

Fusing Two Directions in Cross-domain Adaption for Real Life Person Search by Language

Kai Niu^{1,3} Yan Huang^{1,4} Liang Wang^{1,2,3,4}

¹Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)

²Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)

³University of Chinese Academy of Sciences (UCAS)

⁴Artificial Intelligence Research, Chinese Academy of Sciences (CAS-AIR)

kai.niu@cripac.ia.ac.cn {yhuang, wangliang}@nlpr.ia.ac.cn

Abstract

Person search by language is an important application in video surveillance. The existing huge visual-semantic discrepancy and the cross-domain difficulty of emerging pedestrian images with new identities while no language description for training in real life application make this problem non-trivial to be addressed. In this paper, we first propose a concise and effective framework for image-sentence alignment to deal with the visual-semantic discrepancy. Second, we innovatively **fuse the two opposite directions**, i.e., source target and target source, for cross-domain adaption. Extensive experiments have validated the significant superiority of the proposed method on both source domain and target domain, and we have obtained the state-of-the-art performance and won the 1st place in competition.

1. Introduction

Person search by language [10, 12] is an important task in intelligent surveillance [4], which requires discriminative cross-modal representations to distinguish different people. It is difficult to directly measure the similarities between images and natural language descriptions due to the existing huge visual-semantic discrepancy. And in many practical situations, the problem that the **training data and testing data are in different domains** makes it even harder to accurately search for the matched person.

Although much progress [3] has been achieved for matching images and sentences accurately, it is still non-trivial to address the problem of person search by language, due to the less discriminative situation among images of different pedestrians. Specifically, different from the conventional image-sentence matching problem with images hav-

ing various topics, scenes and styles, all images in the problem of person search by language belong to the pedestrian category, only having fine-grained differences and being much harder to distinguish. Therefore, we have to consider the characteristic of person search rather than only depending on the cross-modal matching approaches. Many solutions [17, 15] to the problem of image-based person search employ the pedestrian identities for classification, which can obtain more discriminative features for distinguishing pedestrians. And appropriately combining the objectives for person search and image-sentence matching may contribute to better visual-semantic embeddings further.

Beyond the visual-semantic discrepancy, there is another problem that the newly emerging pedestrian images have new identities but no language description for training, i.e., **cross-domain difficulty in real life application**. To address this problem, domain adaption is necessary for narrowing the cross-domain gap. Many effective solutions [16, 14] have focused on transferring from the source domain to the target domain, but neglect the opposite direction. In fact, these two directions can contribute complementarily to the final fusion and obtain a model that better addresses the cross-domain problem.

In summary, this paper first introduces a general framework for dealing with the problem of person search by language, which considers identity classification as well as cross-modal matching for better visual-semantic embeddings. Second, a **Cross-domain Bi-directional Adaption** (CBA) method is proposed to alleviate the cross-domain difficulty by innovatively fusing the two opposite adaption directions, which facilitates the practical application of person search by language. Specifically, in the source target (S T) direction, the model is first trained to have pedestrian identity classification and cross-modal matching abilities in

