee et al. 2011), and AKAZE (Alcantarilla, Nuevo, and Bartoli 2013) were selected as the input of the joint decision framework to remove random FMP in the background. The reason for choosing them is that the spatial complementarity of these three algorithms can enrich the number of FMP, making the results of the joint decision-making more accurate.

These algorithms are allowed to have a relaxed constraint on the number of point sets, resulting in a large number of relaxed FMP. Although several FMP are added in the target area, a considerable number of FMP appear randomly in the background. With a rigorous constraint, the three algorithms retrieve fewer FMP, but most of them are concentrated in the motion target area, with extremely few points randomly appearing in the background. Using the rigorous FMP generated by two algorithms, a joint decision is made to remove random FMP that are in the relaxed FMP of the third algorithm but randomly appear in the background.

For one of the three feature detection algorithms, its relaxed FMP is denoted as *R*, whereas the rigorous FMP of the other two algorithms is denoted as *T1* and *T2*, respectively.

In the *k-th* frame, the rigorous FMP subset is defined as $T_m, m = 1, 2, 3 \ldots n$ the relaxed FMP subset is defined as $R_m, m = 1, 2, 3 \ldots n$. Using *t* as an example, we define the position distance dist between two descriptors in the set:

$$dist^k(p_i, p_j) = \left\| V_i^k - V_j^k \right\|, p_i, p_j \in T_m \tag{5}$$

where *V* represents the coordinate position of the descriptor *p*. Then we calculate the average distance $mean_m^k(p_i)$ from the descriptor *p* to the other descriptors in the set T_m, and the variance $u_m^k(p_i)$ between them

$$mean_m^k(p_i) = \frac{\sum_{j=1}^{w} dist_m^k(p_i, p_j)}{w}, j \neq i, p_i, p_j \in T_m \tag{6}$$

$$u_m^k(p_i) = \frac{\sum_{j=1}^{w} \left[dist_m^k(p_i, p_j) - mean_m^k(p_i)\right]^2}{w}, j \neq i, p_i, p_j \in T_m \tag{7}$$

The other *n-1* sets are also calculated using the above formula. We define a decision algorithm *f*, and calculate the average distance $mean_R^k$ from each keypoint p_i^k in the relaxed FMP subset to the rigorous FMP subset. And find the two rigorous WMD sets T_m closest to the keypoint p_i^k.

The $\max\left(u_m^k(p_i)\right)$ represents the maximum variance of T_m, it is the variance value of the descriptor with the largest deviation in T_m. If the variance of the current p_i in the two rigorous FMP subsets is less than the maximum variance $\max\left(u_m^k(p_i)\right)$ of the two subsets, the descriptor is defined as FMP. Finally, this paper divides the feature descriptors in the relaxed FMP subset into two categories: FMP and BG descriptors:

$$WMD = \left\{ p_i^k | \frac{\sum_{j=1}^{w} \left[dist_m^k(p_i, p_j) - mean_m^k(p_i)\right]^2}{w} < \max\left(u_m^k(p_i)\right) \right\} \tag{8}$$

$$BG\,descriptors = R^k - WMD \tag{9}$$

With ΔT in Figure 5 being set to 1 and 2, the FMP results for frame t are obtained, and the two results are superimposed, with the removal of FMP that randomly appear on the background. As shown in Figure 6(a), there are a total of 562 points before the joint decision of the three algorithms; after the joint decision, a total of 756 FMP are acquired. When the acquired FMP is high enough in quantity, the outline of the motion target is already clear. To further enrich the information inside the motion object area, a convex hull-based concave point detection algorithm – a method suited for a set of discrete points – is adopted here to detect the outline of the discrete points. The FMP that had been removed according to the decision are now included within the motion target outline to enrich the motion object saliency, as shown in Figure 7, wherein the number of FMP in Figure 7(a) is 756; however, the number increases to 1,146 inside the motion target area after the enrichment process.

As the feature descriptors used in the three algorithms have a diameter of 32 pixels, it is deemed here that the range of motion object saliency expressed by each FMP should be 32 pixels in diameter. Thus, Gaussian beam spots, each having a diameter of 32 pixels, are used to represent the feature descriptors, and the gray value at the center of each spot is set to 25. Then use

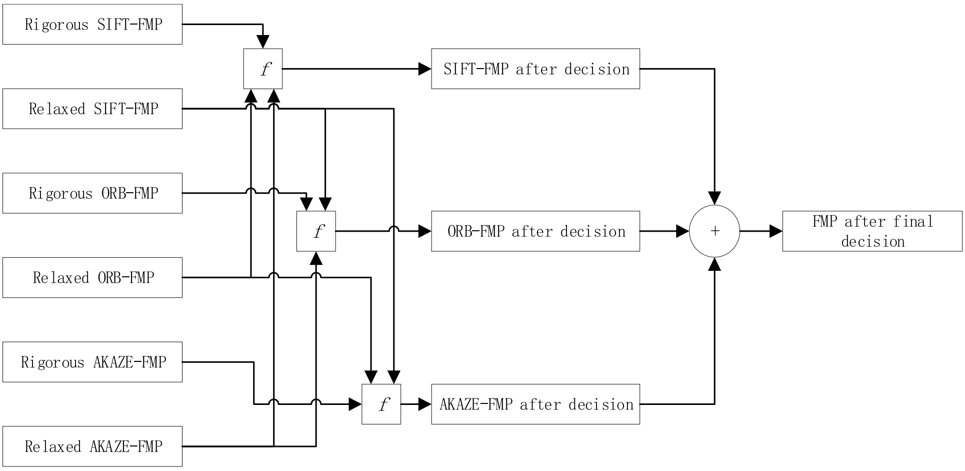

Figure 6. Flowchart of three algorithms decision.

morphological closing operations to fill the target. By doing so, motion object saliency is generated as shown in the above figure, with the highlighted portions being regarded as having relatively obvious motion, as depicted in Figure 7(d).

Experimental Results and Discussion

To evaluate motion object saliency detection model based on multimodal FMP proposed in this paper, the DAVIS and SegTrackv2 benchmark datasets were used in the experiment. In the experiment, the standard precision–recall curve (PR curve), F-score, and average absolute error (MAE) were used as performance indicators, comparing seven recent advanced object saliency detection algorithms subjectively and objectively, including GS (Geodesic saliency using background priors) (Wei et al. 2012), MR (Saliency detection via graph-based manifold ranking) (Yang et al. 2013), RBD (Zhu et al. 2014), SR (Hou and Zhang 2007), SF (Perazzi et al. 2012), SA(Wang et al. 2018), CAG(Chen et al. 2021).

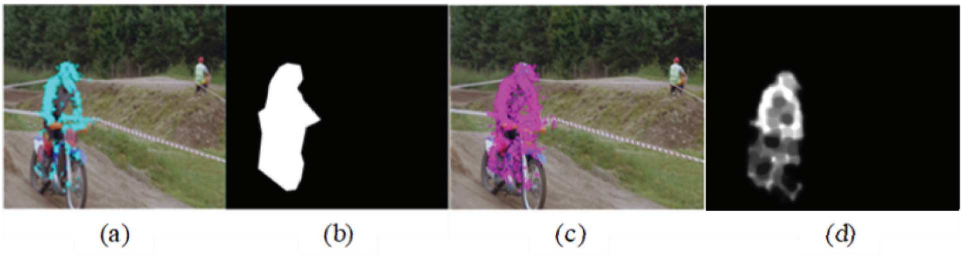

Figure 7. Draws the outline of a set of points and use Gaussian beam spots represent motion object saliency.

Subjective Evaluation

The experimental results for continuous scenes are shown in Figure 8, where the first row shows the original images of the swing sequence, and the third row shows the motocross-bumps sequence, and the second and fourth rows show the results of motion object saliency detection in this paper.

As we can see from Figure 9, in the scene with a relatively prominent background, the partial contrast algorithm cannot detect the significance of motion object well in the video, while the partial algorithm can detect more backgrounds. In this paper, our algorithm can detect the motion object saliency clearly and there is no background noise distributed in the result. Although the intensity of the local detail part of the motor target is weak, the local difference in the movement intensity on the motor target is also appropriately expressed.

Objective Evaluation

For performance evaluation, we used the standard precision–recall curve (PR curve), F score, and MAE (mean absolute errors). PR curve is one of the important indexes used to evaluate the performance of the model, and the F score is the harmonic average of precision and recall, which is used to evaluate the overall performance. We set the to stress the importance of precision (Achanta et al. 2009). MAE is to directly calculate the significance map binarization of the model output and compare it with Ground-truth to obtain the mean absolute error between them.

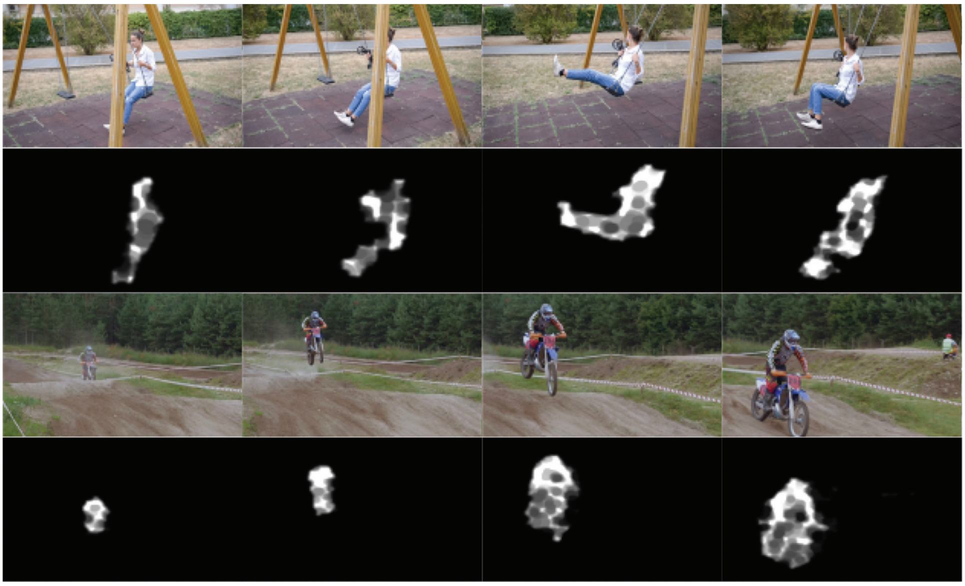

Figure 8. Experimental results for continuous scenes.

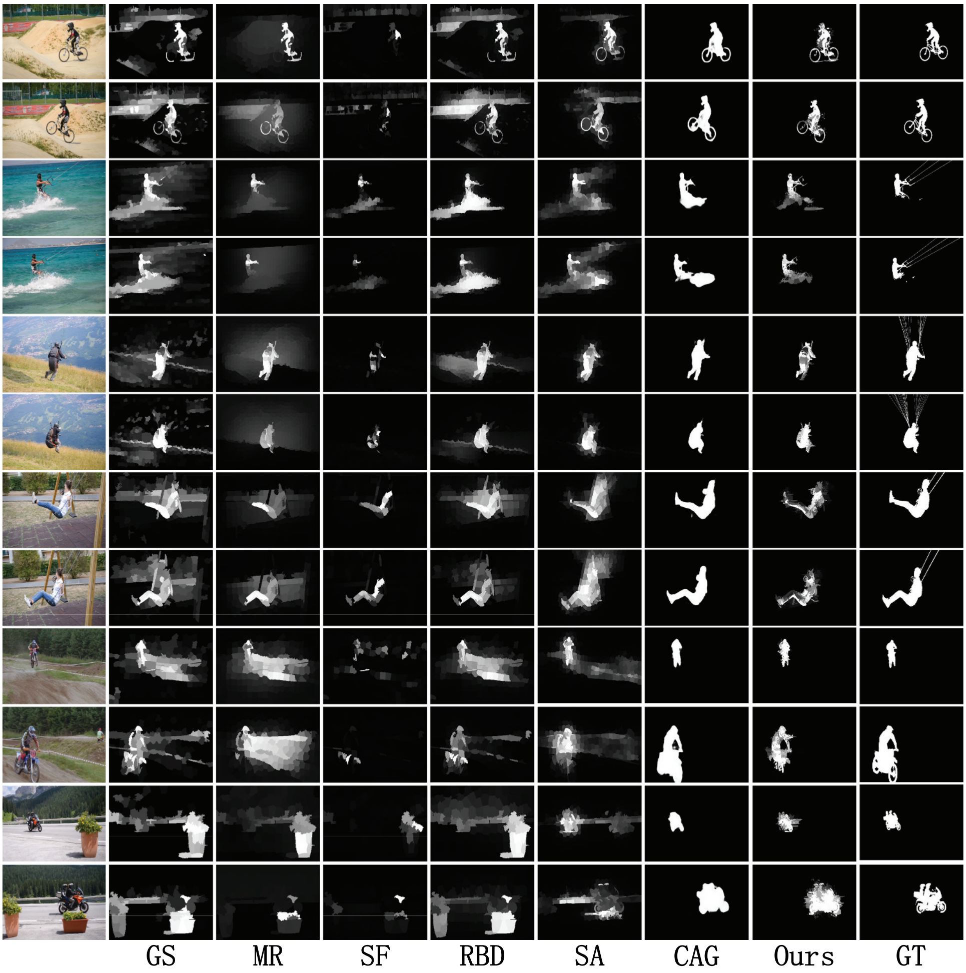

Figure 9. Experimental results of FMP and other six advanced object saliency detection algorithms results in the DAVIS and SegTrackv2 datasets.

From Figure 10, we can see that our results and detection rates remain more stable with the increasing recall rate in (a) than most algorithms, one of which is that our algorithm can successfully suppress background noise. Our average F-score value in (b) is around 0.8, and the average MAE value is below 0.05, approximating the CAG model of fusion color and optical flow features, which means that our results in this paper are close to Ground Truth.

Discussion

The motion object saliency detection model based on multimodal FMP is a low matching degree of failure matching point (FMP) reverse application model, in which FMP is inspired by the innate biological vision mechanism of human attention to complex scenes or complex movements. FMP and the degree of motion object movement are strongly related. The detection model was evaluated

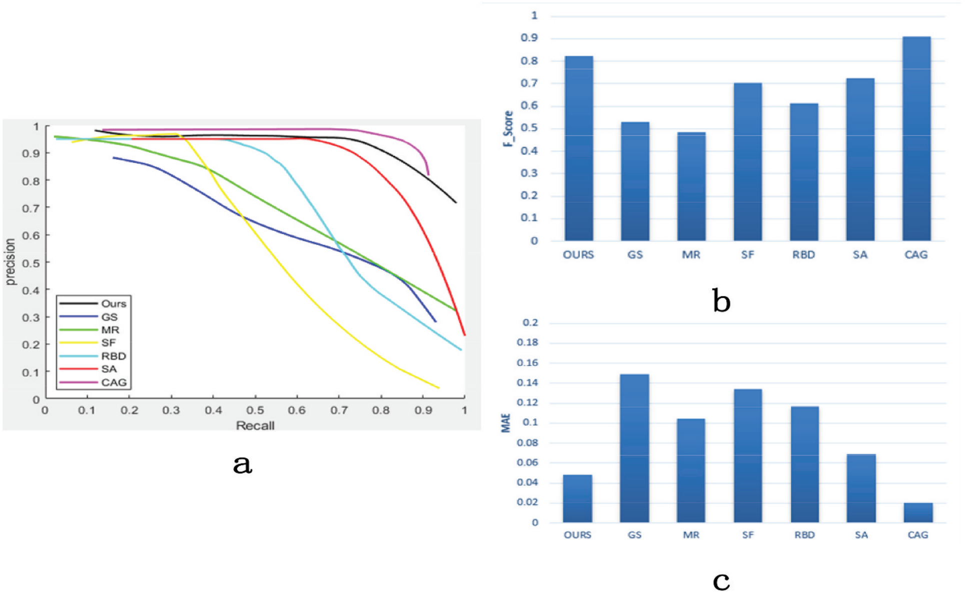

Figure 10. a is the graph of PR, b is the histogram of F score, and c is the histogram of MAE.

on the DAVIS and SegTrackv2 benchmark datasets. Figures 9 and 10 provide the results of the comparison with the other algorithms. Based on these experimental data, we can get an overall conclusion: the multimodal-based motion object saliency detection model has relatively good performance in the DAVIS and SegTrackv2 benchmark datasets, and our algorithm can effectively suppress the noise from the background and clarify the relative saliency of the motion object. Since the underlying features of multimodal FMP originate from classical feature descriptor detection methods, this indicates that further analysis and consideration are needed in feature descriptor selection.

Conclusions and Future Work

Feature descriptor is a classical model that describes the typical features of an image and simplifies the effective information of the image. Based on the mechanism of biological visual saliency perception of motion object, this paper proposes a new idea to effectively use the failure matching point (FMP). Based on the multimodal FMP, we developed a motion object saliency detection model. Essentially, the detection model proposed here is proposed in the practical phenomenon where the FMP and the degree of motor target motion are correlated. By evaluating the DAVIS benchmark dataset with seven models. The comparative results demonstrate the usability and effectiveness of the proposed model in the motion object saliency detection task. In addition to borrowing existing classical feature descriptors, a thought worth exploring is to imitate the innate sensitivity of biological vision to motor targets without extensive training and learning. In future work, we will investigate how to design some new detection models for more computer vision tasks starting from biological vision phenomena.

Acknowledgments

This work is supported by the Major Project Foundation of the Education Department of Anhui (Grant No. KJ2021A1482), the Fund of Anhui Province Project(Grant No. 2020kfkc204).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Anhui Department of Education2020kfkc204; Anhui Department of Education KJ2021A1482.

References

Achanta, R., S. S. Hemami, F. J. Estrada, and S. Süsstrunk. 2009. Frequency-tuned salient region detection. *CVPR*. doi:10.1109/CVPR.2009.5206596.

Alcantarilla, P. F., J. Nuevo, and A. Bartoli. 2013. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *Bmvc*. doi:10.5244/C.27.13.

Chen, P., J. Lai, G. Wang, and H. Zhou. 2021. Confidence-guided adaptive gate and dual differential enhancement for video salient object detection. 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 1–6.

Cheng, M., N. J. Mitra, X. Huang, P. H. Torr, and S. Hu. 2014. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (3):569–82. doi:10.1109/TPAMI.2014.2345401.

Guo, C., and L. Zhang. 2010. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing* 19:185–98. doi:10.1109/TIP.2009.2030969.

Hou, X., and L. Zhang. 2007. Saliency detection: a spectral residual approach. 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 1–8. doi:10.1109/CVPR.2007.383267.

Jian, M., Q. Qi, H. Yu, J. Dong, C. Cui, X. Nie, H. Zhang, Y. Yin, and K. Lam. 2019. The extended marine underwater environment database and baseline evaluations. *Applied Soft Computing* 80:425–37. doi:10.1016/j.asoc.2019.04.025.

Jian, M., R. Wang, H. Xu, H. Yu, J. Dong, G. Li, Y. Yin, and K. Lam. 2022. *Robust seed selection of foreground and background priors based on directional blocks for saliency-detection system*. Multimedia Tools and Applications, , 1-25.

Jian, M., J. Wang, H. Yu, and G. Wang. 2021a. Integrating object proposal with attention networks for video saliency detection. *Information Sciences* 576:819–30. doi:10.1016/j.ins.2021.08.069.

Jian, M., J. Wang, H. Yu, G. Wang, X. Meng, L. Yang, J. Dong, and Y. Yin. 2021b. Visual saliency detection by integrating spatial position prior of object with background cues. *Expert Systems with Applications* 168:114219. doi:10.1016/j.eswa.2020.114219.

Li, F., T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. 2013. Video segmentation by tracking many figure-ground segments. 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 2192–99.

Lin, X., Z. Wang, L. Ma, and X. Wu. 2019. Saliency detection via multi-scale global cues. *IEEE Transactions on Multimedia* 21:1646–59. doi:10.1109/TMM.2018.2884474.

Liu, T., Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. 2007. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33:353–67. doi:10.1109/TPAMI.2010.70.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. nternational Conference on Computer Vision, 1150, Kerkyra, Greece. Ieee. doi:10.1109/ICCV.1999.790410.

Luo, Y., J. Yuan, P. Xue, and Q. Tian. 2011. IEEE transactions on circuits and systems for video technology. *Saliency Density Maximization for Efficient Visual Objects Discovery* 21:1822–34. doi:10.1109/TCSVT.2011.2147230.

Parkhurst, D. J., K. Law, and E. Niebur. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research* 42:107–23. doi:10.1016/S0042-6989(01)00250-4.

Perazzi, F., P. Krähenbühl, Y. Pritch, and A. Sorkine-Hornung (2012). Saliency filters: Contrast based filtering for salient region detection. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 733–40. doi:10.1109/CVPR.2012.6247743.

Pont-Tuset, J., F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. V. Gool. (2017). The 2017 DAVIS challenge on video object segmentation. arXiv preprint arXiv:1704.00675 .

Qian, J., X. Luo, and Y. Xue. 2021. Half-edge composite structure: Good performance in motion matching. *Journal of Ambient Intelligence and Humanized Computing* 1-16. doi:10.1007/s12652-021-03073-4.

Ramanathan, S., H. Katti, N. Sebe, M. Kankanhalli, and T. Chua. 2010. An eye fixation database for saliency detection in images. *ECCV*. doi:10.1007/978-3-642-15561-1_3.

Rublee, E., V. Rabaud, K. Konolige, and G. Bradski (2011). ORB: An efficient alternative to SIFT or SURF. International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011. IEEE. . doi:10.1109/ICCV.2011.6126544.

Shi, R., Z. Liu, H. Du, X. Zhang, and L. Shen. 2012. Region diversity maximization for salient object detection. *IEEE Signal Processing Letters* 19:215–18. doi:10.1109/LSP.2012.2188388.

Wang, W., J. Shen, R. Yang, and F. M. Porikli. 2018. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40:20–33. doi:10.1109/TPAMI.2017.2662005.

Wei, Y., F. Wen, W. Zhu, and J. Sun. 2012. Geodesic saliency using background priors. *ECCV*. doi:10.1007/978-3-642-33712-3_3.

Yang, J., and M. Yang. 2012. Top-down visual saliency via joint CRF and dictionary learning. *CVPR*. doi:10.1109/CVPR.2012.6247940.

Yang, C., L. Zhang, H. Lu, X. Ruan, and M. Yang. 2013. Saliency detection via graph-based manifold ranking. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 3166-3173, Portland, OR, USA. doi:10.1109/CVPR.2013.407.

Zhu, W., S. Liang, Y. Wei, and J. Sun. 2014. Saliency optimization from robust background detection. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2814-2821, Columbus, OH, USA. doi:10.1109/CVPR.2014.360.