Spectral-spatial-temporal Modelling for Hyperspectral Object Tracking

Jun Zhou

Griffith University, 170 Kessels Road, Nathan, Queensland 4111, Australia

Abstract Object tracking is a fundamental task in computer vision, with applications spanning surveillance, autonomous driving, environmental monitoring, and robotics [1]. While traditional tracking methods based on color videos have achieved remarkable success, they remain limited in handling complex environments, such as cluttered backgrounds, significant object deformations, or targets with similar colors and textures. Hyperspectral video tracking addresses these challenges by capturing rich spectral signatures in addition to spatial and temporal information [2]. This joint representation allows trackers to distinguish objects not only by their visual appearance but also by their intrinsic material properties. In this talk, I will begin by reviewing traditional spectral–spatial analysis methods for hyperspectral data processing. I will then introduce recent advances in hyperspectral video dataset construction and spectral–spatial–temporal modeling for object tracking. Finally, I will conclude with a discussion of emerging research directions and their potential impact on real-world applications.

Spectral-spatial Analysis of Hyperspectral Data: Traditional hyperspectral image processing often relies on spectral or combined spectral–spatial analysis to extract features for tasks such as image classification and object detection. Two widely used approaches are band selection [3], which identifies the most informative spectral bands for a given application, and hyperspectral unmixing [4], which estimates both the material spectra in a scene and how they are distributed across the image. Unmixing is typically based on a linear mixture model, where the image is decomposed into an endmember matrix (capturing the spectral signature of each material) and an abundance matrix (indicating the proportion of each material at every pixel) [5]. To make the results physically meaningful and mathematically well-defined, additional constraints such as sparsity [6] and smoothness [7] are often included in the reconstruction process. These models can be solved using iterative optimization methods or reformulated as deep neural networks that mimic the optimization steps [8]. The resulting abundance maps provide rich spatial information, which can be combined with other features from traditional or deep learning techniques to support downstream computer vision tasks such as object detection, boundary detection, and image classification [9].

Object Tracking in Hyperspectral Videos: With advances in sensing technology, snapshot hyperspectral cameras now make it possible to capture hyperspectral videos in real time. This has enabled the creation of hyperspectral video datasets and stimulated research on spectral–spatial–temporal analysis methods. The first high-framerate hyperspectral video dataset was introduced in 2020 [2], and has since been extended into larger benchmark datasets used in hyperspectral object tracking challenges [10]. In a typical hyperspectral object tracking task, the target is specified by a bounding box in the first frame. The tracking process then proceeds through several stages: template initialization, feature extraction, target proposal generation with a tracking model, object detection, and model update. Throughout this pipeline, the joint modeling of spectral, spatial, and temporal information is essential for achieving accurate and robust tracking performance.

Spectral-spatial-temporal video tracking: Most trackers work by comparing features extracted from the first frame with those from subsequent frames to estimate the location and size of the target. However, since hyperspectral images contain many more bands than conventional color images, it is not straightforward to directly adopt powerful color trackers trained on large video datasets for feature extraction and representation learning. To address this issue, a hyperspectral frame can be divided into band groups, which are converted into false-color images based on band importance estimated using an attention mechanism [3]. Pretrained deep color trackers can then be applied to each false-color image in parallel for object detection and tracking.

These parallel deep neural networks collectively form a Siamese fusion network [11]. Each branch extracts features from different layers and band groups, producing a multi-scale, multi-level spectral–spatial representation of the target. Through feature fusion, the network captures both global and local structures, enabling it to model spatial variations as well as changes in material appearance for adaptive online tracking. Within the same Siamese framework, the feature extraction and fusion stages can also be replaced by modern object detection backbones

STRL'25: Fourth International Workshop on Spatio-Temporal Reasoning and Learning, 16 August 2025, Montreal, Canada

jun.zhou@griffith.edu.au (J. Zhou)
https://experts.griffith.edu.au/7205-jun-zhou (J. Zhou)
0000-0001-5822-8233 (J. Zhou)

such as YOLO [12], which have shown excellent performance in complex tracking scenarios. Building on the detected objects, the tracker can be further enhanced with a classifier and a temporal network based on gated recurrent units (GRUs). The classifier helps distinguish between visually similar objects, while the temporal network models frame-to-frame dependencies, improving robustness against challenges such as occlusion and scale variations.

Hyperspectral video datasets captured by different cameras often vary in band numbers and wavelength ranges. For example, the 2025 Hyperspectral Object Tracking contest [10] provides more than 200 videos collected across the visible, near-infrared, and red–NIR ranges. However, hyperspectral trackers developed for one dataset cannot be directly adapted to such multi-modal data. To address this issue, Islam et al. [13] proposed an adaptive band selection strategy combined with a multimodel ensemble approach. Their method begins with a local–global attention-based band selection module that identifies the three most informative bands from any dataset. By doing so, the selected bands become independent of the original number of spectral channels supported by the camera, allowing a single model to handle hyperspectral videos with diverse configurations. Finally, a multimodel ensemble framework refines tracking by selecting the optimal candidate proposals for the target. This is achieved by comparing the similarity between proposals generated by base models and the target's appearance in historical frames.

Future Research Topics: Although hyperspectral video tracking has attracted growing attention since 2020, research in this area remains relatively limited. To advance the field, larger datasets need to be collected and annotated, ideally approaching the scale of existing color video datasets. In this regard, generative AI techniques offer promising opportunities to augment training data. In addition, most existing approaches rely heavily on detection-based tracking, without fully leveraging the temporal information embedded in hyperspectral videos. Future work should investigate advanced temporal modeling techniques, such as optical flow and sequence learning, to better capture object dynamics and scene dependencies across frames. Finally, the application scope of hyperspectral video tracking should be broadened to domains such as agriculture, environmental monitoring, healthcare, and consumer products, thereby demonstrating its real-world impact and societal value.

References

[1] D. A. Forsyth, J. Ponce, Computer Vision - A Modern Approach, Edition 2, Pearson Higher Ed, 2015.

[2] F. Xiong, J. Zhou, Y. Qian, Material based object tracking in hyperspectral videos, IEEE Transactions on Image Processing 29 (2020) 3719–3733.

[3] J. Wang, J. Zhou, W. Huang, Attend in bands: Hyperspectral band weighting and selection for image classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12 (2019) 4712–4727.

[4] P. Ghosh, S. K. Roy, B. Koirala, B. Rasti, P. Scheunders, Hyperspectral unmixing using transformer network, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–16.

[5] X.-R. Feng, H.-C. Li, R. Wang, Q. Du, X. Jia, A. Plaza, Hyperspectral unmixing based on nonnegative matrix factorization: A comprehensive review, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 4414–4436.

[6] Y. Qian, S. Jia, J. Zhou, A. Robles-Kelly, Hyperspectral unmixing via $l_{1/2}$ sparsity-constrained nonnegative matrix factorization, IEEE Transactions on Geoscience and Remote Sensing 49 (2011) 4282–4297.

[7] X. Liu, W. Xia, B. Wang, L. Zhang, An approach based on constrained nonnegative matrix factorization to unmix hyperspectral data, IEEE Transactions on Geoscience and Remote Sensing 49 (2011) 757–772.

[8] F. Xiong, J. Zhou, S. Tao, J. Lu, Y. Qian, SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–16.

[9] J. Liang, J. Zhou, L. Tong, X. Bai, B. Wang, Material based salient object detection from hyperspectral images, Pattern Recognition 76 (2018) 476–490.

[10] 2025. https://www.hsitracking.com/.

[11] Z. Li, F. Xiong, J. Lu, J. Zhou, Y. Qian, Material-guided siamese fusion network for hyperspectral object tracking, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, pp. 2809–2813.

[12] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, p. 7464–7475.

[13] M. A. Islam, J. Zhou, W. Xing, Y. Gao, K. K. Paliwal, Ubstrack: Unified band selection and multimodel ensemble for hyperspectral object tracking, IEEE Transactions on Geoscience and Remote Sensing 63 (2025) 1–15.