

Multi-Adapter RGBT Tracking

Chenglong Li^{1,3}, Andong Lu¹, Aihua Zheng¹, Zhengzheng Tu¹, Jin Tang^{1,2}

¹Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education,
School of Computer Science and Technology, Anhui University, Hefei 230601, China

²Key Laboratory of Industrial Image Processing and Analysis of Anhui Province, Hefei 230601, China

³Institute of Physical Science and Information Technology, Anhui University, Hefei 230601, China

{lcl1314, adlu_ah}@foxmail.com, ahzheng214@qq.com, zhengzhengahu@163.com, tangji@ahu.edu.cn

Abstract

The task of RGBT tracking aims to take the complementary advantages from visible spectrum and thermal infrared data to achieve robust visual tracking, and receives more and more attention in recent years. Existing works focus on modality-specific information integration by introducing modality weights to achieve adaptive fusion or learning robust feature representations of different modalities. Although these methods could effectively deploy the modality-specific properties, they ignore the potential values of modality-shared cues as well as instance-aware information, which are crucial for effective fusion of different modalities in RGBT tracking. In this paper, we propose a novel Multi-Adapter convolutional Network (MANet) to jointly perform modality-shared, modality-specific and instance-aware feature learning in an end-to-end trained deep framework for RGBT tracking. We design three kinds of adapters within our network. In a specific, the general-ity adapter is to extract shared object representations, the modality adapter aims at encoding modality-specific information to deploy their complementary advantages, and the instance adapter is to model the appearance properties and temporal variations of a certain object. Moreover, to reduce computational complexity for real-time demand of visual tracking, we design a parallel structure of generic adapter and modality adapter. Extensive experiments on two RGBT tracking benchmark datasets demonstrate the outstanding performance of the proposed tracker against other state-of-the-art RGB and RGBT tracking algorithms.

1. Introduction

The problem of RGBT tracking could be considered as an extension of visual tracking, and its goal is to estimate target states using the complementary advantages of visible spectrum (called RGB in this paper) and thermal infrared information given the initial state in the first pair of frame. It has been receiving much more attention recently and becoming more and more popular partly due to the following reasons: i) RGB and thermal data have strong complementary advantages and thus could overcome imaging limitations of individual source [18, 30, 31, 21]. ii) Thermal infrared cameras are economically available in recent years [6], making RGBT data easier to access in various applications, such as object segmentation [18], person Re-ID [30] and pedestrian detection [10, 31]. iii) Recent RGBT tracking benchmark datasets [16, 21] provide a flexible evaluation platform of various RGBT trackers. iv) The VOT2019 challenge has announced “VOT-RGBT challenge” to address short-term trackers that use RGB and thermal infrared modalities¹. Although much progress has been achieved, how to make the best of RGB and thermal information for robust RGBT tracking is an open problem.

Existing works focus on modality-specific information integration from two major aspects. One is to introduce modality weights that reflect their reliabilities in tracking prediction to achieve adaptive fusion of different modalities. For example, Li et al. [16] integrate computation of modality weights and sparse representation in a joint model and perform online object tracking in Bayesian filtering framework. Lan et al. [15] learn modality weights and classifiers of different modalities in a max-margin learning framework. Another is to learn robust feature representations of different modalities. For example, Li et al. [21] propose to represent target object using a collaborative graph with local image patches as nodes. They learn a patch-based weighted RGBT features to fuse different modalities and

This research is jointly supported by the National Natural Science Foundation of China (No. 61702002, 61976003, 61976002, 61602006, 61872005), Natural Science Foundation of Anhui Province (1808085QF187), Open fund for Discipline Construction, Institute of Computer Science and Technology, Anhui University, and Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2018A0023).

¹<http://www.votchallenge.net/>

