MULTI-FEATURES INTEGRATION BASED HYPERSPECTRAL VIDEOS TRACKER

Zhe Zhang[1,§], *Kun Qian*[2,§] , *Juan Du*[1], *Huixin Zhou*[1]

[1]Lab of Optoelectronic Imaging and Image Processing, Xidian University
Xi'an 710071, P.R.China.
[2]School of Artifical Intelligence and Computer Science, Jiangnan University
Wuxi Jiangsu 214122, P.R.China.
Jiangsu Key Laboratory of Media Design and Software Technology
Wuxi Jiangsu 214122, P.R.China.

[1]

ABSTRACT

Most target tracking is over visible videos, but in a challenging scene, tracking targets with the same appearance is very difficult on visible videos, due to the limitation of grayscale and color information. Therefore, we use Hyperspectral Videos (HSVs) with rich spectral information for target tracking to distinguish similar targets. In this paper, a multi-features integration based tracking method is proposed over HSV. The feature maps are generated by Histogram of Gradient (HOG) and pretrained VGG-19 network, and then kernelized correlation filter framework is utilized to detect target over HSVs. Specially, More information of spatial, spectral and temporal are all used to extract useful features, and these feature can track the target that can not be tracked in visible videos. The experimental results on HSVs show that the proposed method has better performance than the three existing tracking methods with hyperspectral information.

***Index Terms*—** Target tracking, Hyperspectral video, Multi-features, Correlation filter, Deep learning

1. INTRODUCTION

Object tracking is one of the hotspots and has been applied in the tasks of computer vision, such as military reconnaissance, security monitoring and autonomous driving. The initial tracking methods were developed on grayscale or RGB videos. It is easy to mistake the target for a similar-looking background. In addition, the non-directional movement of the target will cause its appearance change and occlusion during tracking. To solve these problem, Hyperspectral Images (HSIs) is used in remote sensing to use its rich spectral information that can recognize the characteristics of inherent material [1]. Yet, there is little work focusing on object tracking using Hyperspectral Videos (HSVs).

In the field of tracking over RGB videos, Discriminative Correlation Filter (DCF) based framework [2, 3, 4, 5] has gained much attention, because of its good performance and high speed. In DCF, a correlation filter is learned to localize the target in consecutive frames, and then the target location is estimated via the maximum response. Several popular algorithms contains the Minimum Output Sum of Squared Error Filter (MOSSE) tracker [3], Kernelized Correlation Filter (KCF) [4] and Spatial-Temporal Context (STC) [5]. Specially, grayscale features, Histogram of Gradient (HOG), color name features[2] or other traditional feature are utilizied. Although these features are well designed, these tracking algorithms may not obtain high performance over types of background. To enhance the performance of tracking algorithms, two main approaches have been adopted, including feature fusion based methods [6] and deep network based methods [7]. The first approach combines at least two features together to deal with complex environment. The other approach utilizes characteristic of the neural network to obtain a good representation of object.

Based on the above observations,, we propose a Multi-Features Integration based Hyperspectral Videos Tracking (MFI-HVT) method. The tracker is divided into two steps, feature extraction and tracking task (The framework is shown in Figure 1). In the feature extraction part, a VGG-19 network [8] is used to obtain convolutional feature. Considering the different challenging factors such as background clutter and occlusion, HOG feature is also fused to improve the stability of the tracker. In the tracking part, each feature is adopted to obtain a weak response map of the target in KCF, and then an ensemble response map is calculated referring to all weak response maps.

The rest of this paper is organized as follows. In Section 2, the proposed method is described in detail. Besides, Section 3 presents the experimental results of proposed MFI-HVT on hyperspectral videos. Finally, the conclusion is drawn in Section 4.

[1]Author Contributions: § Zhe Zhang and Kun Qian contributed equally.

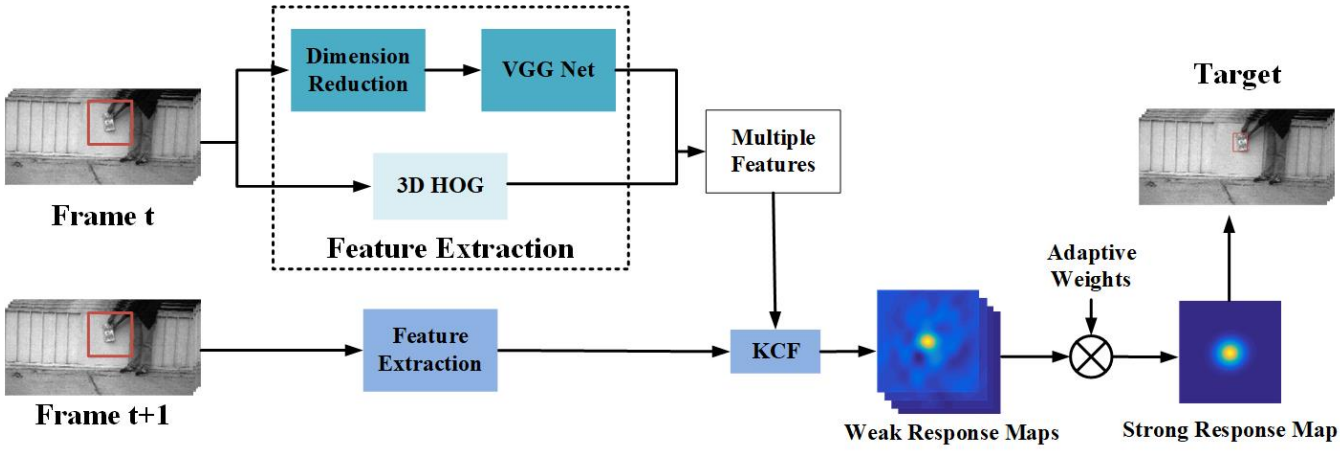


Fig. 1: The proposed framework.

2. THE PROPOSED TRACKING ALGORITHM

In this section, the detail of the KCF tracker [4] is described firstly. Then, the component of feature extraction is given.

2.1. The KCF tracker

The proposed method is built on the KCF tracker, which is trained using an image patch x with size of MN centered around the target. The key of the KCF tracker is to train a classifier using a ridge regression model.

$$min_w(||Xw - y||_2^2 + \lambda||w||_2^2) \tag{1}$$

where w, λ and y represent the regression coefficient, regularization parameter and regression value, respectively. X represents the data matrix by concatenating all the circular shifts of image patch x. X can be decomposed into Eq. 2.

$$X = F \cdot diag(\hat{x}) \cdot F^H \tag{2}$$

where F and F^H represent the Discrete Fourier Transform (DFT) matrix and its Hermitian transpose, respectively. x with the hat symbol denotes the DFT of the vector $x(\hat{x} = F(x))$, and diag() represents the diagonalization function.

After plugging Eq. 2 into Eq. 1, w is obtained by

$$w = F^{-1}(\frac{\hat{x}\hat{y}}{\hat{x}^H\hat{x} + \lambda}) \tag{3}$$

where $F^{-1}(\cdot)$ represents the inverse DFT.

Subsequently, kernel trick is applied in KCF tracker, w is mapped to a high-dimensional feature space by $w = \alpha^H\varphi(x)$, where $\varphi(x)$ denotes the mapping function and α is a coefficient calculated by

$$\alpha = (K + \lambda I)^{-1}y \tag{4}$$

$$\hat{\alpha} = (F(k^{xx}) + \lambda)^{-1}\hat{y} \tag{5}$$

where kernel matrix K is a circulant matrix with k^{xx} denoting the first row. Besides, k can be described as

$$k^{xx} = exp(-\sigma^{-2}(||x||^2 + ||z||^2) - 2F^{-1}(\hat{x}\hat{z})) \tag{6}$$

Finally, the target location is obtained by response map

$$f(z) = F^{-1}(F(k^{xz})\hat{\alpha}) \tag{7}$$

where x and α are learnt in advance. It can be found from Eq. 7 that the response map $f(z)$ is a linear combination of the neighboring kernel value k^{xz} with the weighted coefficient α.

2.2. Feature Extraction

Each feature has its particular advantages in target matching. Specially, the rotation invariant of HOG feature makes it easily to solve most problems in tracking. Deep features can be utilized to solve the occlusion problem that cannot be solved well by HOG feature. As described in [9], the response map constructed by single layer in VGG-19 network is not accurate enough. In the proposed MFI-HVT method, both shallow feature and deep features are utilized for tracking, the shallow feature uses HOG feature as the first level, the deep features uses the outputs of the conv 3-4, conv 4-4 and conv 5-4 layers in VGG-19 network. To improve the performance of the tracking method, adaptive weighting coefficient is applied to describe the different contribution referring to types of response maps.

As mentioned in Section 2.1, a response map is realized by using a filter to correlate an image, and its peek value represents a high degree of similarity between the current frame target and the template frame target. Therefore, the adaptive weighting coefficient w_i is calculated by

$$w_i = R_{pci}/R_{pai} \tag{8}$$

where R_{pci} and R_{pai} represent the i-th level peak response value of current frame and the i-th level peak response value of all time, respectively.

Due to the fact of background variation in image sequences, R_{pai} can be updated via

$$R_{pai} = \begin{cases} R_{pci} & R_{pci} \geq R_{pai} \\ \mu R_{pai} + (1-\mu) R_{pci} & R_{pci} < R_{pai} \end{cases} \quad (9)$$

where μ denotes a coefficient.

It is found in the experiment that the higher the feature level, the stronger the ability to adapt to complex situation. Therefore, the high-level feature is weak for local discrimination and cannot accurately track targets in the case of strong background clutter. Accordingly, an activation function is applied to the fourth level response map instead of adaptive weighting coefficient.

$$w_4 = \begin{cases} w_4 & w_4 \geq v_{th} \\ 0 & else \end{cases} \quad (10)$$

where v_{th} represents a threshold value.

The input of the pretrained VGG-19 network is a 3-channels image, yet a HSI contains 16 bands. Therefore, dimension reduction is adopted in the step of feature extraction. In this paper, the Principal Component Analysis (PCA) and Maximum Difference Selection (MDS) are combined together. Further, the spectral features are also utilized to prevent the problem of information loss, due to the fact of data reduction.

PCA. PCA attempts to provide a set of orthogonal axes along which data can be projected, hoping to account for most data with just the first few axes in the new space. It is proved that the principal components are the maximum likelihood estimators of the sources. The principal axes are found by diagonalize the covariance matrix of centered observations,

$$C = m^{-1} \sum_{i=1}^{m} X X^T, (s.t. \sum_{i=1}^{m} x_i = 0) \quad (11)$$

where C is the covariance matrix, the superscript T is the transpose operation, X is a centered matrix and m denotes the column number.

Note that C is positive definite, and thus it can be diagonalized with nonnegative eigenvalues, which eigenvalues and eigenvectors are obtained using Eq. 12.

$$\lambda \mathbf{v} = C \mathbf{v} \quad (12)$$

where λ are eigenvalues and $\mathbf{v}$ are eigenvectors, which is calculated by

$$\mathbf{v} = (m\lambda)^{-1} \sum_{i=1}^{m} (x_i \mathbf{v}) x_i \quad (13)$$

By sorting λ descending, the largest eigenvalue λ_m is selected as the principal component, and its corresponding eigenvector $\mathbf{v}_m$ is used to calculate a low-dimensional data.

$$D_l = \mathbf{v} X \quad (14)$$

MDS. MDS is a fast and effective dimensionality reduction method, which is used to make up for the weakness of single dimension reduction method. In HSI, the response of different wavebands can be seen directly from the spectral curve. The spectral difference curve can be calculated by the spectral curve of target and background. After calculation, the band with maximum D_{ri} value is select as the low-dimensional data.

$$D_{ri} = |R_{ti} - R_{bi}| \quad (15)$$

where R_{ti} and R_{bi} represent the target and background response, respectively. D_{ri} means the response difference referring to the i-th waveband.

Both two low-dimensional data sets pass through the pretrained VGG-19 network, and the output of the same layer is concatenated as the final feature.

2.3. Scale Estimation

To improve the robustness of tracking, a scale estimation approach is used. Let $m \times n$ denotes the target size we got from the previous frame, $(2 \times S + 1)$ is the number of different scales used to select the best scale. For each $i \in \{-S, -S+1, ..., S-1, S\}$, an image patch I_i with size of $a^i m \times a^i n$ around the target location is extracted, and then a possibility degree P_i is calculated using correction filter. Here, a denotes the step of the scale. Besides, considering the continuity of target size, we penalize the scale different from the target size of the previous frame with a penalty parameter γ.

$$s = \arg\min_i \ \gamma^i \times P_i \quad (16)$$

where s denotes the best scale factor, and the target size is $a^s m \times a^s n$.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we first introduce details of experiment setting, and then give our results, analysis and comparisons.

3.1. Experimental Setup

The proposed MFI-HVT method is implemented with MATLAB R2016b and obtains 3 frame per second on a PC with an Intel i5-8500 CPU (3 GHz), 16 GB RAM and a TITAN V GPU, and the MatConvNet toolbox is used for extracting the deep features in VGG-19 network.

Further, the tracking framework learning rate is set to 0.025, the scale learning rate is set to 1.02, μ is set to 0.98 and v_{th} is set to 0.8.

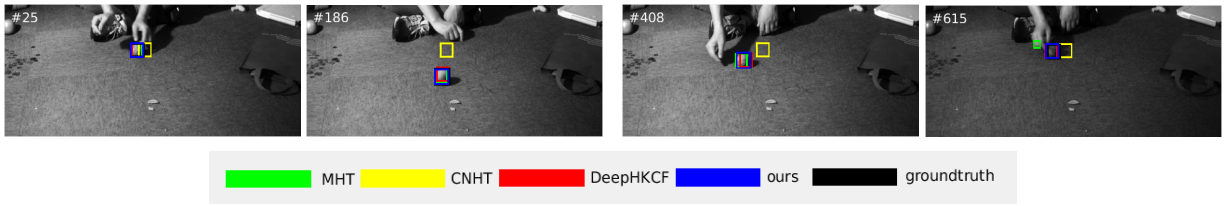


Fig. 2: Qualitative results on the ball sequence.

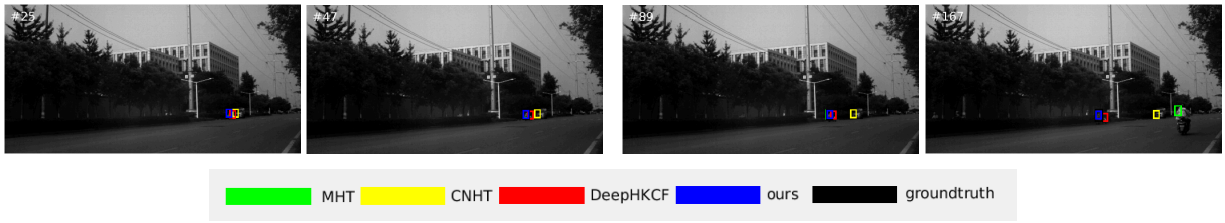
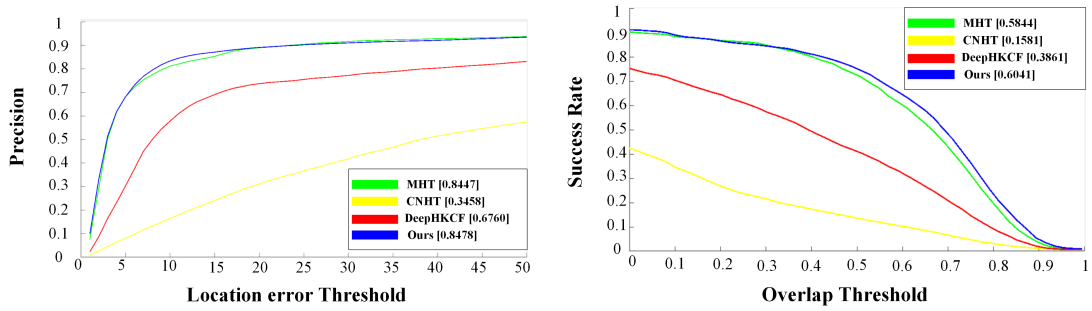


Fig. 3: Qualitative results on the rider sequence.



(a) Precision plot

(b) Success plot

Fig. 4: Quantitative results for all sequences.

3.2. Qualitative Comparison

To validate the performance, three existing hyperspectral trackers, CNHT [10], DeepHKCF[11] and MHT [12] are utilized. All the three trackers are based on KCF with different features. In CNHT method, normalized three-dimensional patches are selected from the target region in the initial frame as fixed convolution kernels for feature extraction in succeeding frames. In DeepHKCF method, HSIs are converted into false-color images to obtained deep features using VGG-19 network. In MHT method, the material feature does well in distinguishing objects in the same color. Experiment results are shown in Figs. 2 and 3. For convenience, HSIs are converted into grayscale images by computing the mean value of all bands.

3.2.1. *Partial Occlusion.*

Fig. 2 shows the situation that the target is partially occluded. In the first few occlusions, nearly all the trackers can re-detect the target when the target re-appears (e.g. #408), but with the increase of occlusion times and occlusion time, several trackers lose the target (e.g. #615).

3.2.2. *Scale Variation.*

The targets in all sequences contain scale variation. In Fig. 3, the rider is particularly small in the initial frame, and as time goes on, the target will become larger. Only the proposed method can track the target accurately (e.g. #167).

The above experimental results show that the proposed method performance well than the other comparing methods.

3.3. Quantitative Comparison

In this section, we compare four methods quantitatively using precision plot and success plot. Precision plot shows the percentages of successful frames whose distance between the center of predicted bounding box and the center of ground truth is smaller than a threshold (in pixels), which reflects

the accuracy of the center positioning. Success plot shows the percentages of successful frames whose overlap ratio is larger than a threshold varied from 0 to 1, which reflects the accuracy of scale selection. The Area Under Curve (AUC) of each plot is an overall evaluation measure to analyze the performance of all trackers.

Fig. 4 shows the quantitative results for all sequences. It can be seen that the proposed method gets the highest AUC score in both plots (0.8478 in precision plot and 0.6041 in success plot). The results show CNHT gives inferior accuracy because its filters are trained used only positive samples, which leads to the inability to obtain a robust model, and then leads to easy tracking failure. As for DeepHKCF, the VGG-Net, which uses both positive samples and negative samples, is utilized to learn robust filters, and therefore it has achieved relatively good results. But DeepHKCF only uses deep feature to track target, which has resulted in a great limitation. MHT adds material information to distinguish targets, which makes full use of spectral information and spatial information, and achieve better results.

In addition, we also provide the mean precision and FPS comparison of different tracking methods in Table 1, and our method can achieve comparable efficiency compared with the other three methods. Different from three trackers mentioned above, the proposed method also utilizes temporal information to track targets on the basis of spatial information and spectral information, which leads to the best precision among these four methods and high speed capturing the target. Both qualitative results and quantitative results confirm this conclusion.

Table 1: Precision and FPS

Algorithm	Video Type	Mean Precision (20px)	Mean FPS
MHT[12]	HSI	86.2%	1.2 (CPU)
CNHT[10]	HSI	30.2%	1 (CPU)
DeepHKCF[11]	HSI	73.2%	3 (GPU)
MFI-HVT(Ours)	HSI	86.3%	2.6 (GPU)

4. CONCLUSIONS

In this paper, a Multi-Features Integration based method is introduced to tracking targets of interest in hyperspectral videos. The integrated feature provides more discriminative information than separate features. The adaptive weights ensure good robustness to complex background. A novel convert method is utilized to make HSI adapt pretrained VGG-19 network. It can prevent more spectral information rather than traditional dimension reduction method. The experiment results show that the proposed method achieve good results in an existing hyperspectral tracking dataset.

5. REFERENCES

[1] Y. Li, W. Xie, and H. Li: Hyperspectral image reconstruction by deep convolutional neural network for classification. Pattern Recognition **63**, 371–383 (2017).

[2] M. Danelljan, F. Khan, M. Felsberg and J. Weijer: Adaptive color attributes for real-time visual tracking. IEEE Conference on Computer Vision and Pattern Recognition, 1090–1097 (2014).

[3] D. Bolme, J. Beveridge, B. Draper and Y. Lui: Visual object tracking using adaptive correlation filters. IEEE Conference on Computer Vision and Pattern Recognition, 2544–2550 (2010).

[4] J. Henriques, R. Caseiro, P. Martins and J. Batista: High-Speed Tracking with Kernelized Correlation Filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 583–596 (2015).

[5] K. Zhang, L. Zhang, and Q. Liu: Fast visual tracking via dense spatio-temporal context learning. European Conference on Computer Vision, 127–141 (2014).

[6] C. Bailer, A. Pagani, and D. Stricker. A superior tracking approach: Building a strong tracker through fusion. European Conference on Computer Vision, 2014.

[7] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking, arXiv preprint arXiv:1704.06036, 2017.

[8] K. Simonyan, A. Zisserman, A. Kelly: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv **1409**, 1556 (2014)

[9] C. Ma, J.B. Huang, X. Yang: Hierarchical convolutional features for visual tracking, IEEE International Conference on Computer Vision, 3074–3082 (2015).

[10] K. Qian, J. Zhou, F. Xiong, and H. Zhou: Object tracking in hyperspectral videos with convolutional features and kernelized correlation filter. International Conference on Smart Multimedia, (2018).

[11] B. Uzkent, A. Rangnekar, and M. J. Hoffman: Tracking in aerial hyperspectral videos using deep kernelized correlation filters. IEEE Transactions on Geoscience and Remote Sensing, **57**(1), 449–461 (2019).

[12] F. Xiong, J. Zhou and Y. Qian: Material Based Object Tracking in Hyperspectral Videos. IEEE Transactions on Image Processing, **29**(1), 3719–3733 (2020).