SPECTRAL-SPATIAL-TEMPORAL ATTENTION NETWORK FOR HYPERSPECTRAL TRACKING

Zhuanfeng Li[1] *Xinhai Ye*[1] *Fengchao Xiong*[1] *Jianfeng Lu*[1] *Jun Zhou*[2] *Yuntao Qian*[3]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] School of Information and Communication Technology, Griffith University, Australia
[3] College of Computer Science, Zhejiang University, China

ABSTRACT

Thanks to the abundant spectral bands, hyperspectral videos (HSVs) are able to describe objects at material level, i.e., the physical property, providing more benefits for object tracking than color videos. Considering limited HSV dataset for training, a band attention aware ensemble network was recently proposed for hyperspectral tracking, which leverages band attention to select several three-channel images for deep hyperspectral tracking. However, it fails to fully consider the joint spectral-spatial-temporal information in HSVs, compromising its tracking performance in challenging scenarios. To this end, we introduce a spectral-spatial-temporal attention neural network (SST-Net) for hyperspectral tracking in this paper. Specifically, the spatial attention with convolution and deconvolution structure focuses on the salient spatial features. Moreover, the temporal attention with an RNN structure is adopted to depict the temporal relationship among adjacent frames. By combining the spatial, spectral, and temporal attention, the band relationship can be better depicted thus valuable hyperspectral bands can be better selected for deep ensemble tracking. Experimental results show the improved effectiveness of SST-Net in tracking over serval alternative trackers.

***Index Terms*—** deep learning, hyperspectral tracking, spectral-spatial-temporal attention

1. INTRODUCTION

Object tracking plays an important role in computer vision applications and has recently made great progress ranging from manual feature-based to deep feature-based [1–3]. However, these object tracking algorithms implemented in color videos have inherent drawbacks in describing the real physical properties of the object. This prevents them from distinguishing between targets and backgrounds that are similar in appearance or texture, resulting in limited tracking performance.

In contrast, hyperspectral videos (HSVs) can alleviate the above problem thanks to their material identification ability enabled by abundant spectral bands. The existing hyperspectral tracking algorithms mainly concentrate on the visual appearance representation of the target. For example, Qian *et al.* [4] extracted 3D local cubes around the object as the convolution kernels for feature extraction. Xiong *et al.* [5] embedded the spectral-spatial structure of hyperspectral images (HSIs) into a traditional histogram of oriented gradients (HOG) which was then combined with global material abundance features to describe the object. The above methods use handcrafted features, which have limited representation ability. In contrast, Uzkent *et al.* [6] converted HSI into three-channel data to pass the VGGNet network [7] for deep feature extraction. Unfortunately, the converted three-channel data unavoidably lose much valuable spectral information. Alternatively, multiple sets of three-channel data can be selected. But how to select these bands is an important issue. Motivated by the ranking based band selection, Li *et al.* [8] regards the importance of each band as a criterion and proposed an autoencoder-like band attention mechanism network (BAE-Net) to learn the nonlinear spectral relationship to better convert HSI into a number of false-color images. These converted images were then passed through several VITAL [9] trackers, yielding several weak trackers for subsequent ensemble learning to obtain the target location.

BAE-Net only considers spectral information of the current frame but ignores valuable spatial and temporal information of HSVs, yielding unstable tracking. In fact, the spatial and temporal information can be used as a supplement to spectral information to form a stronger object representation feature. On the one hand, the spatial information helps to mine the positional relationship of HSI and improves the ability of band selection. For example, Sun *et.al* [10] leveraged manifold learning to preserve the geometric relationship of HSIs in low-dimensional subspace for enhanced band selection. On the other hand, spatial information provides the

This work was supported in part by “111” Program (No. B13022), Jiangsu Provincial Natural Science Foundation of China under Grant BK20200466, the National Key Research and Development Program of China under Grant 2018YFB0505000 and the National Natural Science Foundation of China under Grant 62002169 and 62071421.

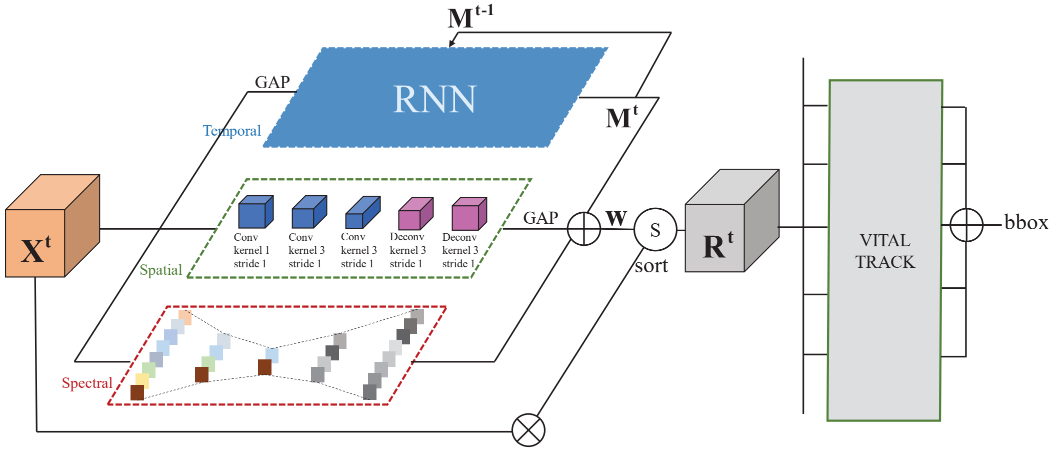


Fig. 1. The architecture of SST-Net. The HSI $\mathcal{X}^t$ at t-th frame are firstly passed into the spectral attention, spatial attention and temporal attention module to obtain the spectral relationship $\mathbf{w}$. According to the relationship, the bands are then ranked and rearranged to yield $\mathcal{R}^t$. Subsequently,$\mathcal{R}^t$ is divided into a number of three-channel false-color images which are then passed into VITAL tracker to yield a number of weak trackers. The state of yielded weak trackers are finally summarized by ensemble learning to produce the object location.

interrelationship between the parts of the current frame and is proven beneficial for many computer vision tasks, e.g, object tracking [11] and person reidentification [12].

Besides spatial and spectral information, HSVs also provide temporal information. Temporal information depicts interrelationships among adjacent frames and offers very conducive clues for accurate object localization in many scene [13]. By learning to fuse useful information over time for converting HSI into a group of false-color images, the appearance of the object can be more robustly represented. This facilitates overcoming drifts compared with only using the spatial and spectral information in a single frame.

To this end, in this paper, we introduce an end-to-end spectral-spatial-temporal attention network named as SST-Net to model the band relationship so as to improve the appearance representation in HSV and achieve robust tracking. As shown in Fig. 1, SST-Net contains three modules including spectral attention module as in [8], additional spatial attention, and temporal attention module. The spatial attention takes advantage of convolution and deconvolution operators to exploit the inter-spatial relationship, enabling to use of the most valuable object regions for band converting. Furthermore, an RNN-like architecture is adopted to capture inter-temporal correspondence and motion changes among adjacent frames. Such spectral-spatial-temporal attention cooperates with each other so that the band relationship can be better modeled, facilitating extracting the most informative hyperspectral features for tracking. Experimental results show that SST-Net surpasses BAE-Net to a large margin and achieves state-of-the-art tracking performance.

2. PROPOSED SST-NET

In this section, we will describe the details of SST-Net, including BAE-Net, spatial attention and temporal attention modules.

2.1. BAE-Net: Spectral Attention Module

Our method is based on the framework of BAE-Net. BAE-Net takes advantage of spectral attention to rank the bands so that bands with higher importance are grouped to formulate three-channel images for feature extraction. The spectral attention is obtained by a sequence of two convolutional operations and two transposed convolutional operations. Let $\mathbf{x_n} \in N^{1\times C}$ be the pixel of the HSI, where C represents the number of hyperspectral bands. The spectral attention module can be formulated as:

$$\mathbf{w}_{\text{spectral}} = s(f_p(\sigma_2(\sigma_1(\mathbf{x}_n\mathbf{A}_1)\mathbf{A}_2\mathbf{A}_2^T)\mathbf{A}_1^T)) \qquad (1)$$

where σ represents the ReLU activation function, $\mathbf{A}_1$ and $\mathbf{A}_2$ represent the parameters of 1×1 convolution operation, likewise $\mathbf{A}_1^T$ and $\mathbf{A}_2^T$, f_p is the global average pooling operation, and $s(\bullet)$ represents the softmax normalization.

However, BAE-Net only takes spectral attention into consideration but ignores the valuable spatial information and temporal information. The ignorance of spatial and temporal information makes BAE-Net prone to many complicated scenes such as occlusion and illumination changes. Therefore, we add spatial attention and temporal attention to BAE-Net and then construct the spectral-spatial-temporal attention network shown in Fig. 1 to better convert HSI into a group of three-channel images for deep hyperspectral tracking.

2.2. Spatial Attention Module

The spatial interference from the surrounding background is not uncommon in object detection. Spatial attention aims to assign different weights to different spatial locations so that the network focuses on the salient region and suppresses learning from annoying backgrounds. As shown in Fig. 1, three convolution operations and two deconvolution operations followed by a ReLU activation function are first applied along the spatial axis to encode the inter-spatial relationship of HSI. After that, an average-pooling operation along the channel axis is used to summarize the relationship across spatial, producing the refined attention map for band converting. In short, the spatial attention can be mathematically expressed as:

$$\mathbf{w}_{\text{spatial}} = f_p(\text{dconv}_2(\text{dconv}_1(\text{conv}_3(\text{conv}_2(\text{conv}_1(\mathcal{X})))))) \tag{2}$$

where conv and dconv respectively represent convolution and deconvolution operations with a ReLU activation function, and f_p represents the global average pooling.

2.3. Temporal Attention Module

In addition to spatial attention and spectral attention, we also integrate the motion changes information into the network by a temporal attention module. As in [14–17], we also adopt an RNN computing unit to model long-term temporal appearance and motion dynamics among adjacent frames. Specifically, we firstly apply a global average pooling layer on $\mathcal{X}^t$ to yield $\mathbf{x}_t$. After that $\mathbf{x}_t$ is passed into an RNN architecture to produce the temporal attention $\mathbf{w}_{\text{temporal}}$. The RNN architecture can be mathematically represented by

$$\begin{aligned}
\mathbf{m}_t &= \sigma(B_m * \mathbf{x}_t + A_m * \mathbf{m}_{t-1}) \\
\mathbf{n}_t &= \sigma(B_n * \mathbf{x}_t + A_n * \mathbf{m}_{t-1}) \\
\widehat{\mathbf{m}}_t &= \sigma(B * \mathbf{x}_t + A * (\mathbf{m}_{t-1} \odot \mathbf{n}_t)) \\
\mathbf{m}_t &= (1 - \mathbf{m}_t) \odot \mathbf{m}_{t-1} + \mathbf{m}_t \odot \widehat{\mathbf{m}}_t
\end{aligned} \tag{3}$$

where A, B, A_m and B_m are fully connected layers whose parameters can be obtained by end-to-end training, and $\odot$ is element-wise multiplication.

Here, $\mathbf{n}_t$ gate masks the previous memory $\mathbf{m}_{t-1}$ to allow the previous state to be forgotten or not which is then merged with the input at current frame to produce a candidate memory $\widehat{\mathbf{m}}_t$. $\mathbf{m}_t$ is a gate corresponding to memory and determines how to combine historical information $\mathbf{m}_{t-1}$ with current frame $\widehat{\mathbf{m}}_t$ to generate a new memory.

2.4. Ensemble Tracking

The spectral weight matrix generated by spatial attention, spectral attention and temporal attention modules are then averaged to produce the importance of bands $\mathbf{w}$. After that the HSI is divided into a number of three-channel false-color images according to $\mathbf{w}$ to pass into state-of-the-art tracker VITAL, generating a set of weak trackers. The determined location is obtained by averagely weighting the locations produced by the weak trackers. Therefore, the loss function for SST-Net is defined by

$$\mathcal{L} = \frac{1}{\lfloor L/3 \rfloor} \sum_{i=1}^{\lfloor L/3 \rfloor} \mathcal{L}_i \tag{4}$$

where $\lfloor L/3 \rfloor$ indicates the number of groups, $\mathcal{L}_i$ represents the VITAL loss function from i-th weak tracker.

3. EXPERIMENTS

We compared SST-Net with the baseline BAE-Net, state-of-the-art color video trackers and hyperspectral trackers to show the advantages of our method. Moreover, an ablation study is also conducted to demonstrate the effectiveness of spatial-temporal-spectral attention network in hyperspectral object tracking.

3.1. Experiment Settings

All the experiments were conducted on the dataset provided by hyperspectral object tracking competition [1]. The dataset contains 40 videos for training and 35 videos for testing each of which includes hyperspectral, false-color and RGB videos under the same scene. The spectral attention and spatial attention module were pre-trained offline on the training set. The temporal attention is trained on the first frame when tracking. If the score is less than a given threshold, obtained through extensive cross-validation experiments, all the three attention modules are adjusted to adapt to the scenario changes. The learning rate for the SST-Net model was set to 0.005. All the competing trackers are evaluated using the precision plot, success plot, and area under the curve (AUC) [18] of one pass evaluation (OPE).

Table 1. Ablation Study of SST-Net. **Red** and **blue** mark the top two values.

Method	AUC
baseline	0.6061
baseline+spatial	**0.6190**
baseline+temporal	0.6101
baseline+spatial+temporal	**0.6230**

3.2. Ablation Study

Here, we perform an ablation study to show the effectiveness of proposed components by evaluating four variants of our approach on hyperspectral videos. The baseline is implemented

[1] https://www.hsitracking.com/

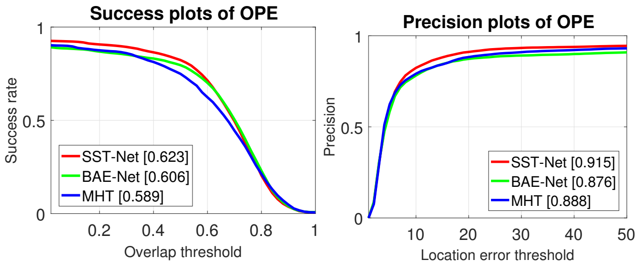
Table 2. AUC comparison with state-of-the-art color trackers. Red and blue mark the top two values.

Video	SST-Net	BACF [19]	fDSST [20]	KCF [21]	VITAL [9]	C-COT [20]	CFNet [3]
Color	n/a	0.5315	0.4639	0.3769	0.5759	**0.6020**	**0.5596**
Hyperspectral/False-color	**0.6230**	0.5440	0.4416	0.4078	**0.6047**	0.5572	0.5426

with only spectral attention. In order to ensure fairness, all the parameters are set to the same for four cases. Table 1 presents the tracking accuracy with respect to the AUC score. Thanks to the merits of suppressing the interference from uninformative regions and promote learning from important regions, the combination of spatial attention and spectral attention surpasses baseline to a large margin. The temporal attention is able to capture the motion changes among adjacent frames, yielding a better AUC than baseline. Moreover, simultaneous consideration of spatial, spectral and temporal attention allows SST-Net to achieve the best tracking performance among all the cases by providing an AUC of **0.6230**. Overall, this study evidently demonstrates the contribution of the proposed modules.

3.3. Comparison with State-of-the-art Color Trackers

We further compare proposed SST-Net with some recent state-of-the-art color trackers, including BAFC [19], KCF [21], fDSST [20], VITAL [9], C-COT [22], CFNet [3]. All the color trackers are run on both false-color and color videos. The results of the comparison are shown in Table 2. The BACF, fDSST and KCF fall behind because they use handcrafted features for tracking, limiting their robust feature representation. Thanks to powerful representation ability, VITAL, C-COT and CFNet provide better tracking results. The proposed SST-Net method achieves the highest AUC because of the comprehensive consideration of spectral-spatial-temporal information.



(a) Success plot (b) Precision plot

Fig. 2. Comparison with hyperspectral trackers.

3.4. Comparison with Hyperspectral Trackers

We further compare SST-Net against two hyperspectral trackers including BAE-Net [8], and MHT [18] to thoroughly demonstrate the advantages of proposed SST-Net. Fig. 2 shows the comparative results of these trackers. Compared with the MHT method, the BAE-Net method based on data-driven deep feature has a more robust feature representation. So, BAE-Net obtains higher tracking accuracy than MHT. But unlike the BAE-Net, which only considers spectral information, the SST-Net simultaneously takes the spectral-spatial-temporal information to convert HSI and can more effectively depict the object, yielding the best tracking performance. Fig. 3 visualizes the tracking results on forest2, pedestrian2 and playground sequences. As can be seen, the proposed SST-Net also provides the best visual results, implying the effectiveness of spatial and temporal modules.

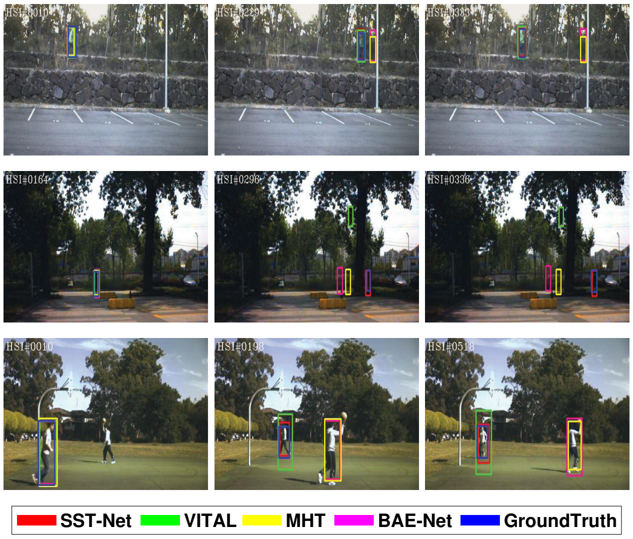


Fig. 3. Demonstrations of visual tracking results.

4. CONCLUSION

Limited training samples make it difficult to train a deep model for hyperspectral tracking. In order to tackle this problem, we propose a spectral-spatial-temporal attention network that takes advantage of available feature extraction based on three-channel color images for training. The spatial attention module makes the network focus more on the salient object. The temporal attention module models motion changes over time. These two modules are then combined with spectral attention to depicting the relationship among bands so that the HSI can be better converted for deep feature extraction. The experimental result shows that our proposed SST-Net algorithm surpasses the baseline BAE-Net and other hyperspectral and color trackers, reflecting the effectiveness of SST-Net in hyperspectral object tracking.

5. REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr, "Fully-convolutional siamese networks for object tracking," in *ECCV*, 2016, pp. 850–865.

[2] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang, "Learning dynamic siamese network for visual object tracking," in *IEEE ICCV*, 2017.

[3] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip HS Torr, "End-to-end representation learning for correlation filter based tracking," in *IEEE CVPR*, 2017.

[4] Kun Qian, Jun Zhou, Fengchao Xiong, Huixin Zhou, and Juan Du, "Object tracking in hyperspectral videos with convolutional features and kernelized correlation filter," in *ICSM*, 2018.

[5] Fengchao Xiong, Jun Zhou, Jocelyn Chanussot, and Yuntao Qian, "Dynamic material-aware object tracking in hyperspectral videos," in *IEEE WHISPERS*, 2019.

[6] Burak Uzkent, Aneesh Rangnekar, and Matthew J Hoffman, "Tracking in aerial hyperspectral videos using deep kernelized correlation filters," *IEEE TGRS*, vol. 57, no. 1, pp. 449–461, 2018.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *IEEE CVPR*, 2014.

[8] Zhuanfeng Li, Fengchao Xiong, Jun Zhou, Jing Wang, Jianfeng Lu, and Yuntao Qian, "BAE-Net: A band attention aware ensemble network for hyperspectral object tracking," in *IEEE ICIP*, 2020, pp. 2106–2110.

[9] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang, "VITAL: Visual tracking via adversarial learning," in *IEEE CVPR*, 2018.

[10] W. Sun and Q. Du, "Graph-regularized fast and robust principal component analysis for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3185–3195, 2018.

[11] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11920–11929.

[12] Haoran Wang, Yue Fan, Zexin Wang, Licheng Jiao, and Bernt Schiele, "Parameter-free spatial attention network for person re-identification," *IEEE CVPR*, 2018.

[13] Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan, "End-to-end flow correlation tracking with spatial-temporal attention," in *IEEE CVPR*, 2018.

[14] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei, "Towards high performance video object detection," in *IEEE CVPR*, 2018.

[15] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei, "Flow-guided feature aggregation for video object detection," in *IEEE CVPR*, 2017.

[16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Detect to track and track to detect," in *IEEE CVPR*, 2017, pp. 3038–3046.

[17] Fanyi Xiao and Yong Jae Lee, "Video object detection with an aligned spatial-temporal memory," in *ECCV*, 2018.

[18] Fengchao Xiong, Jun Zhou, and Yuntao Qian, "Material based object tracking in hyperspectral videos," *IEEE TIP*, vol. 29, pp. 3719–3733, 2020.

[19] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey, "Learning background-aware correlation filters for visual tracking," in *IEEE ICCV*, 2017.

[20] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg, "Discriminative scale space tracking," *IEEE TPAMI*, vol. 39, no. 8, pp. 1561–1575, 2016.

[21] Joao F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "High-speed tracking with kernelized correlation filters," *IEEE TPAMI*, vol. 37, no. 3, pp. 583–596, 2014.

[22] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *ECCV*, 2016.