A Biologically Inspired Appearance Modeling and Sample Feature-based Approach for Visual Target Tracking in Aerial Images

Lili Pei[*1], Xiaohui Zhang[2]
College of Architectural Engineering, Tangshan Polytechnic College, Tangshan, 063299, China[1]
College of Information Engineering, Tangshan Polytechnic College, Tangshan, 063299, China[2]

***Abstract*—Visual tracking in uncrewed aerial vehicles is challenging because of the target appearance. Various research has been fulfilled to overcome appearance variations and unpredictable moving target issues. Visual saliency-based approaches have been widely studied in biologically inspired algorithms to detect moving targets based on attentional regions (ARs) extraction. This paper proposes a novel visual tracking method to deal with these issues. It consists of two main phases: spatiotemporal saliency-based appearance modeling (SSAM) and sample feature-based target detection (SFTD). The proposed method is based on a tracking-by-detection approach to provide a robust visual tracking system under appearance variation and unpredictable moving target conditions. Correspondingly, a semi-automatic trigger-based algorithm is proposed to handle the phases' operation, and a discriminative-based method is utilized for appearance modeling. In the SSAM phase, temporal saliency extracts the ARs and coarse segmentation. Spatial saliency is utilized for the object's appearance modeling and spatial saliency detection. Because the spatial saliency detection process is time-consuming for multiple target tracking conditions, an automatic algorithm is proposed to detect the region saliences in a multithreading implementation that leads to low processing time. Consequently, the temporal and spatial saliencies are integrated to generate the final saliency and sample features. The generated sample features are transferred to the sample feature-based target detection (SFTD) phase to detect the target in different images based on samples. Experimental results demonstrate that the proposed method is effective and presents promising results compared to other existing methods.**

Keywords*—Visual tracking; biologically inspired; visual saliency detection; appearance modeling; attention region; spatiotemporal

I. Introduction

Visual tracking is mainly investigated as an active research topic according to its wide range of applications, including smart surveillance systems, intelligent remote sensing technologies, action recognition, and robotic and human-computer interaction [1]. Particularly, visual target tracking is broadly studied on uncrewed aerial vehicles (UAV) to detect and track targets on aerial images. As opposed to applications with fixed cameras, for example, traffic monitoring, aerial videos have the favorable circumstances of higher portability and superior reconnaissance and surveillance [2]. However, visual tracking systems are still suffered from various challenges and difficulties. Target appearance variations and tracking the target under uncontrollable and unpredictable conditions are as main challenges of these systems [3, 4]. In this regard, online learning-based tracking methods that can incrementally update feature representations have received more attention for achieving a reliable and robust visual tracking algorithm [5]. Therefore, an online target feature representation is crucial for preserving an efficient appearance model to describe and identify the target from background [3, 6].

Recently, biologically inspired and cognition-based approaches have become a topic of interest by many researchers [7]. These approaches are inspired by human biological mechanisms, meaningfully indicating that human perception is sensitive to attentional regions (ARs) [5, 8]. By adopting biologically inspired approaches, visual saliency detection methods have been presented to detect attentional regions based on spatial and temporal information, resulting in the design of a significant number of saliency-based object detection methods [2]. Hence, a fascinating question is how we can exploit and take advantage of these approaches to develop a more powerful visual tracking algorithm [9]. Current visual saliency detection approaches can be categorized as task-driven attention (top-down) and data-driven attention (bottom-up) [10-13]. The top-down approach is a result of long-term visual simulation with prior knowledge [10]. The top-down approaches focused on high-level information investigation, such as the sky, faces, and humans [14]. This approach's drawback is a hard generalization because it is not simply obtained from images [15].

Furthermore, they are slow and computationally expensive [10]. On the other hand, the bottom-up approaches are based on low-level visual features simulating the formation of short-term visual attention [16]. In contrast to the top-down method, the bottom-up-approaches are rapid [14, 17]. As promising results indicated on bottom-up based approaches, this study focuses on the bottom-up approach.

II. Related Works

In this section, existing related works discuss two categories, visual saliency-based, and appearance modeling-based methods.

A. Visual Saliency-Based Approach

As discussed previously, saliency-based methods are categorized as bottom-up and top-down methods. The bottom-

up methods are categorized into temporal, spatial, and combined-based approaches [2]. The details of each approach discuss to address the advantages and disadvantages.

1) Temporal saliency: Detection of salient regions is highly dependent on the recognition of moving objects since motions attract more attention [18]. For moving object detection from a video, moving object detection methods are mainly based on temporal information, such as background subtraction [19, 20], frame difference [21, 22], and optical flow [23].

Background subtraction is based on background modeling. It is widely used for moving object detection. It segments foreground objects from the image background to detect objects that are not moving [24]. However, the background subtraction method suffers from some limitations. These methods are sensitive to the fixed background, and the object extraction fails when the background has changed [10].

Optical flow is also used for moving object detection, even under moving camera conditions. However, optical flow is computationally expensive and sensitive to noise. Thus, optical flow algorithms are not robust in real-time visual tracking systems [2, 25].

Frame differencing as a practical approach for moving object detection is based on pixel-wise difference extraction among the image frames. The frame difference does not require background modeling and is not sensitive to fixed background-like background subtraction methods. Frame differencing is adaptive to dynamic backgrounds and has a low computational cost [2]. However, one of the major drawbacks of frame differencing is moving the objects during the frame capturing process. Because a target may have unpredictable motions, such as stop-and-go periods, the frame difference method is not robust under uncontrollable and unpredictable target movement conditions [26]. Therefore, it is required to propose a temporal saliency detection method to overcome the mentioned challenges [25].

2) Spatial saliency: Spatial saliency detection is based on low-level feature representation and focuses on salient region extraction from images. These features are investigated to describe and identify the region of interest as salient regions [2,40]. Several spatial-based methods have been developed, such as region low rank matrix recovery [27], region covariance [28, 29], color-based [41], contrast-based [30], frequency domain [31] and graph [32] methods. As mentioned earlier, visual object tracking in aerial has difficulties in target appearance variations and background changes. To deal with difficulties, spatial saliency detection can extract the salient regions that are moving targets. This method is not sensitive to background changes and abrupt motion.

3) Spatiotemporal saliency: Spatiotemporal saliency detection methods calculate the temporal and spatial saliencies [10] separately., They used motion cues investigation for moving object detection. Generally, the results only based on motion cues are not undesirable because of the lack of spatial distribution [2]. Therefore, it is required to integrate the temporal and spatial features to detect the salient regions more accurately [9, 10,46].

B. Appearance Modeling-based Approach

Appearance modeling approaches are normally used to deal with appearance variations challenges. These approaches are categorized into generative and discriminative-based methods [5].

1) Generative-based methods: Generative-based methods are used to generate a model of an object during appearance changes in scenes. The generated model exploits the discriminative features to handle the target's appearance variations. The mechanism of appearance model generation is frequently updated online to describe the appearance variations. Some generative-based methods are as follows. Lee and Kriegman [33] proposed a generative-based algorithm to update a model for target detection dynamically. Wu and Wang [34] proposed a real-time generative method integrated with an incrementally updating covariance modeling approach for visual tracking. Tianxiang and Li [35] presented an appearance modeling approach based on a generative method and structured sparse representation for tracking an object in a video. However, even though various generative-based methods have been proposed, these methods still have not fully exploited spatial identification within the images efficiently.

2) Discriminative-based methods: The discriminative-based method has also been utilized to overcome the challenges related to appearance changes during visual tracking. The discriminative method are called tracking-by-detection. The mechanism for discriminative-based methods is a separate set of features that are extracted such that they distinguish the target from the background image. A binary separation approach is used for target identification from the background in successive frames. Various studies have been conducted based on discriminative-based methods, such as the discriminative learning method based on graph embedding proposed by Zhang et al. [37]. Fan et al. [38] presented an approach of discriminative region attention that describes the target from the background in terms of spatial features. Their proposed method aimed to overcome the spatial distraction in the visual appearance changes challenging. Tang et al. [39] proposed a robust visual tracking method, DRLTracker, based on a discriminative ranking list approach. DRLTracker utilizes the ranking lists and two-scale features to generate a model of the target and recognize it from the background based on the ranking lists of generated patches. However, the proposed method is limited to two-scales DRLTracker and suffers from high processing time.

This study investigates an enhanced discriminative-based appearance modeling approach to overcome the noise shortcoming and appearance variations difficulty. This study uses appearance modeling instead of discriminative-based appearance modeling term in this paper. The appearance

modeling approach details are discussed in the Material and Methods section.

Finally, the core of this paper is the proposal of visual object tracking based on a combination of spatiotemporal saliency appearance modeling and sample feature-based approaches. The proposed method is based on a tracking-by-detection approach to provide a robust visual tracking system in appearance variation conditions. Correspondingly, a semi-automatic trigger-based algorithm is proposed to handle the phases' operation. Furthermore, an automatic algorithm is proposed to detect the region saliences in a parallel implementation that leads to low processing time. Consequently, the temporal and spatial saliencies are integrated to generate the final saliency and sample features. Our contributions can be summarized as follows,

- A visual tracking method based on spatiotemporal saliency-based appearance modeling (SSAM) and sample feature-based target detection (SFTD) to preserve the visual target tracking robustly under appearance variation conditions, unpredictable motion, and low processing time. The proposed method is efficient in both camera and target moving platforms.
- Develop a novel algorithm for switching automatically between phases to handle their operation based on trigger activation.
- An algorithm is proposed for multiple target detection based on dynamic multithreading implementation of SLIC segmentation algorithm.

The remainder of this paper is organized as follows. Material and Methods section discusses the proposed framework and details of the material and methods. The results and discussion section presents our experimental and performance analysis. Finally, the conclusion section concludes this study.

III. Material and Methods

This section presents an overview and the details of the proposed approach. The underlying goal of the proposed approach is to take advantage of saliency values and appearance modeling in an efficient manner for target detection.

In this study, the proposed method consists of two main phases, spatiotemporal saliency appearance modeling (SSAM) and sample feature-based target detection (SFTD). To handle the phases' operation, a semi-automatic trigger-based algorithm is proposed to switch between the two phases; a phase operation is started when that phase receives a trigger activation. For example, when the saliency-slot time is reached, the SSAM phase activates a trigger to switch to the SFTD phase. The proposed method defines the saliency slot to activate the trigger. The SFTD phase activates a trigger when it cannot detect any objects. The overall architecture of the proposed framework is shown in Fig. 1. The details of the proposed approach are presented in the following sections.

A. Spatiotemporal Saliency Appearance Modeling (SSAM) Phase

This phase involves three stages, temporal saliency and localization detection, spatial and final saliency detection, and, finally, sample feature generation and target detection stages.

1) Temporal saliency and localization detection (TSLD) stage: This stage consists of temporal saliency detection and localization modules. In order to extract the moving target, salient regions are extracted using motion cues detection. For motion cue detection, temporal saliency is investigated with the following details.

2) Temporal saliency detection module: The purpose of temporal saliency is for attention region (ARs) extraction and coarse segmentation. The extracted attentional regions are called Candidate Motion Regions (CMRs). To extract the CMRs, we propose the following steps,

Frame differencing is used for temporal saliency detection. Frame differencing is utilized to identify moving objects in consecutive frames. This technique employs the image subtraction operator, which takes two images (or frames) as input and produces the output [42].

Image Enhancement. Morphological operations are generally applied for image enhancement [43]. This proposed method uses these operators, which are dilation, erosion, and opening and closing; the morphological operators are inspired by [44, 45] with adapted structuring elements paraments.

3) Localization module: Once the temporal saliency detects the moving target region and enhances the CMRs, a localization module is applied to localize the extracted CMRs based on connected components and blob identification methods [47]. This module involves the following steps.

Thresholding. Thresholding assists us in reducing the number of false positives and avoiding missed valid objects. The thresholding is based on the variation of intensity consideration between the object pixels and the background pixels, as inspired by [48]. Setting a determined value to identify those pixels to implement the thresholding. In this matter, THRESH_OTSU is used to determine the optimal threshold value using Otsu's algorithm [49].

Edge segmentation. Canny edge segmentation is then run on the binarized image for further improvement of the extracted region.

Blob Identification. After all the region's edges are extracted, we need to detect the blobs (connected components). To identify the blobs, active contour features, such as the one proposed by [50], are utilized to detect regions of interest and localize them. In this paper, we also use active contour features to detect the contours of regions. The detected contour features from CMRs regions are used for connected component detection, blob area, and bounding box determination. Moreover, removing unwanted blobs with a pixel area smaller than A_{low} or a bounding box with dimensions larger than B_{max}.

Candidate Mask Generation. In this step, geometrical features are extracted from the regions to recognize their location in each frame. Extracting X_{pos}, Y_{pos} as the centroid of

each object based on moment features, as described in the spatial saliency detection module section later. Furthermore, we experimentally found the appropriate width and height values to generate the candidate mask (CM), extracting the regions based on the region of interest (ROI) function. Fig. 2 shows the generated candidate mask by the proposed temporal saliency and localization stage.

4) Spatial saliency detection (SSD) stage: The result of the TSLD stage is one or multiple CM regions, which are ARs. However, as shown in Fig. 2, some regions are incorrectly extracted that are unrelated to a target object or include useless regions. The spatial saliency detection (SSD) is used to overcome this fault detection. Furthermore, because candidate masks are compact and informative, we also investigate SSD to extract the saliency over them to provide further information and generate sample features. Our proposed SSD algorithm is based on integrating the proposed methods in [2, 29] with several modifications in feature extraction, feature representation, and spatial distribution measurement to improve the efficiency of spatial saliency extraction.

In brief, the input image is first decomposed into perceptually homogeneous segments as patches based on a SLIC superpixel algorithm presented in [51]. Second, we extract visual features, including color and moment, to measure the uniqueness and compactness of the spatial distribution. Finally, the temporal and spatial information are integrated to generate a final saliency map named spatiotemporal saliency.

However, since the SLIC is time-consuming for spatial saliency, we implement spatial saliency detection via parallel processing based on multi-threading programming. The use of multi-threading assists us in processing all CM regions in parallel. It can impressively decrease the overall processing time of the SSD stage. In this regard, each thread captures a candidate mask and performs the following processes to determine the spatial saliency and sample feature generation. For instance, if the result of the TSLD stage includes four objects, we assign each object to a thread, totaling four. OpenMP multi-threading is used as a tool to implement our spatial saliency detection algorithm. The steps for the SSD stage are as follows.

5) Patch generation module: Super-pixels segmentation as an effective region-based analysis algorithm is increasingly investigated by many researchers in computer vision communities [36,52]. As proposed in [51], this study uses a SLIC algorithm to segment the CM regions into homogeneous regions. Fig. 3 shows patch generation for a moving object using the SLIC superpixel algorithm.

6) Spatial saliency module: In this study, we use spatial uniqueness and compactness to compute the spatial saliency detection inspired by Perazzi et al. [53]. Moreover, we take advantage of other features, such as image moments and different metrics, to improve efficiency. In our method, we investigate pixel intensity for dissimilarity measurement of a patch compared to other regions. Compactness spatial distribution also contributes to detecting salient objects based on image moments for uniqueness measurement. Details of the proposed spatial saliency detection are explained in the following.

Spatial uniqueness measurement. Similar to [54], each region's color similarity with other regions is measured. However, in [54], they implemented the color feature in a static image. In contrast, we investigate saliency detection in a dynamic environment. Furthermore, as reported in [55], Earth Mover's Distance (EMD) yielded excellent retrieval performance for the small sample size; we also use the EMD distance metric instead of Euclidean.

Spatial compactness measurement. Because the salient patches are spatially compact, the pixels with high saliency values are also expected to be spatially close [56]. Spatial moments are efficient and powerful in describing spatial distribution and compactness. In this study, we investigate spatial moments to estimate spatial compactness. Our work employs first- and second-order spatial moments.

7) Final saliency map generation: Generally, it is necessary to collaborate the temporal and spatial saliencies in a meaningful way to produce the final spatiotemporal saliency maps [10]. Therefore, the temporal and spatial information are integrated to generate a final saliency map named spatiotemporal saliency.

8) Sample generation and target detection (SGDT) stage: As shown in Fig. 1, the sample feature generation and target detection stage involve feature extraction, sample feature generation, and target detection. According to the feature extraction step and the result of the SSD stage, we collect appropriate features, such as color contrast and region compactness. As mentioned previously, these features are dynamically updated per frame and normalized, generating the sample features. Based on the sample features, we can detect the target.

B. Sample Feature-based Target Detection (SFTD) Phase

A trigger is activated upon the sample features being generated, and the sample features are transferred from the SSAM phase to this phase. The advantage of this phase is that it covers both moving and non-moving object detection conditions to detect objects with uncontrollable and unpredictable target movement conditions and overcome the difficulty of frame difference. The steps for this phase are mostly similar to the previous operation's steps, i.e., frame differencing, Image Enhancement, Feature Matching, Object Segmentation, and, finally, Target Detection.

IV. Results and Discussion

This section presents the implementation details and experimental results. Additionally, we compare the results with existing methods based on qualitative and quantitative performance evaluations to test and evaluate the proposed method. The qualitative analysis presents the image results from the proposed and other methods. In contrast, quantitative analysis involves precision and recall calculation and processing time. To validate the efficacy of the proposed method, the experiment was conducted on the VIVID public

dataset [57]. The VIVID dataset was collected at Eglin during DARPA VIVID and involves aerial images in video sequences. Several videos have been collected in VIVID, of which we use the EgTest01 and EgTest02 videos. The EgTest01 video involves moving cars that pass each other, with an image size of 640*480 pixels and 1800 frames, whereas the EgTest02 video involves 1300 frames with two sets of three civilian vehicles passing each other on a runway.

A. Qualitative Analysis

Qualitative analysis is implemented to demonstrate the result of each phase and compare the proposed method with others. Fig. 4 shows the results of the qualitative analysis. The saliency-based methods considered for comparison are Itti [58], MD [21], GBVS [59], and SD [2]. Fig. 5 shows the comparison of the proposed method with other existing methods. The first row is the original raw images (Raw), the second, third and fourth are the results for the TSLD, SSD, and MOED phases, respectively, and the final row represents the feature-based object detection phase.

In the following sections, quantitative analysis for precision, recall, and f-measure calculation is discussed.

B. Precision and Recall Measurement

Similar to Refs. [2, 31, 60], precision and recall measures are used to evaluate the performance of the proposed method. In our evaluation, the target is the exciting object, whether moving or not. To measure the precision, recall, and f-measure, we need to define the following terms,

- True Positive: Detected salient regions that correspond to a target.
- False Positive: Detected salient regions that do not correspond to a target.
- False Negative: No detection of salient regions where there is, in fact, a target,

$$\text{Precision} = \left(\sum_{i=1}^{n} \frac{\text{TP}}{\text{TP} + \text{FP}}\right) \times 100\ \%$$

$$\text{Recall} = \left(\sum_{i=1}^{n} \frac{\text{TP}}{\text{TP} + \text{FN}}\right) \times 100\ \%$$

$$\text{F}_1 - \text{score} = 2 \times \left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right)$$

Details of the measurements for TP, FP, FN, precision, recall, and F1-score are presented for the proposed method in Table I.

The precision and recall rates of different numbers of frames are illustrated in Fig. 6. As shown in Fig. 6 and 7, precision and recall rates are increased when the number of frames is increased.

Furthermore, to validate the proposed method, we compare our model with state-of-the-art visual object tracking methods, such as the FMD [2], DMM [60], HSC [10], RD [2], and SD [2]. The comparison was conducted based on precision, recall (PR), and F1-score. Table II and Fig. 8 show the comparison results. Based on the obtained experimental results, we show that the proposed approach can be effectively employed for the extraction of moving objects.

C. Processing Time

Our experiment was implemented in Visual Studio and performed on a Windows 8 platform with an Intel 2.6 GHz CPU and 4 GB of Memory. The processing time is measured based on wall-clock time computation because, when measuring the performance of parallel programs, the wall-clock time needs to be considered, then using the tick_count class, which is located in tbb/tick_count.h. A tick_count is an absolute timestamp. The average processing time for the proposed method is approximately 78 and 24 milliseconds for SSAM and SFTD, respectively, which is suitable for near-real-time visual tracking applications.

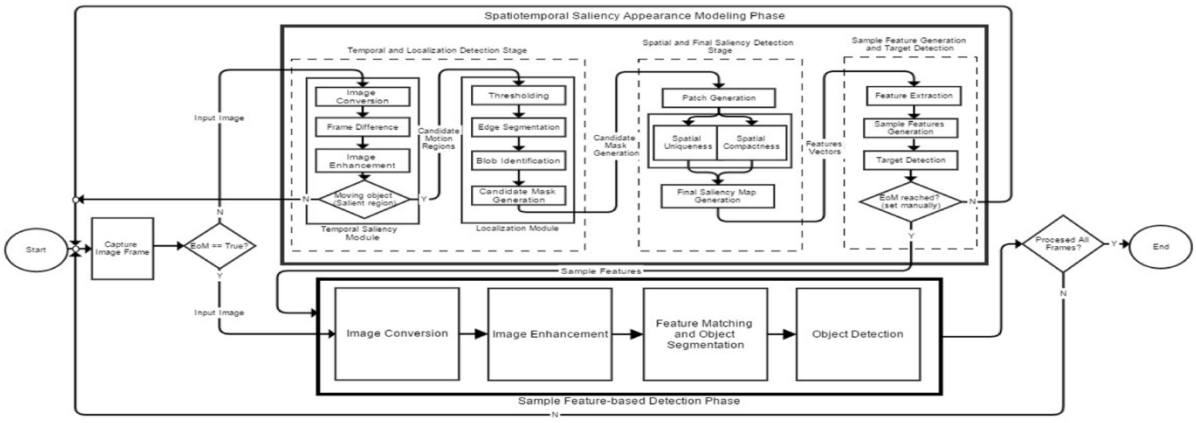

Fig. 1. Our proposed framework for visual tracking system

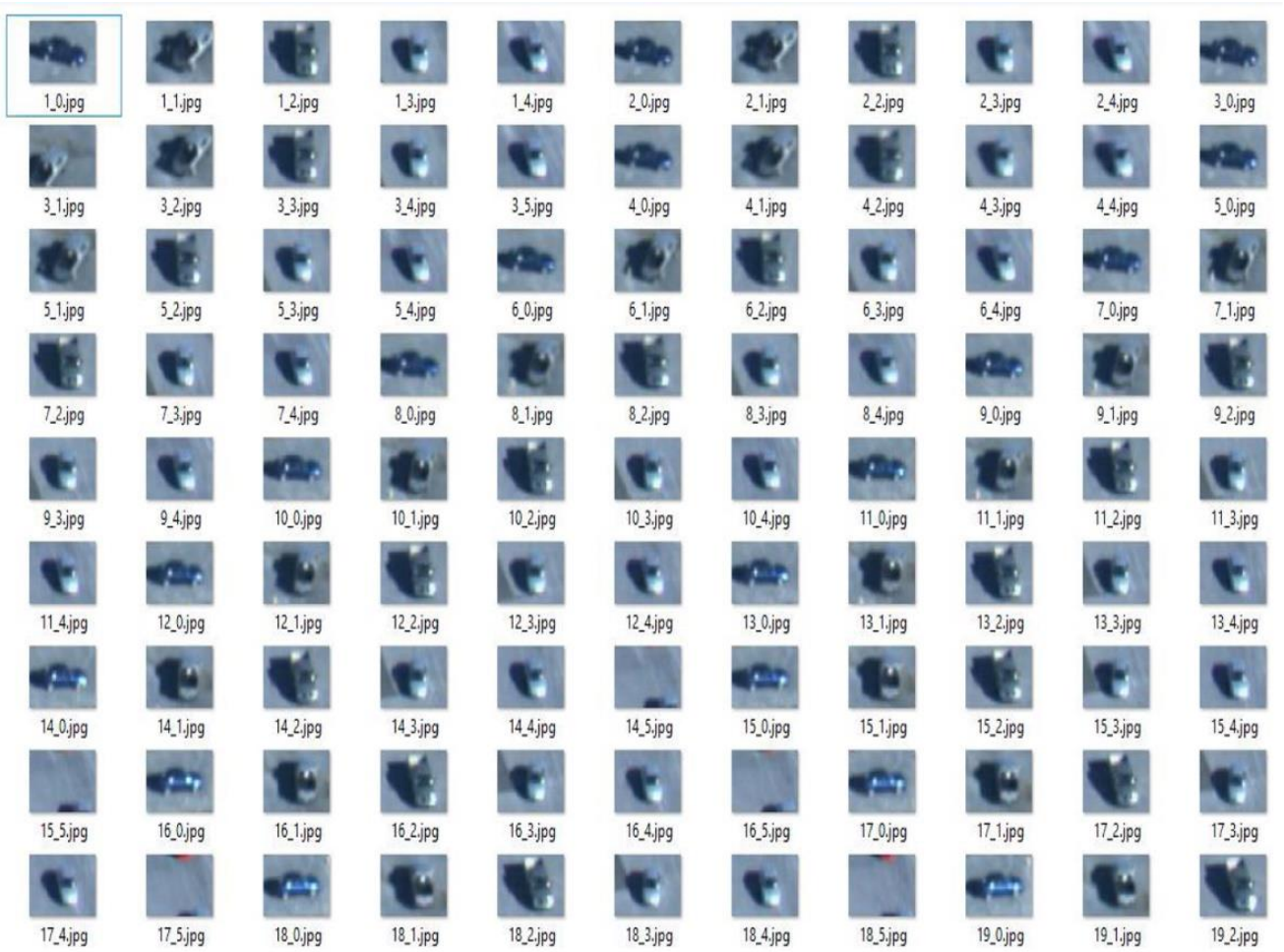

Fig. 2. Candidate mask generation images

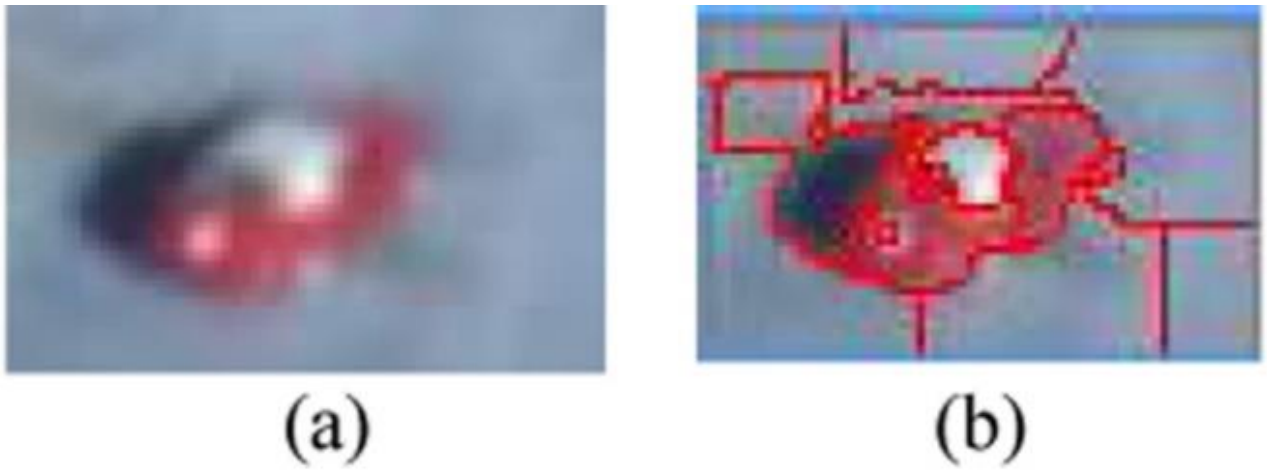

Fig. 3. Patch generation for a moving object using a parallel SLIC superpixel algorithm. (a) An original candidate mask, (b) generated patches

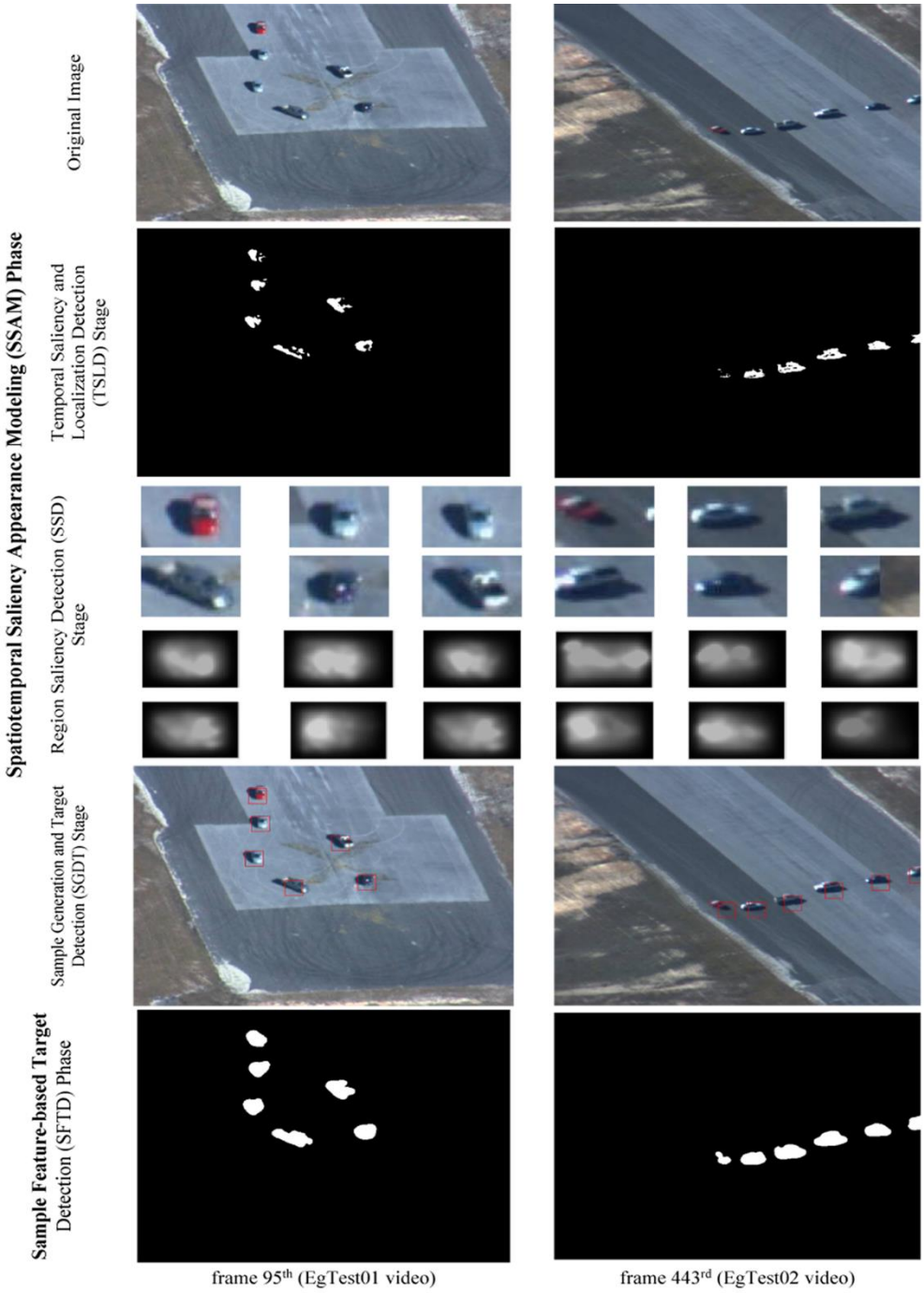

Fig. 4. Image results for each phase and stage of the proposed method

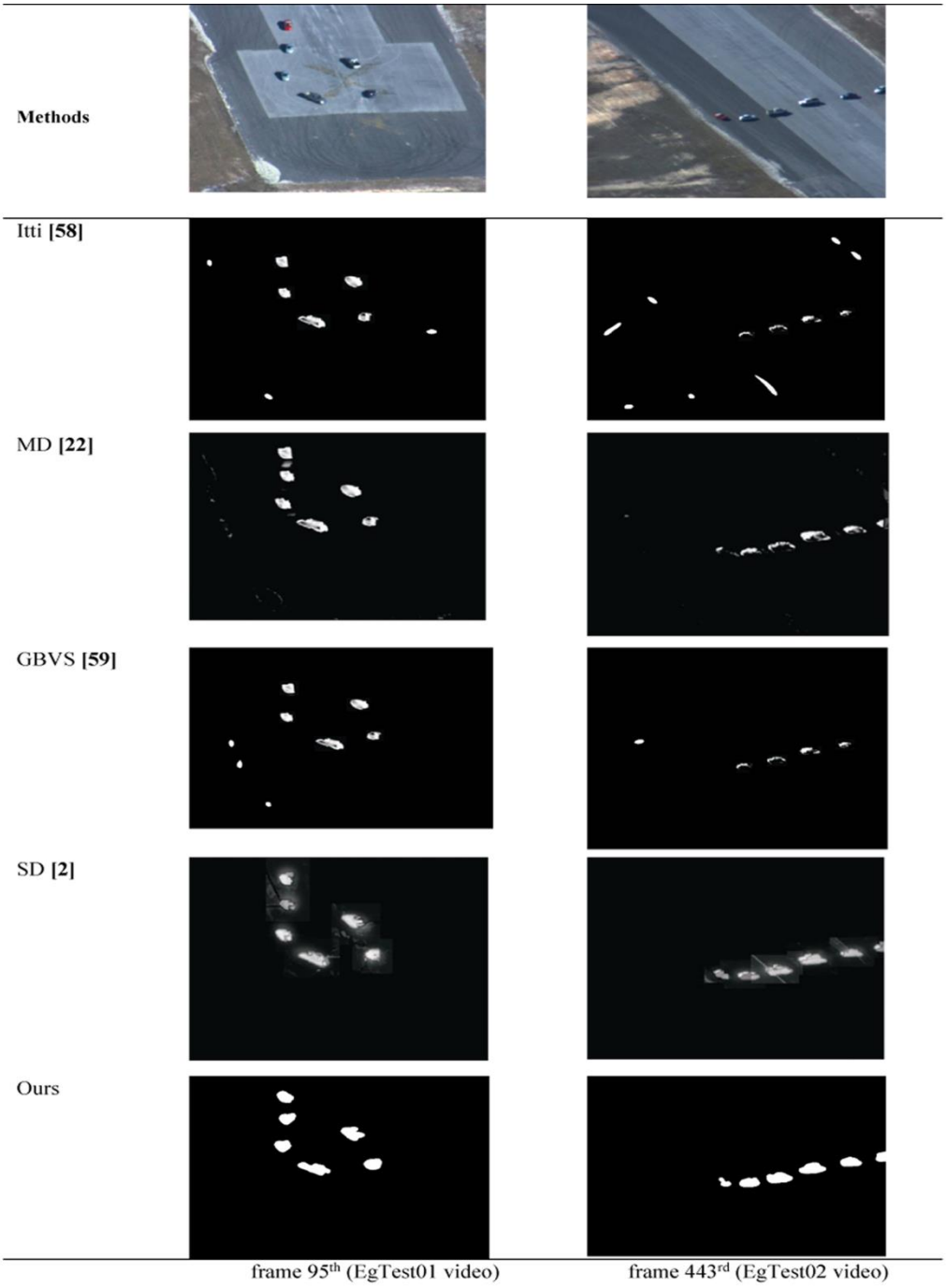

Fig. 5. Comparison of the proposed method with four state-of-the-art saliency methods

TABLE I. DETAILS OF THE MEASUREMENTS OF THE TRUE POSITIVE, FALSE POSITIVE, FALSE NEGATIVE, PRECISION, AND RECALL RATES

Dataset	Number of Frames	TP	FP	FN	Precision (%)	Recall (%)
EgTest01	50	32	12	6	0.73	0.84
	450	307	97	46	0.76	0.87
	1150	846	214	93	0.80	0.90
	1800	1381	298	121	0.82	0.92

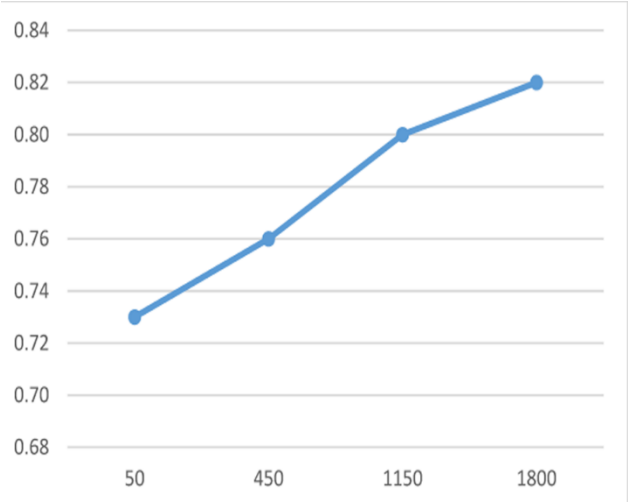

Fig. 6. Precision metric comparison for different numbers of frames

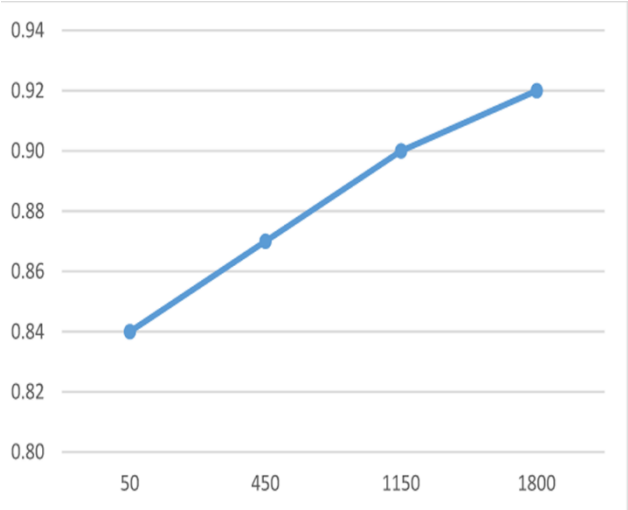

Fig. 7. Comparison of different numbers of frames based on the recall measures

TABLE II. COMPARISON OF VISUAL TRACKING METHODS AND THE PROPOSED METHOD

Method	Recall (%)	Precision (%)	F_1-score (%)
FMD	0.49	0.34	0.40
DMM	0.68	0.48	0.56
HSC	0.69	0.51	0.59
RD	0.74	0.69	0.71
SD	0.79	0.48	0.60
Ours	0.82	0.73	0.77

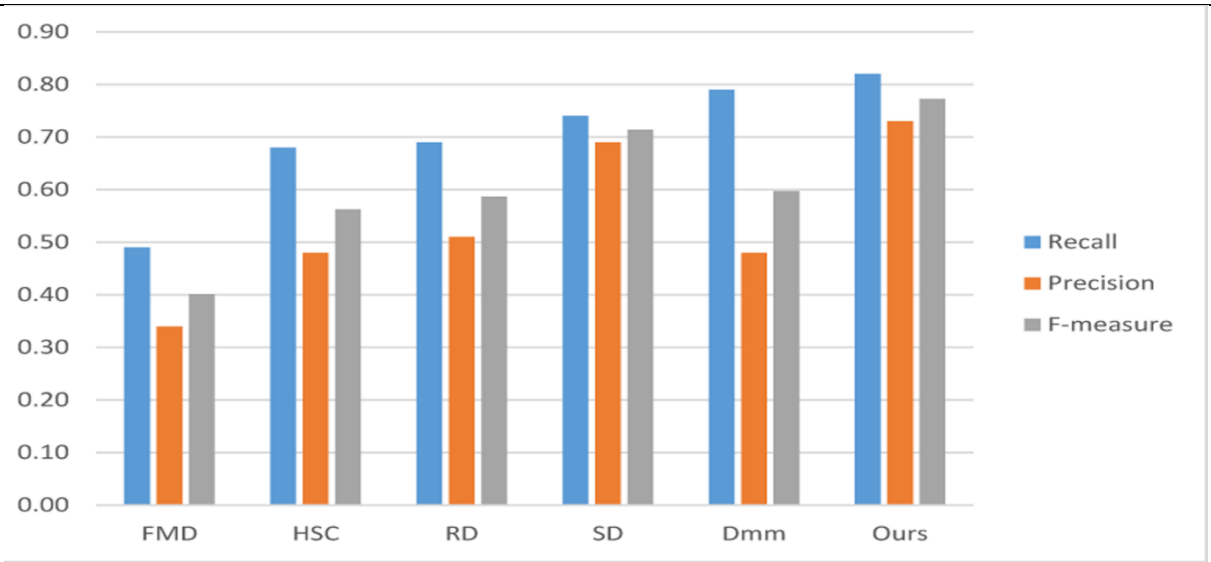

Fig. 8. Precision, recall and F-measure for visual tracking methods and our method

V. CONCLUSION

This paper addresses the significant problems facing visual tracking, such as appearance variations and unpredictable moving targets, for aerial images. The proposed method uses spatial and temporal saliencies to address these challenges by adopting biologically inspired approaches to detect the attentional regions (ARs). Furthermore, a biologically inspired approach integrated with an appearance modeling-based approach is investigated to overcome visual tracking challenges. In this regard, the proposed method consists of two main phases, spatiotemporal saliency-based appearance modeling (SSAM) and sample feature-based target detection (SFTD). The proposed method uses a tracking-by-detection approach to provide a robust visual tracking system under appearance variation conditions. Correspondingly, a semi-automatic trigger-based algorithm is proposed to handle the phases' operation, and a discriminative-based method is utilized for appearance modeling. In the spatiotemporal saliency phase, temporal saliency is used to extract the attentional regions (ARs) and coarse segmentation. Spatial saliency is utilized to obtain the object's appearance details in ARs regions. By combining temporal and spatial saliencies, we can obtain refined detection results and track the target. During the spatial saliency detection, prominent features are collected, and a sample feature is generated to describe the target.

Consequently, a target detection process is performed to recognize the target in images. Experiments were conducted on the VIVID dataset. Moreover, the proposed method compared with other state-of-the-art methods. The analyses demonstrate that the proposed method is superior to most state-of-the-art methods and presents an effective visual tracking method which is robust in appearance variation difficulties.

Future works can be conducted to address other difficulties and challenges in visual tracking, such as when complicated backgrounds or backgrounds with partial and/or full occlusion are present.

ACKNOWLEDGMENT

The authors would like to appreciate Assoc. Prof. Dr. Anton Satria Prabuwono, Dr. Ang Mei Choo, and Teck Loon Lim for helpful advice and suggestions. We also thank all the anonymous reviewers for their comments which assisted us in improving the quality of this paper.

REFERENCES

[1] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang (2014), Fast Visual Tracking via Dense Spatio-temporal Context Learning, In: Computer Vision–ECCV 2014: Springer, pp. 127-141.

[2] H. Shen, S. Li, C. Zhu, H. Chang, and J. Zhang (2013), Moving object detection in aerial video based on spatiotemporal saliency, Chinese Journal of Aeronautics, vol. 26, pp. 1211-1217.

[3] F. Chen, Q. Wang, S. Wang, W. Zhang, and W. Xu (2011), Object tracking via appearance modeling and sparse representation, Image and Vision Computing, vol. 29, pp. 787-796.

[4] S. Zhang, H. Yao, H. Zhou, X. Sun, and S. Liu (2013), Robust visual tracking based on online learning sparse representation, Neurocomputing, vol. 100, pp. 31-40.

[5] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song (2011), Recent advances and trends in visual tracking: A review, Neurocomputing, vol. 74, pp. 3823-3831.

[6] M. Yang, J. Yuan, and Y. Wu, Spatial selection for attentional visual tracking, in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007, pp. 1-8.

[7] C. Siagian and L. Itti (2007), Rapid biologically-inspired scene classification using features shared with visual attention, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 300-312.

[8] Y. Kashiwase, K. Matsumiya, I. Kuriki, and S. Shioiri (2013), Temporal Dynamics of Visual Attention Measured with Event-Related Potentials, PloS one, vol. 8, p. e70922.

[9] Y. Zhai and M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in Proceedings of the 14th annual ACM international conference on Multimedia, 2006, pp. 815-824.

[10] C. Li, J. Xue, N. Zheng, X. Lan, and Z. Tian (2013), Spatio-temporal saliency perception via hypercomplex frequency spectral contrast, Sensors, vol. 13, pp. 3409-3431.

[11] L. Itti, Models of bottom-up and top-down visual attention, California Institute of Technology, 2000.

[12] F. Göschl, A. K. Engel, and U. Friese (2014), Attention Modulates Visual-Tactile Interaction in Spatial Pattern Matching, PloS one, vol. 9, p. e106896.

[13] V. Mahadevan and N. Vasconcelos (2013), Biologically Inspired Object Tracking Using Center-Surround Saliency Mechanisms, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, pp. 541-554.

[14] A. Borji and L. Itti (2013), State-of-the-art in visual attention modeling, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, pp. 185-207.

[15] W. Kim, C. Jung, and C. Kim (2011), Spatiotemporal saliency detection and its applications in static and dynamic scenes, Circuits and Systems for Video Technology, IEEE Transactions on, vol. 21, pp. 446-456.

[16] D. Kerzel, J. Schönhammer, N. Burra, S. Born, and D. Souto (2011), Saliency changes appearance, PloS one, vol. 6, p. e28292.

[17] Y. Zhang, Z. Mao, J. Li, and Q. Tian (2014), Salient Region Detection for Complex Background Images Using Integrated Features, Information Sciences,

[18] H. R. Tavakoli, E. Rahtu, J. Heikkil, and #228 (2013), Temporal saliency for fast motion detection, presented at the Proceedings of the 11th international conference on Computer Vision - Volume Part I, Daejeon, Korea.

[19] T. Crivelli, P. Bouthemy, B. Cernuschi-Frías, and J.-f. Yao (2011), Simultaneous motion detection and background reconstruction with a conditional mixed-state markov random field, International journal of computer vision, vol. 94, pp. 295-316.

[20] O. Barnich and M. Van Droogenbroeck (2011), ViBe: A universal background subtraction algorithm for video sequences, Image Processing, IEEE Transactions on, vol. 20, pp. 1709-1724.

[21] Z. Yin and R. Collins, Moving object localization in thermal imagery by forward-backward MHI, in Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, 2006, pp. 133-133.

[22] C. Benedek, T. Szirányi, Z. Kato, and J. Zerubia (2009), Detection of object motion regions in aerial image pairs with a multilayer markovian model, Image Processing, IEEE Transactions on, vol. 18, pp. 2303-2315.

[23] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia (2001), Event detection and analysis from video streams, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 873-889.

[24] K. K. Ng and E. J. Delp, Background subtraction using a pixel-wise adaptive learning rate for object tracking initialization, in IS&T/SPIE Electronic Imaging, 2011, pp. 78820I-78820I-9.

[25] K. K. Ng and E. J. Delp, Object tracking initialization using automatic moving object detection, in IS&T/SPIE Electronic Imaging, 2010, pp. 75430M-75430M-12.

[26] H. J. Min, Multi-Robot Formation and Cooperation Using Visual Tracking, 2013.

[27] X. Shen and Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 853-860.

[28] E. Erdem and A. Erdem (2013), Visual saliency estimation by nonlinearly integrating features using region covariances, Journal of vision, vol. 13, p. 11.

[29] W. Wang, D. Cai, X. Xu, and A. Wee-Chung Liew (2014), Visual saliency detection based on region descriptors and prior knowledge, Signal Processing: Image Communication, vol. 29, pp. 424-433.

[30] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, Global contrast based salient region detection, in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 409-416.

[31] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, Frequency-tuned salient region detection, in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1597-1604.

[32] J. Harel, C. Koch, and P. Perona, Graph-based visual saliency, in Advances in neural information processing systems, 2006, pp. 545-552.

[33] K.-C. Lee and D. Kriegman, Online learning of probabilistic appearance manifolds for video-based recognition and tracking, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 852-859.

[34] Y. Wu, J. Cheng, J. Wang, H. Lu, J. Wang, H. Ling, et al. (2012), Real-time probabilistic covariance tracking with efficient model update, IEEE Transactions on Image Processing, vol. 21, pp. 2824-2837.

[35] T. Bai and Y. F. Li (2012), Robust visual tracking with structured sparse representation appearance model, Pattern Recognition, vol. 45, pp. 2390-2404.

[36] B. Zhong, Y. Chen, Y. Shen, Y. Chen, Z. Cui, R. Ji, et al. (2014), Robust tracking via patch-based appearance model and local background estimation, Neurocomputing, vol. 123, pp. 344-353.

[37] X. Zhang, W. Hu, S. Maybank, and X. Li, Graph based discriminative learning for robust and efficient object tracking, in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1-8.

[38] J. Fan, Y. Wu, and S. Dai (2010), Discriminative spatial attention for robust tracking, In: Computer Vision–ECCV 2010: Springer, pp. 480-493.

[39] M. Tang and X. Peng (2012), Robust tracking with discriminative ranking lists, IEEE Transactions on Image Processing, vol. 21, pp. 3273-3281.

[40] C. Huang, Q. Liu, and S. Yu (2011), Regions of interest extraction from color image based on visual saliency, The Journal of Supercomputing, vol. 58, pp. 20-33.

[41] W. Chen, Y. Q. Shi, and G. Xuan, Identifying computer graphics using HSV color model and statistical moments of characteristic functions, in Multimedia and Expo, 2007 IEEE International Conference on, 2007, pp. 1123-1126.

[42] R. C. Gonzalez and R. E. Woods, Digital image processing, ed: Prentice hall Upper Saddle River, NJ:, 2002.

[43] K. Sreedhar and B. Panlal (2012), Enhancement of images using morphological transformation, arXiv preprint arXiv:1203.2514,

[44] E. R. Dougherty, R. A. Lotufo, and T. I. S. f. O. E. SPIE, Hands-on morphological image processing vol. 71: SPIE press Bellingham, 2003.

[45] J. Serra (1986), Introduction to mathematical morphology, Computer vision, graphics, and image processing, vol. 35, pp. 283-305.

[46] X. Bai, F. Zhou, and B. Xue (2012), Image enhancement using multi scale image features extracted by top-hat transform, Optics & Laser Technology, vol. 44, pp. 328-336.

[47] A. Das, M. Diu, N. Mathew, C. Scharfenberger, J. Servos, A. Wong, et al. (2014), Mapping, Planning, and Sample Detection Strategies for Autonomous Exploration, Journal of Field Robotics, vol. 31, pp. 75-106.

[48] W. OpenCV. (2014). Basic Thresholding Operations. Available: http://docs.opencv.org/3.0-alpha/doc/py_tutorials/py_imgproc/py_thresholding/py_thresholding.html?highlight=adaptive%20thresholding

[49] W. OpenCV. (2014). Miscellaneous Image Transformations. Available: http://docs.opencv.org/modules/imgproc/doc/miscellaneous_transformations.html?highlight=threshold#threshold

[50] S. Sclaroff and J. Isidoro (2003), Active blobs: region-based, deformable appearance models, Computer Vision and Image Understanding, vol. 89, pp. 197-225.

[51] C. Y. Ren and I. Reid (2011), gSLIC: a real-time implementation of SLIC superpixel segmentation, University of Oxford, Department of Engineering, Technical Report,

[52] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk (2012), SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 2274-2282.

[53] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 733-740.

[54] L. Shuhua, L. Zhi, L. Lina, Z. Xuemei, and O. Le Meur, Efficient saliency detection using regional color and spatial information, in Visual Information Processing (EUVIP), 2013 4th European Workshop on, 2013, pp. 184-189.

[55] Y. Rubner, C. Tomasi, and L. J. Guibas (2000), The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision, vol. 40, pp. 99-121.

[56] Z. Chi and W. Weiqiang, Object-level saliency detection based on spatial compactness assumption, in Image Processing (ICIP), 2013 20th IEEE International Conference on, 2013, pp. 2475-2479.

[57] S. o. C. S. Robotics Institute, Carnegie Mellon University. . (2013). VIVID Tracking Evaluation Web Site. Available: http://vision.cse.psu.edu/data/vividEval/datasets/PETS2005/EgTest01/index.html

[58] L. Itti, C. Koch, and E. Niebur (1998), A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 1254-1259.

[59] J. Harel, C. Koch, and P. Perona, Graph-based visual saliency, in Advances in neural information processing systems, 2007, pp. 545-552.

[60] F. M. S. Saif, A. S. Prabuwono, and Z. R. Mahayuddin (2014), Moving Object Detection Using Dynamic Motion Modelling from UAV Aerial Images, The Scientific World Journal, vol. 2014, p. 12.