Attention-based Cross-modality Interaction for Multispectral Pedestrian Detection

Tianshan Liu, Rui Zhao and Kin-Man Lam

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hong Kong, China

Email: tianshan.liu@connect.polyu.hk, rick10.zhao@connect.polyu.hk, enkmlam@polyu.edu.hk

Abstract—Multispectral pedestrian detection has attracted extensive attention, as paired RGB-thermal images can provide complementary patterns to deal with illumination changes in realistic scenarios. However, most of the existing deep-learning-based multispectral detectors extract features from RGB and thermal inputs separately, and fuse them by a simple concatenation operation. This fusion strategy is suboptimal, as undifferentiated concatenation for each region and feature channel may hamper the optimal selection of complementary features from the different modalities. To address this limitation, in this paper, we propose an attention-based cross-modality interaction (ACI) module, which aims to adaptively highlight and aggregate the discriminative regions and channels of the feature maps from RGB and thermal images. The proposed ACI module is deployed into multiple layers of a two-branch-based deep architecture, to capture the cross-modal interactions from diverse semantic levels, for illumination-invariant pedestrian detection. Experimental results on the public KAIST multispectral pedestrian benchmark show that the proposed method achieves state-of-the-art detection performance.

Index Terms—Multispectral pedestrian detection, attention mechanism, cross-modal representation

I. INTRODUCTION

In order to highlight and aggregate the discriminative regions and channels of the feature maps from the RGB and thermal modalities, we propose an Attention-based Cross-modality Interaction (ACI) module for multispectral pedestrian detection. The ACI module is designed based on a two-stage strategy. In the first stage, we employ a spatial-wise attention mechanism to generate modality-specific gates, which assign different importance weights for each modality at each position. In the second stage, we compute two channel-attended vectors via a channel-wise attention mechanism, and utilize them to aggregate the spatially gated feature maps obtained in the first stage. The whole multispectral pedestrian detection network consists of two sub-branches, which take RGB images and thermal images as input, respectively. Each branch is built based on a Region Proposal Network (RPN)-based model, i.e., Faster R-CNN [1]. The overall network architecture is illustrated in Fig. 1.

II. EXPERIMENTAL RESULTS

The proposed method is compared with several state-of-the-art detectors on the public KAIST multispectral pedestrian data set [2]. The curves of MR against FPPI are shown in Fig. 2. We can find that the proposed model outperforms the other state-of-the-art methods, in all of the all-day, daytime and nighttime evaluation settings.

REFERENCES
