Attention-based Model with Attribute Classification for Cross-domain Person Re-identification

Simin Xu, Lingkun Luo, Shiqiang Hu
siminxu0613@sjtu.edu.cn, lolinkun@gmail,com, sqhu@sjtu.edu.cn
School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Person re-identification (re-ID) which aims to recognize a pedestrian observed by non-overlapping cameras is a challenging task due to high variance between images from different viewpoints. In this paper, we propose an attention-based model with attribute classification (AMAC) to facilitate a well trained model transferring across different data domains, which further enables an efficient cross-domain video-based person re-ID.

RELATED WORKS

Attention Model. The concept of attention model imitates the human perception scheme in which we tend to concentrate on discriminative local parts.

Attribute Learning. The attribute-semantic features can be considered as a complement of the main identity-discriminative features, thereby improving the accuracy of person re-ID.

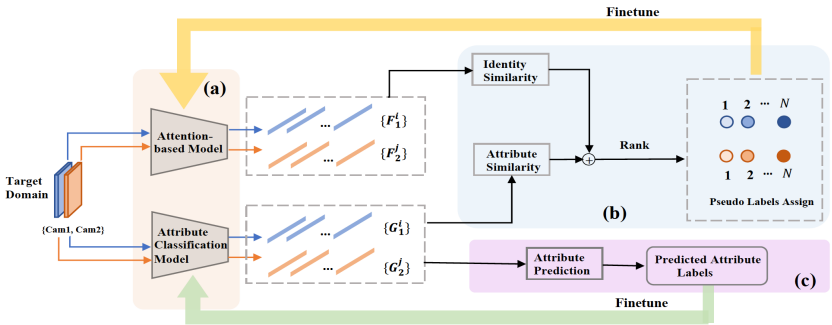
REFERENCES

[1] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," 2017.

[2] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," pp. 868–884, 2016.

METHODOLOGY

The figure below illustrates an overview of our network, where the input target video clips firstly go through the attention-based model pretrained on the source dataset and the attribute classification model pretrained on the PETA dataset simultaneously. Then we assign the pseudo labels to the target dataset according to feature similarities between all the images from two cameras. Finally, we finetune the two pretrained models according to the pseudo labels.



ATTENTION-BASED MODULE

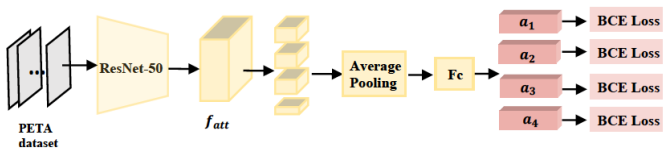
Given a pedestrian video $V = \{I_t\}_{\{t=1:T\}}$, where T is the number of frames and It denotes the t-th frame. Take the anchor set as an example, we denote $\mu^a = \{\mu_u^a(I_t), \mu_m^a(I_t), \mu_l^a(I_t)\}$ as the attention map and $\{f_u^a(I_t), f_m^a(I_t), f_l^a(I_t)\}$ as the middle representation of the triplet. Then the final representation of the anchor set can be denoted as $F_a = \{F_u^a, F_m^a, F_l^a\}$. We generate the final representation of each local part by the formulations below:

$$F_{part}^a = \sum_{t=1}^{T} \mu_{part}^a(I_t)\, f_{part}^a(I_t)$$

where $\mu_{part}^a(I_t)$ and $f_{part}^a(I_t)$ represent different parts (upper, middle, lower) scores and features, respectively.

ATTRIBUTE CLASSIFICATION MODULE

Since there are no available video-based datasets containing attribute labels, our proposed method pretrained a simple CNN model on the PETA attribute dataset and then use this model to extract attribute-semantic features of the target dataset.



CONCLUSION

In this work, we develop an attention-based model with attribute classification (AMAC) for joint learning identity discriminative features and attribute-semantic features under an unsupervised setting in order to alleviate the limitation of existing methods in real-word large-scale person reidentification.