SCIENCE CHINA
Information Sciences

• LETTER •

January 2023, Vol. 66 119103:1–119103:2
https://doi.org/10.1007/s11432-021-3518-y

Learning cross-modal interaction for RGB-T tracking

Chunyan XU, Zhen CUI*, Chaoqun WANG, Chuanwei ZHOU & Jian YANG

Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094*, China*

Received 12 August 2021/Revised 26 February 2022/Accepted 7 June 2022/Published online 1 November 2022

Citation Xu C Y, Cui Z, Wang C Q, et al. Learning cross-modal interaction for RGB-T tracking. Sci China Inf Sci, 2023, 66(1): 119103, https://doi.org/10.1007/s11432-021-3518-y

Dear editor,

Visual object tracking, which has attracted increasing attention in the field of general visual understanding, aims to track each temporally changing object in a video sequence, with the target specified only in the first frame. Although most tracking algorithms have facilitated significant advances in RGB video sequences, object tracking using only RGB information is unreliable under extreme lighting conditions (e.g., dark night, rain, and foggy). With the development of hardware devices, infrared cameras have been widely used to capture the contour information of objects by sensing the radiation intensity of their surfaces, which effectively compensates for RGB images for identifying objects. Therefore, the integration of complementary data from visual and thermal infrared spectrum is essential for dealing with appearance changes and background distractions in the RGB-T object tracking problem. According to the different fusion methods of multimodal data, previous RGB-T tracking algorithms can be divided into three categories: modality weight learning for multimodal trackers [1], sparse representation-based multimodal trackers [2], and RGB-T tracking based on deep learning [3, 4] (Appendix A). Although the above RGB-T tracking approaches have achieved promising performance, several challenges remain regarding the utilization of the interactive characteristics between heterogeneous modalities to solve the RGB-T tracking problem.

In this study, we propose a cross-modal interaction (CMI) learning framework for the RGB-T tracking task, which explores the interrelated characteristics of the two modalities to promote heterogeneous information fusion. The CMI architecture is composed of a basic feature encoding/extraction network, a cross-modal interaction module, and a binary classification network (Figure 1(a)). According to the object tracking state in the previous frame, we sampled multiple region samples from the input RGB and thermal image pairs, and then adopted a pre-trained CNN for extracting hierarchical representations of all region proposals. Inspired by the recent transformer mechanism, we introduce two specially designed interaction learning modules from the pixel-/relation-level cross-modal feature representation, which can be seamlessly integrated in our RGB-T tracking network. Specifically, the pixel-level CMI approach mainly considers pixel-pairwise affinities in the spatial domain, whereas the relation-level CMI method employs these pattern associations between the two modalities. The built cross-modal correlations can be used to guide the propagation of these convolutional features from one modality to another during the process of multimodal fusion. These fused features of RGB and thermal data can be aggregated and then fed into the binary classification network. The binary classification network has three fully connected layers followed by the cross-entropy loss function to estimate the possibilities of proposal regions as the background or foreground, and the top-k confident candidates can be used to regress the final location of the tracking object.

This study aims to learn the cross-modal interaction in the feature-learning process to boost the performance of the RGB-T tracking problem. Given a pair of samples in RGB and thermal modalities (recorded as $I_{\mathcal{R}}$ and $I_{\mathcal{T}}$), we can choose CNNs to extract deep feature representations. By mining these cross-modal interactions/correlations between two modalities, we can then perform cross-modal information propagation from one modality to another modality to enhance the feature representation of tracking objects. Formally,

$$\begin{aligned} &\mathcal{F}_{\mathcal{R}} = \mathrm{Conv}(I_{\mathcal{R}}), \mathcal{F}_{\mathcal{T}} = \mathrm{Conv}(I_{\mathcal{T}}),\\ &\mathcal{W}_{\mathrm{corr}} = f_{\mathrm{corr}}(\mathcal{F}_{\mathcal{R}}, \mathcal{F}_{\mathcal{T}}),\\ &\mathcal{F}_{\mathcal{T}\to\mathcal{R}} = f_{\mathcal{T}\to\mathcal{R}}(\mathcal{W}_{\mathrm{corr}}, \mathcal{F}_{\mathcal{T}}),\\ &\mathcal{F}'_{\mathcal{R}} = \psi(\mathcal{F}_{\mathcal{R}}, \lambda_{\mathcal{R}}\mathcal{F}_{\mathcal{T}\to\mathcal{R}}), \end{aligned} \tag{1}$$

where $\mathcal{F}_{\mathcal{R}}$ and $\mathcal{F}_{\mathcal{T}}$ denote convolutional features of the input RGB and thermal images, respectively. The function f_{corr} is used to explore the intra-modal correlation in the spatial domain or the cross-modal interaction between two modalities (Figure 1(b)). With the guidance of the cross-modal correlation $\mathcal{W}_{\mathrm{corr}}$, the function $f_{\mathcal{T}\to\mathcal{R}}$ is adopted to perform feature propagation from the thermal modality to the RGB modality. Finally, we obtain the enhanced feature of the RGB modality (i.e., $\mathcal{F}'_{\mathcal{R}}$), where the hyperparameter $\lambda_{\mathcal{R}}$ refers to a learned balance factor and ψ denotes an aggregated concat function by considering the original feature (i.e., $\mathcal{F}_{\mathcal{R}}$) and the propagated feature from another modal-

* Corresponding author (email: zhen.cui@njust.edu.cn)

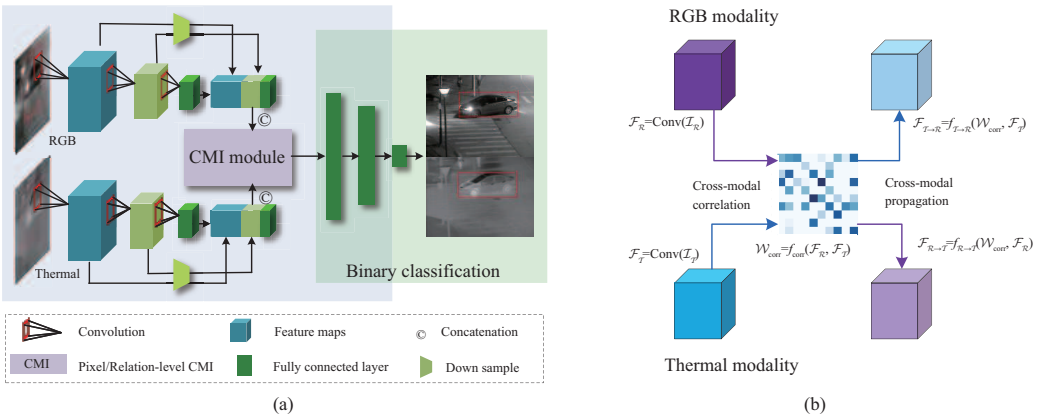

Figure 1 (Color online) Illustration of our proposed CMI network architecture. Given the input RGB and thermal image pair, we first adopt a classic convolutional network to extract multi-scale feature representation. In the CMI module, we can mine these cross-modal correlations between two heterogeneous modalities, which can be then employed to promote/guide the propagation from one modality to another modality. These fused features of RGB and thermal data can be then fed into the binary classification network. Finally, we can address the RGB-T tracking task in an end-to-end network. (a) Network architecture; (b) feature propagation in the CMI module.

ity (i.e., $\mathcal{F}_{\mathcal{T}\to\mathcal{R}}$). Similarly, we can perform cross-modal interaction learning among different modalities to obtain a robust feature representation for a specific object.

By employing the convolutional features of the two modalities $\{\mathcal{F}_{\mathcal{R}}, \mathcal{F}_{\mathcal{T}}\}$, we can first obtain the pixel-pairwise correlation of each modality at the feature level:

$$\begin{aligned} \mathcal{W}_{\mathcal{R}}^{\text{pixel}} &= \sigma(Q_{\mathcal{R}} K_{\mathcal{R}}^{\mathrm{T}}/\sqrt{d_{\mathcal{R}}}), \\ \mathcal{W}_{\mathcal{T}}^{\text{pixel}} &= \sigma(Q_{\mathcal{T}} K_{\mathcal{T}}^{\mathrm{T}}/\sqrt{d_{\mathcal{T}}}), \end{aligned} \tag{2}$$

where $\mathcal{W}_{\mathcal{R}}^{\text{pixel}}$ and $\mathcal{W}_{\mathcal{T}}^{\text{pixel}}$ denote the pixel-pairwise correlations for the RGB and thermal modality, respectively. $Q_{\mathcal{R}} \in \mathbb{R}^{(wh)\times d_{\mathcal{R}}}$ and $K_{\mathcal{R}} \in \mathbb{R}^{(wh)\times d_{\mathcal{R}}}$ are the reshaped feature matrices of convolutional maps. w, h and $d_{\mathcal{R}}$ are the width, height, and channel of convolutional feature map, respectively. $Q_{\mathcal{R}}$ and $K_{\mathcal{R}}$ can be learned from $\mathcal{F}_{\mathcal{R}}$ though one or more 1×1 convolutional layers; σ refers to a standard normalization (e.g., softmax function). Similarly, we can learn the feature representations $Q_{\mathcal{T}}$ and $K_{\mathcal{T}}$ of the thermal image and then obtain the corresponding pixel-pairwise spatial affinity $\mathcal{W}_{\mathcal{T}}^{\text{pixel}}$. By employing pixel-pairwise correlation in the spatial domain, we can then perform feature propagation across different modalities to complete pixel-level fusion. Second, we consider the relation-level interaction between RGB and thermal modality from a high semantic perspective, which can be formulated as

$$\mathcal{W}_{(\mathcal{R},\mathcal{T})}^{\text{relation}} = \sigma(\mathcal{W}_{\mathcal{T}}^{\text{pixel}} \odot \mathcal{W}_{\mathcal{R}}^{\text{pixel}}), \tag{3}$$

where $\odot$ refers to the dot product operation of the two matrices. The relation-level CMI is mined from the pixel-pairwise correlations for the RGB and thermal modalities, where $\mathcal{W}_{\mathcal{R}}^{\text{pixel}}$ and $\mathcal{W}_{\mathcal{T}}^{\text{pixel}}$ belong to the low-order pattern correlations in the pixel level, and then are employed to infer a high-order pattern interaction (i.e., $\mathcal{W}_{(\mathcal{R},\mathcal{T})}^{\text{relation}}$) between heterogeneous modalities in the semantic level. Under the guidance of relation-level interactions, we can promote the transmission of information from one modality to another modality. Finally, we perform feature propagation across different modalities to complete the cross-modal fusion, which is then fed into the subsequent binary classification network for the RGB-T tracking task.

Experiments. Comprehensive evaluations of the GTOT [2], RGBT234 [1], and VOT-RGBT2019 [5] datasets showed that our CMI performs better than the compared baselines (see Appendix B). The performance improvements may be primarily due to the complementarity between the heterogeneous data sources, which allows the moving objects to be well located during the tracking process in the RGB-T video sequences. The component-wise analysis demonstrates that by accounting for the multi-level cross-modal interactions, the complementary characteristics between RGB and thermal images can be well captured in the information fusion process. Through a qualitative comparison, we can observe that cross-modal fusion can help capture the effective representation of tracking objects and generate more precise predictions in challenging video sequences. Both the tracking speed and robustness under the modality-missing situation are superior when using our CMI learning mechanism. Therefore, extensive experimental results clearly demonstrate the effectiveness of the proposed CMI method in solving the RGB-T tracking problem.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61972204, 62072244) and in part by Natural Science Foundation of Jiangsu Province (Grant Nos. BK20191283, BK20190019).

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Li C, Liang X, Lu Y, et al. RGB-T object tracking: benchmark and baseline. Pattern Recogn, 2019, 96: 106977
2 Li C, Cheng H, Hu S, et al. Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Trans Image Process, 2016, 25: 5743–5756
3 Yu Y, Xiong Y, Huang W, et al. Deformable Siamese attention networks for visual object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 6728–6737
4 Wang C, Xu C, Cui Z, et al. Cross-modal pattern-propagation for RGB-T tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 7062–7071
5 Kristan M, Matas J, Leonardis A, et al. The seventh visual object tracking VOT2019 challenge results. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019