Material Based Object Tracking in Hyperspectral Videos

Fengchao Xiong, Student Member, IEEE, Jun Zhou, Senior Member, IEEE, and Yuntao Qian, Member, IEEE

Abstract—Traditional color images only depict color intensities in red, green and blue channels, often making object trackers fail in challenging scenarios, e.g., background clutter and rapid changes of target appearance. Alternatively, material information of targets contained in large amount of bands of hyperspectral images (HSI) is more robust to these difficult conditions. In this paper, we conduct a comprehensive study on how material information can be utilized to boost object tracking from three aspects: dataset, material feature representation and material based tracking. In terms of dataset, we construct a dataset of fully-annotated videos, which contain both hyperspectral and color sequences of the same scene. Material information is represented by spectral-spatial histogram of multidimensional gradients, which describes the 3D local spectral-spatial structure in an HSI, and fractional abundances of constituted material components which encode the underlying material distribution. These two types of features are embedded into correlation filters, yielding material based tracking. Experimental results on the collected dataset show the potentials and advantages of material based object tracking.

Index Terms—Object tracking, hyperspectral imaging, material unmixing.

I. INTRODUCTION

OBJECT tracking is one of the fundamental tasks in computer vision, particularly for video surveillance and robotics. It is a challenging task to detect the size and location of an object in video frames given a bounding box of the object in the first frame. Many efforts have been made to obtain informative visual cues, such as color intensities [1], color names [2], texture [3], [4], deep feature representation [5]–[7] and more for object tracking. Alternatively, many researches focus on imposing temporal consistency to achieve robust tracking [8], [9]. However, tracking in traditional color videos has its inherent limitations. Trackers tend to drift in some challenging scenarios, e.g., background clutter, similar color between tracking target and background, object rotation and deformation [10]–[13], as also exemplified by the recent failure of autonomous cars.¹ In these cases, however, the underlying material information of foreground and background is distinct, which can be used for distinguishing the target from the surrounding environment [14], [15].

Material information can be broadly obtained by two groups of methods, discriminative and generative. Discriminative approaches extract material information using visual appearance, e.g., color, shape, local textures, etc., in a single RGB image [15]–[18]. Such methods fail to identify similar appearances of different materials [15], for example artificial plastic fruits and real fruits, due to the limitations of color images in depicting the full physical property of surface reflectance. Alternatively, generative methods represent the material information based on the intrinsic reflectance of the material. Therefore, optical images such as spectral images and time-of-flight images are vastly used thanks to more captured information on the material of the recorded objects [19]–[21]. Hyperspectral images (HSI) are such a kind of spectral images that record continuous spectrum information instead of monochrome or color intensities on each object. The spectrum information provides details on the material constitution of contents in the scene and increases the inter-object discrimination capability [22]. Fig. 1 shows a

¹https://www.tesla.com/blog/tragic-loss
sample HSI for material identification from two toy minions built from different materials. Though pixels 1 and 2 are both black in color, their recorded spectral responses are different. The same observation can be made from Figs. 1(d) and 1(e). Benefiting from the superior advantages of material identification along the spectral dimension, HSIs have enabled many unique applications in remote sensing [23], computer vision [14], [22], [24], and more [25]. For example, thanks to the valuable spectral information, HSIs have been adopted to chemical gas plume tracking [26], [27].

Despite the material identification ability of HSIs, how to effectively extract and use this information for visual object tracking is facing many challenges. First, the high dimensionality nature of HSIs caused by large number of spectral bands and lower spatial resolution because of the imaging mechanism bring difficulties for robust material information extraction. Traditional feature extractors developed for monochrome or color images may not provide highly discriminative material information for hyperspectral data because of the ignorance of valuable spectral information. Pixel-wise spectral reflectance was adopted as the feature for object tracking in early attempts [28]–[30], but the spatial structure is ignored. Alternatively, Uzken et al. [31] proposed a deep kernelized correlation filter based method (DeepHKCF) for aerial object tracking at the sacrifice of valuable spectral information, in which an HSI was converted to false-color image before passing to a deep convolutional neural network. Qian et al. [32] selected a set of patches as convolutional kernels for each band to extract features, but the correlations among bands were neglected. It is known that local structure inside an object region facilitates exploiting material information [17]. Therefore, spectral-spatial feature extraction methods, which simultaneously consider local spatial information, local spectral information and joint local spectral-spatial information in an HSI, are more favorable for material information representation.

Second, there is a lack of datasets with high diversity to support hyperspectral object tracking. One reason is that with the limitation of most existing hyperspectral sensors, it is difficult to collect real-time hyperspectral videos at high frame rate with high signal-to-noise ratio and high spatial and spectral resolutions. Recently, Uzken et al. [31] introduced a synthetic aerial dataset generated by Digital Imaging and Remote Sensing (DIRSIG) software at 1.42 fps. However, the objects are described at very low frame rate and in the remote sensing setting, making it difficult to cover the challenges in close-range computer vision setting, i.e., rotation, deformation, illumination variation, etc.

In this paper, we tackle object tracking problem from a new perspective in which the material properties are considered. As analyzed in [33], a good powerful feature representation is always important for tracking. To this end, this paper tackles material based object tracking by extracting material features from HSIs. Specifically, instead of extracting features in 2D images, we develop two spectral-spatial feature extractors, local spectral-spatial histogram of multidimensional gradients (SSHMG) and spatial distribution of materials, to capture the material appearances in an HSI. SSHMG captures the local spectral-spatial texture information in terms of the spatial and spectral gradients orientations. The distributions of underlying constituent materials in the scene are encoded by fractional abundances. Abundances behave much more like bag-of-words where they implicitly discover recognition-related structures in the spectral space, making them superior to spectral representation. They are obtained by hyperspectral unmixing which decomposes an HSI into constituent spectral (or endmembers) and their corresponding fractions (or abundances). As shown in Fig. 2, two visual feature descriptors are further embedded to background-aware correlation filter, yielding material based tracking. To evaluate the effectiveness of material features in object tracking, we introduce a fully annotated dataset with 50 hyperspectral videos captured by a commercial high-speed hyperspectral camera. To our best of knowledge, this is the first large-scale fully-annotated hyperspectral video dataset in close-range computer vision setting. We also provide color videos captured and annotated on the same scene for fair evaluation purpose. Extensive experiments with detailed analysis are carried out to explore a better understanding of how material information can be used to promote object tracking. The experimental results show that material based tracker even out-performs several state-of-the-art deep learning based trackers, indicating the effectiveness of proposed material features and the promising potentials of HSIs in object tracking.

The remainder of this paper is organized as follows: Section II gives the detailed description on the collected dataset and comparison with other well-know color and color-depth datasets. Proposed local SSHMG and global abundances are addressed in Section III. In Section IV proposed features are embedded to correlation filter framework to yield material based tracking method. Section V presents the experimental results on the collected dataset. The paper is concluded in Section VI.

II. HYPERSONTAL TRACKING DATASET

In this section, we provide the details about the hyperspectral tracking dataset including data collection, data annotation and attribute statistics. The differences between datasets with other modalities and advantages of our dataset are also presented.

A. Dataset Construction

1) Data Collection: Traditionally, hyperspectral images are mostly used in remote sensing and collected by sensors mounted on satellite. Recent progress on sensors makes it possible to collect hyperspectral sequences at video rate in close range. In our research, a snapshot mosaic hyperspectral camera\(^2\) was used to collect videos. This camera can acquire videos up to 180 hyperspectral cubes per second, each of which contains 512 \(\times\) 256 pixels and 16 bands in the wavelength from 470nm to 620nm. In our data collection, we captured videos at 25 frames per second (FPS), where a frame refers to a 3D hyperspectral cube with two dimensions indexing the spatial location and the third dimension indexing the spectral band. In order to more evidently verify the

\(^2\)https://www.imec-int.com/en/hyperspectral-imaging
advantages of material features in object tracking and also ensure a fair comparison with color-based tracking methods, RGB videos were also acquired at the same frame rate in a very close view point as the hyperspectral videos.

2) Data Preprocessing: Hyperspectral cameras require preprocessing steps to ensure high-quality HSIs. The preprocessing steps include spectral calibration, image registration and color conversion. Their details are listed as follows.

Spectral Calibration: Our calibration process involves two steps: dark calibration and spectral correction. Dark calibration aims to remove the effect of noises produced by the camera sensor. Following the manual provided by IMEC company, we performed dark calibration by subtracting a dark frame from the captured image, in which the dark frame was captured when lens was covered by a cap. The goal of spectral calibration is to suppress the contributions of unwanted second order responses, for example response to wavelengths leaking into the filters. By applying the sensor-specific spectral correction matrix on the acquired reflectance, the resulted spectrum is more consistent with the expected spectrum. An example is given in Fig. 3. The blue curve is acquired spectrum without correction and the red curve is corrected spectrum.

Color to Hyperspectral Image Registration: We registered the hyperspectral sequences and color sequences to make them describe almost the same scene. Specifically, we manually selected some points in the initial frame of both hyperspectral and color videos as key points. These points were matched and then used for the geometrical transformation. Finally, the resulted matrix was used to transform the color images in all the subsequent frames to align with the corresponding hyperspectral frames.

Hyperspectral to Color Image Conversion: Although color videos and hyperspectral videos were shot from almost the same viewpoint at the same frame rate and two videos describe almost the same content, there might be subtle differences between two videos. To ensure fair comparison, the hyperspectral videos were converted to false color videos using CIE color matching functions (CMFs). The CMFs indicate the weights of each wavelength in a hyperspectral imagery (HSI) when they are used to generate red, green, and blue channels. Given an HSI $\mathbf{H} \in \mathbb{R}^{N \times K}$ with $N$ pixels and $K$ bands, this step converts the HSI to a CIE XYZ image, formulated as

$$\mathbf{I} = \sum_{k=1}^{K} \mathbf{H}(::, k) \mathbf{A}(k, :)$$

where $\mathbf{A} \in \mathbb{R}^{K \times 3}$ are the CMFs. Here we used 10-deg XYZ CMFs transformed from the CIE (2006) with a step size of 0.1nm. After that, $\mathbf{I}$ is converted to the default RGB color space sRGB. Furthermore, the color transformation method in [34] was applied to make the color intensity of converted image close to that of the corresponding collected color frame.

3) Data Annotation: We manually drew a single upright bounding box for each frame to obtain the ground truth location of the target. The bounding box is represented by the centre location and its height and width. Though we tried

---

3http://cvrl.ioo.ucl.ac.uk/cmfs.htm
our best to make the hyperspectral video and color video capture the same scene, the detailed location of the target may differ slightly in two types of videos. To make the ground truth as precise as possible, we labelled hyperspectral and color videos independently. Every pair of hyperspectral video and color video in the same scene was labeled by the same person to ensure that the videos were annotated under the same protocols. In this way, there is nearly no difference in annotation between two types of videos. Every video was labelled by three volunteers and one of them was an expert to check the bounding boxes and adjust them if necessary. With the above strategy, the quality of data labels is guaranteed.

4) Data Attributes: Here, we show the first frame of all the sequences in the collected dataset in Table I. Due to limited space, only color videos are presented. The whole dataset contains 50 color videos and 50 hyperspectral videos with an average of around 425 frames in each sequence. Every video is also labelled with associated challenging factors out of eleven attributes [35], including illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutters (BC), and low resolution (LR). Following the challenging factors listed in [35], we carefully collected videos to include multiple target categories, diverse scenarios, rich activities and diverse content etc., guaranteeing the generality and complexity of the dataset. For example, the tracked objects are of high diversity, including vehicles, faces, people, generic objects, animals, etc. Additionally, we collected objects in different scenes (indoor and outdoor) and different viewpoints. To account for the impact of color on object tracking, we also collected the same object class with different colors, by which, the researchers are able to more comprehensively exam the role of material information in object tracking.

Table II shows the number of coincident attributes across all videos. As can be seen, all the tracking factors are considered and the most common challenging factors are SV, OPR, BC, IV and OCC. These factors remain difficult for color tracking. With the usage of the material identification ability of HSIs, it is expected that our dataset will enable improved tracking performance in some scenes, such as DEF, OPR and BC, where the shape of objects changes but not the materials. The full dataset including the sequences and annotations have been available online.

B. Dataset Comparison

In recent years, many other modality datasets have been developed for object tracking, which significantly boost object

4https://bearshng.github.io/mht/
TABLE II

| Attributes | SV | MB | OCC | FM | LR | IP | OPR | DEF | BC | IV | OV |
|------------|----|----|-----|----|----|----|-----|-----|----|----|----|
| SV         | 35 | 5  | 19  | 4  | 9  | 10 | 15  | 5   | 8  | 14 | 4  |
| MB         | 5  | 6  | 3   | 2  | 2  | 2  | 3   | 1   | 1  | 0  | 0  |
| OCC        | 19 | 3  | 25  | 4  | 9  | 3  | 5   | 2   | 6  | 11 | 3  |
| FM         | 4  | 2  | 4   | 6  | 4  | 1  | 1   | 0   | 2  | 1  | 0  |
| LR         | 9  | 2  | 9   | 4  | 12 | 0  | 0   | 0   | 2  | 6  | 0  |
| IP         | 10 | 2  | 3   | 1  | 0  | 15 | 4   | 14  | 6  | 1  | 0  |
| OPR        | 15 | 3  | 5   | 1  | 0  | 14 | 19  | 7   | 9  | 2  | 2  |
| DEF        | 5  | 1  | 2   | 0  | 1  | 4  | 7   | 8   | 3  | 1  | 0  |
| BC         | 8  | 1  | 6   | 2  | 2  | 6  | 9   | 3   | 17 | 1  | 1  |
| IV         | 14 | 0  | 11  | 1  | 6  | 1  | 2   | 1   | 16 | 3  | 4  |
| OV         | 4  | 0  | 3   | 0  | 0  | 0  | 2   | 0   | 1  | 3  | 4  |

TABLE III

| Dataset     | Videos | Min Frames | Mean Frames | Max Frames | Total Frames | Per-frame Annotation | Data Modality | Data Source |
|-------------|--------|------------|-------------|------------|--------------|----------------------|---------------|-------------|
| OTB50       | 50     | 71         | 578         | 3872       | 59K          | ✓                    | Color&gray   | -           |
| TC128       | 128    | 71         | 429         | 3872       | 55K          | ✓                    | Color        | -           |
| VOT2017     | 60     | 41         | 536         | 1500       | 21K          | ✓                    | Color        | -           |
| TrackingNet | >30K   | -          | 471         | -          | 509K         | ×                    | Color Internet | -           |
| GOT-10k     | 10K    | -          | -           | -          | 1.5M         | -                    | Color Youtube | -           |
| LaSOT       | 1.4K   | 1000       | 2506        | 13937      | 3.52M        | ✓                    | Color Youtube | -           |
| UAV123      | 123    | 109        | 915         | 3085       | 112K         | ×                    | Color UAV Platform | RGBG sensors |
| CDTB        | 80     | 406        | 1274        | 2501       | 101K         | ✓                    | RGBD sensors | -           |
| Our dataset-HSI | 50 | 53        | 425        | 1111       | 21K          | ✓                    | Hyperspectral | Hand-hold camera |
| Our dataset-False Color | 50 | 53        | 425        | 1111       | 21K          | ✓                    | False-Color | -           |
| Our dataset-RGB | 50 | 53        | 427        | 1209       | 21K          | ✓                    | Color | -           |

tracking research. Here, we compare our dataset with several representative datasets to show the differences and advantages of our dataset. OTB50 [36], TC128 [37], VOT2017 [38], TrackingNet [39], GOT-10k [40], LaSOT [41], UAV123 [42] and CDTB [43] are selected for comparison.

Table III summaries the statistical properties of these datasets. As can be seen, our dataset contains three kinds of videos, hyperspectral videos, color videos and false-color videos converted from hyperspectral videos. The shortest video consists of 53 frames while the maximum number of frames in a sequence is over 1000.

Our dataset is unique and essentially different from other datasets. First, unlike recently-introduced very large-scale datasets such as LaSOT, TrackingNet and GOT-10k which are collected from the Internet, our videos are collected by our own cameras in a real-world setting. As far as we know, this is the first hyperspectral dataset captured by a hyperspectral video camera in a close range computer vision setting. We expect that this dataset helps us better understand the role of HSIs in object tracking and provides a platform for researchers to develop more advanced trackers that take materials of target into account. Second, our dataset contains different modalities of videos, i.e., hyperspectral, color and false-color. This helps to analyze the importance of different types of information in object tracking. In addition, individual video modality can also provide complementary information for other types of videos, and different video modes can be combined to facilitate tracking. Third, different from datasets like UAV123 and TrackingNet, all the frames in our videos are annotated frame by frame, taking into account the valuable temporal information between adjacent frames. Fourth, the number of frames in single modality in our dataset is smaller than most of other datasets, but the data volume is still very large as HSI has much more spectral bands. Furthermore, during data collection, we fully consider different scenes, different categories and different activities to describe the movement in the real world as much as possible.

III. MATERIAL FEATURE REPRESENTATION

We focus on visual representation in appearance modeling, i.e., constructing robust object descriptors using different types of visual features. SSHMG and abundances are introduced in this section to respectively encode the local and global appearance of the target.

A. Spectral-Spatial Histogram of Multidimensional Gradients

As analysis earlier, an HSI is a 3D cube, comprising two spatial dimensions and one spectral dimension. Therefore, it can be naturally modelled by a third-order tensor without structural information loss. Previous works show that 3D spectral-spatial features are effective for HSI processing [44]. Instead of 2D patches, these features are constructed in a local 3D spectral-spatial neighborhood. Therefore, spectral information, spatial information and joint spectral-spatial information are simultaneously taken into consideration. In addition, since local information is beneficial to object tracking [45], [46] and
material information representation [15], we build a 3D local SSHMG descriptor for an HSI to exploit local information of materials, which is also less sensitive to illumination condition. As shown in Fig. 4, given an HSI $\mathcal{T} \in \mathbb{R}^{W \times H \times K}$ containing $W \times H$ pixels and $K$ bands, SSHMG is constructed as follows:

First, we convolve $\mathcal{T}$ in three modes using a finite difference filer $[-1, 0, 1]$ to obtain its multidimensional gradients $\nabla T_x, \nabla T_y, \nabla T_k$. Transferring the gradients to a spherical coordinate system, they can be identically represented by $(M, \theta, \phi)$, where $M$ represents the spectral-spatial multidimensional gradient magnitude, $\theta$ indicates the spatial gradient orientation, and $\phi$ denotes the spectral gradient orientation. Their definitions are given as follows:

$$M(x, y, k) = \sqrt{\nabla T_x^2 + \nabla T_y^2 + \nabla T_k^2}$$

$$\theta(x, y, k) = \arctan\left(\frac{\nabla T_y}{\nabla T_x}\right)$$

$$\phi(x, y, k) = \arctan\left(\frac{\nabla T_k}{\sqrt{\nabla T_x^2 + \nabla T_y^2}}\right)$$

(2)

Subsequently, the tensor is divided into a number of 3D cubes with the size of $z \times z \times z$. For each position in a local cube, its gradients orientations are then quantized in both spatial dimension ($B_\theta$) and spectral dimension ($B_\phi$), i.e.,

$$B_\theta(x, y, k) = \text{round}\left(\frac{n_\theta \phi(x, y, k)}{2\pi}\right) \mod n_\theta$$

$$B_\phi(x, y, k) = \text{round}\left(\frac{n_\phi \phi(x, y, k)}{\pi}\right) \mod n_\phi$$

(3)

where $n_\theta$ and $n_\phi$ respectively denote the number of bins in the spatial and spectral directions. Based on such quantization, we can define point-level spatial feature $F_\theta(x, y, k)$ and spectral feature $F_\phi(x, y, k)$ for each position as follows by

$$F_\theta(x, y, k)_b = \begin{cases} M(x, y, k), & \text{if } b = B_\theta(x, y, k) \\ 0, & \text{otherwise} \end{cases}$$

(4)

$$F_\phi(x, y, k)_b = \begin{cases} M(x, y, k), & \text{if } b = B_\phi(x, y, k) \\ 0, & \text{otherwise} \end{cases}$$

(5)

where $b$ indexes gradient orientation. Then we aggregate point-level features to obtain cube-level features $C_\theta(i, j, s)$ and $C_\phi(i, j, s)$ by summarizing point-level features in a cube, where $0 \leq i \leq \lfloor(W-1)/z\rfloor$, $0 \leq j \leq \lfloor(H-1)/z\rfloor$ and $0 \leq s \leq \lfloor(K-1)/z\rfloor$.

Afterwards, we get the block-level feature $\mathbf{v}$ by concatenating the cube-level features over overlapping $2 \times 2$ spatial blocks to reduce the sensitivity of feature descriptor to illumination and foreground-background contrast. $\mathbf{v}$ is then normalized to unit length by being divided with its $L_2$ norm. An element-wise truncation function $L_\alpha$ is applied to the normalized block-level feature $\mathbf{v}$ to limit the maximum values in a feature vector to $\alpha$, which is then renormalized. With such step, the proposed description is less sensitive to the illumination changes. Finally, the block-level features are concatenated in the spectral direction, yielding proposed SSHMG.

**B. Fractional Abundances of Constituted Materials**

Here, we introduce a useful feature, abundances from hyperspectral unmixing, to describe detailed material distribution in
a scene. Low spatial resolution of sensor and long distance from camera to imaging targets cause overlapped spectral responses of neighboring but different surface materials, leading to “mixed” pixels. Hyperspectral unmixing decomposes these pixels into a collection of spectral signatures, or endmembers, and associated proportions or abundances at each pixel. Fig. 5 provides an example of hyperspectral unmixing. In this scene, the abundances produced by hyperspectral unmixing clearly demonstrate the underlying distributions of three materials, namely, plastic lemon, background, and real lemon.

Modern unmixing methods can be broadly categorized into three groups, two-step, blind, and spare regression based. Two-step unmixing breaks this problem into two tasks, endmember extraction [47] and abundance estimation [48]. Alternatively, blind unmixing methods integrate both tasks into a single problem to simultaneously estimate endmembers and abundances [49], but their performance is limited by time-consuming optimization steps. Sparse regression based unmixing assumes the endmembers are a part of a predefined spectral library [50], but the spectral signatures in library must be consistent to the real endmembers in HSIs. However, all of the above methods can not address the standard permutation problem of unmixing in hyperspectral videos: the extracted endmembers may be not in the same order in adjacent frames [51].

Fig. 6 gives the framework of our online unmixing method. We first collect large scale of additional HSIs to formulate a hyperspectral database. The database is of high diversity, including multiple categories, scenes and imaging conditions, etc., to contain the spectrum of different materials. In this way, the library is likely to cover the endmembers in hyperspectral sequences processed.

Every image in the database is firstly unmixed by vertex component analysis (VCA) [47] to extract endmembers. Before VCA, hyperspectral signal subspace identification by minimum error (HySime) [52] method is utilized to estimate the number of underlying endmembers. The extracted endmembers are collected to form the endmember database. Afterwards, the elements in the endmember database are grouped by a $K$-means clustering algorithm. In our research, we set $K$ as 300. The spectral signatures of each cluster center is considered as an atom in the library. The whole process is offline.

As for tracking, we first use the collaborative sparse unmixing by variable splitting and augmented Lagrangian (CLSUnSAL) [53] method to select endmembers contained in the first frame. These endmembers are considered as the constituted materials of target. Specifically, let $H \in \mathbb{R}^{L \times N}$ be the target object in the bounding box, covering $L$ bands and $N$ pixels, CLSUnSAL selects the endmembers from the above library by enforcing group sparsity on the abundance matrix $S$, which is formulated as:

$$\arg \min_S \|H - AS\|_F^2 + \lambda \|S\|_{2,1} \quad \text{s.t.} \quad S \geq 0$$

where $A \in \mathbb{R}^{L \times M}$ is the predefined spectral library with $M$ materials. When the CLSUnSAL is converged, the sum of each row of $S$ is calculated, yielding a vector $s$ with $M$ entries. Each element in $s$ can be regarded as the total contributions of a specified material in $H$. The spectral signatures of top $R$ elements are then selected as the constituted endmembers in the scene, and $R$ is also determined by HySime.

During tracking procedure, we assume that the spectrum of target material is consistent, i.e., the endmembers are fixed for subsequent unmixing. The optimization of CLSUnSAL is iterative and needs a large number of steps to converge, which is time-consuming and unsuitable for online tracking. Therefore, it is not realistic to use CLSUnSAL to estimate abundance in all the frames. To this end, we use the analytic abundance estimation method, simplex-projection unmixing (SPU) [54], to get the abundance maps in all the frames thanks to its computational efficiency.

IV. MATERIAL BASED OBJECT TRACKING

Discriminative correlation filter (DCF) is widely used in object tracking due to its competitive performance and computational efficiency enabled by fast Fourier transform (FFT). DCF produces filters by minimizing the output sum of squared error [1] for all circular shifts of a training sample. The periodic assumption on training samples causes unwanted boundary effects, which can be alleviated by adding spatial regularization [4], [10], [55]. Concretely, background-aware correlation filter (BACF) considers all background patches as negative samples by using a rectangular mask covering central part of the circular samples [55]. Benefiting from alternating direction method of multipliers (ADMM) and FFT, BACF is also computationally efficient.

We adopt BACF [55] as our tracking method due to its computational efficiency. Under the framework of BACF, our material based tracking learns the filters $f_k$ by the following objective function:

$$\min_{f_k} \frac{1}{2} \sum_{d=1}^D \|y_d^d - \sum_{i=1}^N \sum_{k=1}^{K_i} f_k \mathbf{P} x_i^d \|_2^2 + \frac{\lambda}{2} \sum_{i=1}^N \sum_{k=1}^{K_i} \|f_k\|_2^2$$

where $x$ is the material feature, represented by SSHMG and abundances, $w_i$ contains their weights for determining the...
location of target, \( P \) is a binary matrix which crops the central patch of \( x_k \), \( \lambda \geq 0 \) is the regularization parameter, and \( D \), \( K_i \) represent feature dimensionality and number of pixels, respectively.

Taking advantages of FFT, the problem in Equation (7) can be reformulated in the frequency domain as:

\[
\min_{f} \frac{1}{2} \| \tilde{y} - \tilde{X} \sqrt{D}(Q \mathbf{P}^T \otimes \mathbf{I}_K) f \|_2^2 + \frac{\lambda}{2} \| f \|_2^2
\]  

(8)

where \( \tilde{X} \) is a \( D \times KD \) matrix concatenated with \( K \) diagonal matrices \( \text{diag}(w_j x_k) \) and \( f = [f_1^T, \ldots, f_f^T] \). \( Q \) is a orthonormal \( D \times D \) matrix, which maps any \( D \) dimensional vectorized signal to the Fourier domain. \( \mathbf{I}_K \) is a \( K \times K \) identity matrix, and \( \otimes \) denote Discrete Fourier Transform of a signal and Kronecker product respectively. By introducing auxiliary matrix \( \tilde{g} = \sqrt{D}(Q \mathbf{P}^T \otimes \mathbf{I}_K) f \), the Augmented Lagrangian function is:

\[
L(\tilde{g}, \tilde{h}, \zeta) = \frac{1}{2} \| \tilde{y} - \tilde{X} \tilde{g} \|_2^2 + \frac{\lambda}{2} \| f \|_2^2 + \frac{\mu}{2} \| \tilde{g} - \sqrt{D}(Q \mathbf{P}^T \otimes \mathbf{I}_K) f \|_2^2 + \frac{\mu}{2} \| \tilde{f} - \sqrt{D}(Q \mathbf{P}^T \otimes \mathbf{I}_K) f \|_2^2
\]

(9)

where \( \mu \) is regularization factor and \( \zeta \) is the multiplier. This problem can be divided into three subproblems with closed form solutions given by:

\[
f \leftarrow (\mu + \frac{\lambda}{\sqrt{D}})^{-1} P Q^T \otimes \mathbf{I}_K (\mu \tilde{g} + \zeta)
\]

\[
\tilde{g} \leftarrow \frac{1}{\mu} (D \tilde{y} d - \tilde{x} d + \mu \tilde{f} d)
\]

\[
\tilde{f} \leftarrow \mu (\tilde{y} d^{\top} \tilde{x} d - (\tilde{x} d)^{\top} \zeta + \mu \tilde{f} d)
\]

\[
O(B_i^t, B_i^t) = \frac{B_i^t \cap B_i^t}{B_i^t \cup B_i^t}
\]

(11)

Then, the overlap reliability score is determined by:

\[
O_i^t = \exp(-(1 - O(B_i^t, B_i^t))^2)
\]

(12)

In terms of distance reliability score, we also compare the differences between the predicted central location \( C_i^t \) from one particular feature and the final position \( C_i^t \), formulated as:

\[
D_i^t = \exp(-\|C_i^t - C_i^t\|_2^2).
\]

(13)
In addition, self-reliability score measures the trajectory smoothness degree of feature. It records the shift between the previous bounding box \( B_t^{i-1} \) and the current bounding box \( B_t^i \), which is given by:

\[
S_t^i = \exp \left( -\frac{\| C_t^i - C_t^{i-1} \|_2^2}{W(B_t^i) + H(B_t^i)} \right),
\]

(14)

where \( C_t^{i-1} \) is the central location determined by the \( i \)-th feature at the \((t-1)\)-th frame. \( W(B_t^i) \) and \( H(B_t^i) \) represent the width and height of the bounding box, respectively. With the above reliability measure, the final reliability weights of different features are given by:

\[
w_t^i = \frac{O_t^i + D_t^i + S_t^i}{\sum_{i=1}^{N} O_t^i + D_t^i + S_t^i}
\]

(15)

The model optimization, target detection and model update are as follows.

**Model Optimization.** The feature reliability is initialized with the same value for all the features at the initial frame. After each frame, the feature reliability is computed by Equation (15). With these features, the correlation filters are learned by Equation (10).

**Model Detection.** SSHMG and abundances extracted from the candidate area are re-scaled by their reliability weight. To deal with scale changes, multiple resolutions of the candidate regions are used. Afterwards, the weighted features are convoluted by filter \( G \) to derive the channel-wise responses, which are then summed, resulting the final output response. The interpolation strategy in [55] is employed to get the desired location and scale with sub-grid precision.

**Model Update.** The group reliability is computed by Equation (15). To account for appearance changes, for example scale and pose changes and rotation during tracking, group reliability scores and models are updated by an autoregressive method with learning rate \( \eta \), formulated as

\[
\begin{align*}
    w_{\text{model}}^i &= (1 - \eta)w_{i-1}^i + \eta w_t^i \\
    s_{\text{model}}^i &= (1 - \eta)s_{i-1}^i + \eta s_t^i
\end{align*}
\]

(16)

V. EXPERIMENTS

In this section, we first investigate our material-based tracking method from a feature representation perspective, then compare it with some state-of-the-art methods, including hand-crafted feature based trackers, deep feature based trackers and hyperspectral trackers. The parameters of competing methods are automatically set as suggested in the original implementation. Moreover, the attribute-based and quality-based comparison are also presented.

**A. Experimental Setting**

In our experiments, the parameters of SSHMG were set as \( \alpha = 0.2, \ z = 4, \ n_a = 9, \ n_b = 4 \). The learning rate \( \eta \) in MHT was set to 0.0023. All the other parameters were set to the same as those in BACF [55]. Three evaluation protocols, including precision plot, success plot and area under curve (AUC), were adopted to describe the performance of all the trackers. Precision plot records the fractions of frames whose estimated location is within a given distance threshold to the ground truth. The average distance precision rate is reported at a threshold of 20 pixels. Success plot shows the percentages of successful frames whose overlap ratio between the predicted bounding box and ground-truth is larger than a certain threshold varied from 0 to 1. AUC of each success plot is also selected as an evaluation measure to rank all trackers. All the results are presented with one-pass evaluation (OPE), i.e., a tracker is run throughout a test sequence with initialization from the ground truth position in the initial frame.

**B. Effectiveness of Proposed Material Feature**

In this experiment, we compare the effectiveness of six feature extractors, including spectrum, histogram of oriented gradients (HOG), band-wise HOG (BHOG) [31], abundances, SSHMG, and SSHMG combined with material abundances (abbreviated as MHT). For spectrum, the raw spectral response at each pixel was employed as the feature for tracking. HOG was calculated following [57], whose cell size and orientations were set to 4 × 4 and 9 respectively. BHOG was constructed by concatenating the HOG feature across all the bands. In this experiment, except for MHT in which the reliability weights were assigned in BACF, all other tracking steps were based on the framework of the original BACF.

Fig. 7 compares object tracking performance using different features. Spectral feature provides the worst accuracy among all the compared methods, as the raw spectrum is sensitive to illumination changes. Abundances exploit the underlying material distribution information, giving better performance. HOG and BHOG consider the local spatial structure information which is crucial for object tracking, therefore, produce more favorable results. SSHMG depicts local spectral-spatial structure information, yielding a gain of 4.4% in AUC compared with HOG. Notably, replacing the SSHMG with the combination of abundances and SSHMG allows MHT to dominate SSHMG and HOG by a relatively large margin in average distance precision scores. This is owing to the hybrid benefit from SSHMG and abundances. From this experiment, we can infer that generic descriptors proposed in color images may not adapt well to robust material information representation of HSIs. In contrast, spectral-spatial material features are more effective in providing more discriminative information for target tracking.

**C. Quantitative Comparison With Hand-Crafted Feature-Based Trackers**

In this experiment, we compare the proposed MHT tracker with ten state-of-the-art color trackers with hand-crafted features, including KCF [3], iDSST [58], SRDCF [4], MUSTer [59], SAMF [60], Struck [61], CNT [62], BACF [55], CSR-DCF [56], and MCCT [63]. KCF is a kernelized version of correlation filter in which kernel trick is applied to achieve non-linear classification boundaries. SAMF uses multi-scale resolution of the template to deal with the fixed template size problem in KCF. iDSST is a scale adaptive tracking approach in which two discriminative correlation filters are respectively learned for both object location and scale estimation.
SRDCF, BACF and CSR-DCF aim to mitigate the boundary effect of KCF caused by periodic assumptions on training samples via regularizing the correlation filters according to the spatial distribution. Structured support vector machine is adopted in Struck to predict the target location based on the training samples. MUSTer uses short-term memory and long-term memory stores to process target appearance memories, where long-term component is based on keypoint matching-tracking and RANSAC estimation. MCCT ensembles multiple experts to track the target via different levels of features.

The MHT tracker was tested on hyperspectral videos. Since all the alternative trackers were developed for color videos, they were run on both color videos and false-color videos generated from hyperspectral videos. Fig. 8 and Fig. 9 show the tracking performance of all methods. The results show that KCF and Struck give unsatisfying success scores because of limited consideration of scale estimation. MUSTer also produces inferior performance due to the fact that it fails to detect the key points when the object shares similar appearance as the background. BACF and SRDCF integrate the background information to learn more discriminative filters, resulting in much better tracking performance. It is worth mentioning that the proposed MHT tracker achieves competitive performance with C-COT on color videos and significantly outperforms the other trackers on both color and false-color videos. The reason of lower performance of other methods is the challenging nature of the dataset which contains background cluster, rotation and deformation, etc. In addition, most of deep trackers present lower AUCs in hyperspectral dataset. This means useful spectral information contained in hyperspectral videos is lost in the false-color videos.

D. Quantitative Comparison With Deep Feature-Based Trackers

In this section, we select several state-of-the-art deep feature-based trackers for comparison, including ECO [65], CF2 [66], TRACA [71], CFNet [67], HDT [68], DeepSRDCF [70], DSiam [69], and C-COT [64]. ECO adopts factorized convolution operator to reduce the number of parameters in deep network, and uses a DCF like tracker. CF2 trains correlation filters on each convolutional layer in CNNs and the target state is hierarchically obtained by a coarse-to-fine searching method. TRACA uses context-aware autoencoders to compress the convolutional features generated by VGGNet, facilitating fast tracking. CFNet is an end-to-end tracker in which correlation filter is interpreted as a differentiable layer in a deep neural network. HDT adopts an adaptive Hedge method to fuse several trackers trained by different CNN layers into a more robust tracker. DeepSRDCF replaces the hand-crafted features in SRDCF by convolutional features to achieve deep tracking. DSiam is equipped with a fast general transformation learning model to consider the temporal variations of both foreground and background during online tracking. C-COT trains convolution operators in the continuous spatial domain, enabling integration of multi-resolution feature maps rather than single-resolution feature maps in DCF.

As reported in Table IV, the proposed method achieves competitive performance with C-COT on color videos and significantly outperforms the other trackers on both color and false-color videos. The reason of lower performance of other methods is the challenging nature of the dataset which contains background cluster, rotation and deformation, etc. In addition, most of deep trackers present lower AUCs in hyperspectral dataset. This means useful spectral information contained in hyperspectral videos is lost in the false-color videos.

E. Quantitative Comparison With Hyperspectral Trackers

We also compared our method with two recent hyperspectral trackers, CNHT [32] and DeepHKCF [31]. Both CNHT and DeepHKCF are based on KCF but use different features. In CNHT, normalized three-dimensional (3D) patches were selected from the target region in the initial frame as fixed convolution kernels for feature extraction in succeeding frames. In terms of DeepHKCF, an HSI was converted into false-color image to learn deep features by VGGNet.
Fig. 8. Comparisons with hand-crafted feature-based trackers on RGB videos. MHT outperforms all the other trackers.

Fig. 9. Comparisons with hand-crafted feature-based trackers on hyperspectral videos or corresponding false-color videos. MHT achieves the best accuracy with an AUC of 59.5%.

**Table IV**

| Video            | MHT | C-COT [64] | ECO [65] | CF2 [66] | CNet [67] | HDT [68] | DStam [69] | DeepSRDCF [70] | TRACA [71] |
|------------------|-----|------------|----------|----------|-----------|----------|------------|----------------|------------|
| Color            | n/a | 0.581      | 0.586    | 0.449    | 0.590     | 0.391    | 0.552      | **0.594**    | 0.566      |
| Hyperspectral/FALSE-color | 0.595 | 0.568 | **0.587** | 0.423 | 0.519 | 0.402 | 0.464 | 0.569 | 0.517 |

Fig. 10 presents the tracking results of all competing hyperspectral trackers. The results show that CNHT gives inferior accuracy due to the fact that it only considers fixed positive samples in learning convolutional filters. Without negative patches in the surrounding background, the features produced by the fixed convolutional filters are not discriminative enough to learn a robust model, significantly deteriorating the prediction of object location. In contrast, VGGNet uses both positive and negative samples to learn discriminative feature representation. Therefore, DeepHKCF shows more competitive performance than CNHT. However, since HSIs are converted into three-channel false-color images before passing through the VGGNet, the complete spectral-spatial structural information in an HSI is not fully explored. This makes DeepHKCF fail to outperform the proposed MHT tracker. Combining local spectral-spatial texture information...
and detailed material information, the proposed MHT achieves the most appealing performance. This experiment again suggests that material information facilitates object tracking.

**F. Attribute-Based Evaluation**

In this experiment, we report the tracking effectiveness with respect to different video attributes. For simplicity, we only present the performance of top 9 color trackers on color videos and MHT on hyperspectral videos. Table V reports the AUCs of all the trackers. MHT ranks the first on 4 out of 11 attributes: IPR, OPR, BC, and DEF. In BC situations, it is difficult to extract robust features from the target in a color image. In contrast, the material information can be captured by an HSI and represented by the proposed SSHMG and abundance features, helping our tracker to discriminate the object from the background. On the videos with IPR, OPR, DEF attribute, the tracking target is partly or fully deteriorated, which makes the spatial structure information unreliable. Compared with spatial information, the underlying material information is more robust. Thanks to the superior advantages of the proposed SSHMG and abundances in spectral-spatial material structure representation and underlying material information exploitation, MHT tracker is more capable of separating target from surrounding environment. It is also observed that MHT fails to outperform color trackers in IV. The main reason is that the endmembers are fixed during tracking, which are in fact related to illumination condition.

**G. Running Time Comparison**

Table VI reports tracking speed of top eight trackers on CPUs with respect to frame per second (FPS). The Windows machine equipped with Intel(R) Core(TM) i5-4440 CPU@3.10GHz and 16GB RAM is used. BACF achieves the best speed on both videos. MHT cost more time than correlation filter based trackers like MCCT, SRDCF and BACF, but obtained similar speed with deep trackers.

**H. Extension to Other Trackers**

In the above experiments, we compare MHT with state-of-the-art color and hyperspectral trackers. Here, we further integrate proposed material features into KCF and ECO to
more evidently show the merits of material information in object tracking. Table VII compares the AUCs of the extended trackers with original trackers.

I. Qualitative Comparisons

Here we provide qualitative evaluations of the competing trackers on sample hyperspectral or false-color videos, as shown in Fig. 11. Due to ensemble trackers, MCCT provides better performance in rotation (drive), deformation (rubik) and clutter conditions (forest). BACF drifts away when a target is in low-resolution (worker) and shares similar color as the background (forest). Thanks to the high robustness of material properties, MHT shows higher robustness in tracking the objects in these scenarios.

VI. Conclusion

In this paper, we study the object tracking task from a new perspective where the material information is explored and introduce a dataset for object tracking in hyperspectral videos. The material information is embodied in the proposed SSHMG and abundance features. SSHMG encodes the local spectral-spatial texture information by summarizing the occurrences of spectral-spatial multidimensional gradients orientation in local cubes of an HSI. Abundance describes detailed constituent material distribution using hyperspectral unmixing approach. Extensive experiments on the hyperspectral dataset demonstrate that the material properties contribute to moving object tracking especially in background cluster, low-resolution, object rotation and deformation scenes, confirming that material information in HSIs has great potential in object tracking. In our future work, we will develop a material based convolutional neural network to investigate deep spectral-spatial material information for object tracking.

TABLE VI
RUNNING TIME COMPARISON (FPS)

| Videos                  | MHT | C-COT [64] | MCCT [83] | CPNet [67] | ECO [65] | DeepSRDCF [70] | SRDCF [8] | BACF [55] |
|-------------------------|-----|------------|-----------|------------|----------|----------------|-----------|-----------|
| RGB                     | n/a | 0.319      | 8.241     | 2.772      | 1.737    | 2.294          | 8.401     | 27.134    |
| Hyperspectral/False-Color| 2.612| 0.307      | 8.543     | 3.733      | 0.897    | 2.246          | 9.489     | 34.210    |

Fig. 11. Qualitative evaluation on 4 video sequences (i.e., rubik, drive, worker, forest).

TABLE VII
AUCS OF EXTENDED MATERIAL BASED TRACKERS

| Trackers | RGB   | False-Color | Hyperspectral |
|----------|-------|-------------|---------------|
| ECO      | 0.586 | 0.587       | 0.600         |
| KCF      | 0.386 | 0.352       | 0.393         |

REFERENCES
[1] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2544–2550.
[2] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, “Adaptive color attributes for real-time visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1090–1097.
[3] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596, Mar. 2015.
