

Multi-Scale Body-Part Mask Guided Attention for Person Re-identification

Honglong Cai Zhiguan Wang Jinxing Cheng
Suning R&D Center USA

fhongl ong. cai , dori s. wang, j i m. chengG@ussuni ng. com

Abstract

Person re-identification becomes a more and more important task due to its wide applications. In practice, person re-identification still remains challenging due to the variation of person pose, different lighting, occlusion, misalignment, background clutter, etc. In this paper, we propose a **multi-scale body-part mask guided attention network** (MMGA), which jointly learns whole-body and part-body attention to help extract global and local features simultaneously. In MMGA, **body-part masks** are used to guide the training of corresponding attention. Experiments show that our proposed method can reduce the negative influence of variation of person pose, misalignment and background clutter. Our method achieves rank-1/mAP of 95.0%/87.2% on the Market1501 dataset, 89.5%/78.1% on the DukeMTMC-reID dataset, outperforming current state-of-the-art methods.

1. Introduction

Person re-identification (re-ID) aims at identifying the presence of same person in different cameras with different backgrounds, poses and positions. It is still a challenging task due to large variations on persons like pose, occlusion, clothes, background clutter and detection failure which are shown in Figure 1.

Low-level features like colors, shapes, contours and local descriptors are used to train traditional re-ID models with low accuracy [8, 12]. Nowadays, with the fast development of deep neural networks, deep features of human image learned through convolutional neural network (CNN) is demonstrated to have better discrimination and robustness to represent the image, which has made significant improvement on the re-ID problem [6, 24, 26, 33]. The features learned from deep learning network should capture the most salient clues that can represent identities of different persons. However, most of the existing deep learning methods learn features from the whole image that contains not only the human body parts, but also the background regions [40, 37]. The background regions containing clutter

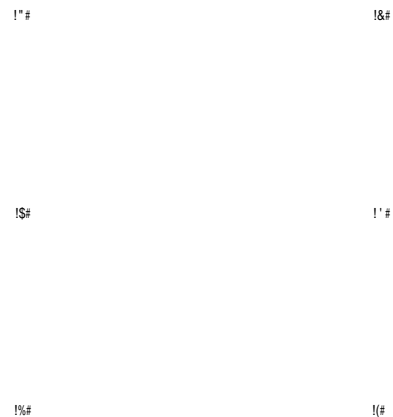


Figure 1. Examples of challenges in person re-identification and how our attention mechanism can handle the challenges. (a-b) occlusions, (c-d) inaccurate bounding-box detection, (e-f) variation of pose. The second to forth images in each group are the global attention map in original image, upper-body attention map and bottom-body attention map generated by our proposed MMGA network.

and occlusions may lead to a misalignment problem. To address this issue, some recent works [31, 38, 45, 6, 34, 18] show that locating the significant body parts and learning the discriminative features from these informative regions can reduce the negative effects of clutter and occlusions, and thus improve the re-ID accuracy.

Visual attention has shown its success in re-ID tasks [42, 22, 29, 19], as the mechanism conforms to the human visual system that a whole image is not likely to be processed in its entirety at once, but only the salient parts of the whole visual space are focused when and where needed. Visual attention module can help to extract dynamic features from salient parts mostly like human body parts in a image by guiding the learning towards informative image regions [29]. Given the human body information, attention maps
