Signal Processing: Image Communication 68 (2018) 207–217

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Two-stage modality-graphs regularized manifold ranking for RGB-T tracking

Chenglong Li [a,b], Chengli Zhu [a], Shaofei Zheng [a], Bin Luo [a], Jing Tang [a,*]

[a] School of Computer Science and Technology, Anhui University, Hefei, China
[b] Institute of Physical Science and Information Technology, Anhui University, Hefei, China

ARTICLE INFO

Keywords:
Visual tracking
Two-stage scheme
Modality-graphs
Manifold ranking
Thermal information

ABSTRACT

This paper proposes a two-stage modality-graphs regularized manifold ranking algorithm to learn a robust object representation for RGB-Thermal tracking. The bounding box of the tracked object is represented with a set of patches, which are described by RGB-thermal features. We assign each patch with a weight to specify its importance in describing the object, and also each modality with a weight to reflect modal reliability. These weights are jointly optimized via manifold ranking on the modality-graphs with patches as nodes. Moreover, we develop a two-stage ranking strategy to mitigate the effects of inaccurate patch weights initialization. The object representation is then updated by imposing the modality weights and the patch weights on the extracted patch features, and the object location is finally predicted by adopting the structured SVM. Extensive experiments on the standard benchmark dataset GTOT suggest that the proposed tracker outperforms several state-of-the-art RGB-T tracking methods.

1. Introduction

RGB-T object tracking is to estimate the states of the target in RGB-Thermal videos, i.e., utilizing both RGB and thermal information for state estimation of the target object. It can leverage the complementary benefits [1–3] of different source data for robust tracking in challenging scenarios, and thus plays a critical role in numerous vision applications, such as self-driving cars, robotics and visual surveillance systems. The related research is limited until a comprehensive RGB-T tracking benchmark, GTOT [3], is proposed recently to provide a evaluation platform for this direction.

Recent works on RGB-T tracking mainly focus on sparse representation because of its capability of suppressing noises and errors [4–6,3]. One direct strategy is to concatenate the image patches from RGB and thermal sources into a feature vector to represent the object, and then employ the conventional sparse representation methods to achieve RGB-T tracking [4]. Further, some works [5,6] pursued a multi-task joint sparse representation on both RGB and thermal modalities and fused the resultant tracking results using the sparse representation coefficients [5] or reconstruction residues [6]. Recent work proposed by Li et al. [3] employed the collaborative model to adaptively fuse RGB and thermal modalities by assigning each modality with a reliability weight in the sparse representation.

In this paper, we propose a general algorithm, called two-stage modality-graphs regularized manifold ranking, to learn a more robust object representation for RGB-T tracking. The pipeline of our work is shown in Fig. 1. Given the object bounding box, we first partition it into a set of non-overlapping patches, which are described with RGB and thermal features. These patch features are concatenated into a feature vector to represent the object for dealing with object deformations and partial occlusions. Then, we assign each patch with a weight to suppress background information in representing the object, and each modality with a weight to exploit different source data adaptively, and integrate these weights into the object features to construct a robust object representation. In particular, we model the patch weights and the modality weights in a joint manner, and optimize them efficiently with closed-form solutions. To improve the robustness of patch weight, we apply two-stage ranking strategy. At the first stage, the patch weight is computed based on the initial seeds. And at the second stage, the patch weight computation is based on the result of first stage. The object location is finally predicted by applying the structured Support Vector Machine (SVM).

The major contributions of this paper can be summarized in three aspects.

[*] Corresponding author.
E-mail address: jtang99029@foxmail.com (J. Tang).

https://doi.org/10.1016/j.image.2018.08.004
Received 15 May 2018; Received in revised form 6 August 2018; Accepted 8 August 2018
Available online 30 August 2018

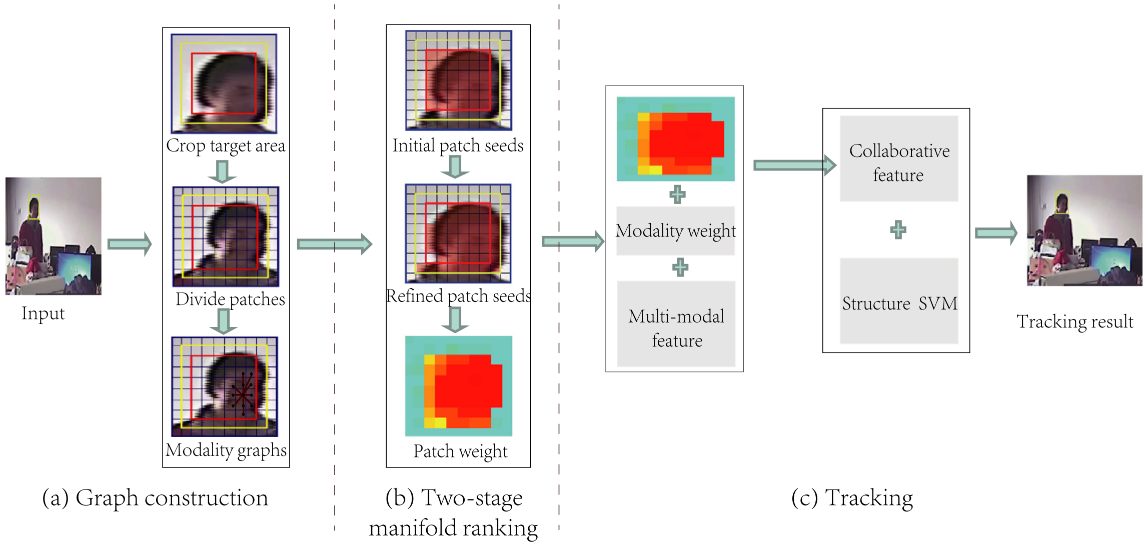

Fig. 1. The pipeline of our work is detailed in this figure.

- We propose a novel two-stage modality-graphs regularized manifold ranking model for RGB-T object tracking. The modality-graphs is better to capture the intrinsic relationship among patches. The proposed model collaboratively and adaptively leverages multiple source data to compute the patch weights and the modality weights in a joint manner, which are utilized to construct a robust object representation for RGB-T object tracking.
- We present an alternating algorithm to optimize the patch weights and the modality weights, each subproblem associating with closed-form solutions. The theoretical proof and the empirical evidence on the standard benchmark GTOT [3] presented in the experimental section suggests that the proposed algorithm has strong and stable convergence behavior.
- We introduce a two-stage scheme to improve the robustness of weight computation and object tracking. To demonstrate the efficiency and superior performance of the proposed method over other state-of-the-art RGB-T tracking methods, extensive experiments on the standard benchmark GTOT are conducted. We also analyze the performance of our main components to justify their respective contributions.

2. Related work

This section first reviews the relevant visual tracking methods and then recent RGB-T object tracking methods.

Constructing or learning good object features are important for visual tracking [7–11]. Song [8] introduced the entropy as a judgment to select informative features for robust visual tracking, and Song et al. [7] proposed to learn self-similarity information among the local features extracted from the different regions. A simple yet effective Boolean map based representation was proposed by Zhang et al. [9] to effectively encode multi-scale connectivity cues, which are able to capture fine-grained structural details and coarse-grained global shape information. Zhang et al. [10] proposed a simple two-layer convolutional networks to learn a robust representations for visual tracking but without online training, and the proposed tracker is thus efficient and effective.

However, bounding boxes sometimes cannot represent target objects well when shapes of objects are irregular or tracking results contain many noises. Assigning weights to different pixels or patches in feature representations has been proved to be an effective way for boosting tracking performance [12–17,7,18]. Comaniciu et al. [12] employed the kernel-based method to assign smaller weights to boundary pixels during the histogram construction. He et al. [13] also assumed that pixels far from a box center should be less important. These methods may fail when a target object has a complicated shape or is occluded. Some works [14,15] integrated segmentation results into tracking to alleviate the effects of background. These algorithms, however, are sensitive to segmentation results. Kim et al. [16] developed a random walk restart algorithm on 8-neighbor graph to compute patch weights within target object bounding box. To make the best use of the intrinsic relationship between patches, Li et al. [18] proposed a novel representation model to jointly learn the graph that infers the graph structure, edge weights and patch weights. Li et al. [19] proposed a dynamic graph learning algorithm based on the low-rank and sparse representation. These works motivate us to learn a robust object representation for RGB-T tracking.

RGB-T tracking has drawn a lot of attentions in the community with the popularity of thermal infrared sensors [3–6,20]. Recent works on RGB-T tracking mainly focus on sparse representation because of its capability of suppressing noises and errors [21]. Wu et al. [4] concatenated the image patches from RGB and thermal sources into a one-dimensional vector that was then sparsely represented in the target template space. Liu et al. [5] performed joint sparse representation calculation on both RGB and thermal modalities and fused the resultant tracking results using min operation on the sparse representation coefficients. Li et al. [6] proposed a Laplacian sparse representation to learn a multi-modal feature model that encodes both the spatial local information and occlusion handling. Li et al. [20] proposed a novel convolutional neural network architecture to achieve adaptive fusion of different source data for RGB-T tracking. The collaborative sparse representation based trackers were proposed by Li et al. [3] to adaptively fuse RGB and thermal modalities by assigning each modality with a reliability weight, and they jointly optimized the sparse coefficients and modality weights online to adapt modal qualities.

3. Modality-graphs regularized manifold ranking

Given the object bounding box, we first partition it into a set of non-overlapping patches, which are described with RGB and thermal features. These patch features are concatenated into a feature vector to represent the object for dealing with object deformations and partial occlusions. Then, we assign each patch with a weight to suppress background information in representing the object, and each modality with a weight to exploit different source data adaptively, and integrate these weights into the object features to construct a robust object representation. These weights are jointly optimized via manifold ranking on