ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

Based semantic parts for cross domain person re-identification

Hui Li

School of Computer Science and Technology, Qingdao University, Qingdao 266071, China.

Abstract

Cross-domain person re-identification is an important issue that restricts the application of person re-identification in practice, and the huge difference between the source domain and the target domain is the most critical factor that affects the generalization ability of the model. In response to this problem, we found that semantic components play an important role in cross-domain re-identification. The semantic analysis model is used to extract the features of multiple semantic components of persons, remove the complex and changeable background interference, and improve the cross-domain adaptive ability of the model. At the same time, the use of fine semantic component features to achieve the purpose of component alignment, but also improves the expressive ability of features, thereby improving the generalization ability of the model. Finally, we conducted a large number of cross-domain person re-identification verification experiments between the two person re-identification data sets of Market1501 and DukeMTMC-reID, which proved the effectiveness of our method.

Keywords

Cross-domain person re-ID, semantic alignment, local feature, global feature.

1. Introduction

In video surveillance, re-identification of people is a basic task, which serves person retrieval and cross-camera tracking. Due to its important application in security and surveillance, person re-identification (person re-ID) has always received extensive attention from academia and industry. Its purpose is to predict whether two images from different cameras belong to the same person.

Person re-ID by the same ID are affected by illumination, background, human posture changes, and camera angle changes. The feature re-ID between different views is significantly reduced. Therefore, the person re-ID technology starts from the global features based on deep convolutional neural networks. More fine-grained local features develop in the direction. Based on the different ways of dividing person's partial areas, the person component model can be divided into hard partition and soft partition. Hard segmentation does not require component labels, and is relatively simple. For example, Zhang [1] uses hard segmentation to extract local features of persons. In order to solve the problem of unaligned distance between local features caused by hard segmentation, dynamic programming is used to find the shortest path. To achieve the purpose of aligning local features, and adopt a mutual learning framework to enhance the effect of the model, but this hard partition alignment is still too rough and the error is large. Starting from the problem of excessive background and missing parts in the image, Zheng [2] et al. used affine transformation to process the image to achieve the purpose of image relative alignment, but the effect was not particularly ideal. Hard partitioning can only be applied to spatially aligned person images without occlusion. For person images with excessive posture changes and unaligned spatial distribution, the error is relatively large. Many soft segmentation methods are proposed based on the pose change problem. For example, Zhao [3] uses pose estimation to extract features of different semantic levels at different stages, and then merge the features of different semantic levels to align the features of the human body in different

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

images and enhance local details information presentation ability. Huang [4] and others based on PCB[5] implicitly align component features through the pose estimation model, and improve the generalization ability of the model through the segmentation model. But the model training is too complicated, which increases the training time. In order to reduce the training time, Li [6] et al. designed an attention convolutional neural network, which can simultaneously learn pixel-level global attention features and hard-region-level attention features within the bounding box, and re-identify feature representations. Maximize attention to the complementary information between selection and feature re-identification. Starting from pixel-level features, Huang [7] et al. obtained the segmentation labels of the re-ID data set by training additional component segmentation models, and then added segmentation heads to different re-ID models to generate segmentation loss, and incorporated component awareness into the re-ID model. Enhance the component awareness of the model.

The above soft division and hard division methods based on local features have achieved certain results, but there is a serious performance degradation in cross-data set testing. This is because the test identity has never been seen during training. The difference in data distribution also has a great impact. For example, persons in the Market1501 [8] data set usually wear shorts, while tops and pants frequently appear in the DukeMTMC-reID [9] data set. Moreover, it is expensive to establish a training set for each scene, and it is costly and infeasible to label all the images in the target data set. One of the most popular solutions is unsupervised domain adaptation (UDA).

At present, the common UDA has been extensively studied in image classification, target detection, face recognition and semantic segmentation. For example, Deng [10] transferred the image style of the source domain to the target domain, and kept the ID unchanged during the migration process to achieve the purpose of unsupervised domain adaptation of person re-ID. Some methods seek the optimal optimization algorithm through metric learning methods. For example, ECN [11] designs three loss functions through three invariances to optimize the cross-domain person re-ID model. However, these methods largely ignore the interference of the background. We propose a cross-domain person re-ID model based on semantic components. The semantic local features are extracted through the semantic analysis model to obtain more refined component features and achieve the purpose of semantic alignment. And remove the influence of the background, and improve the cross-domain adaptive ability.

Our contributions are as follows:

(1) We propose a cross-domain person re-ID framework based on semantic components, which reduces the problem of component misalignment caused by hardening points by extracting more refined person semantic features.

(2) By removing the background features, we reduce the influence of the background, improve the cross-domain adaptive ability and the generalization ability of the model.

(3) Our cross-domain experiment results on two person re-ID datasets, Market1501 and DukeMTMC-reID, prove the effectiveness of our method.

2. Relate Work

2.1 Section Headings

2.1.1 Semantic Component Features

Due to the problems of person image background clutter, pose diversity, and occlusion, such as PCB [5] through hardening points to extract the features of person components, there are often problems such as spatial dislocation and feature mismatch. Local feature matching problems, such as Zhao [3] extract features of different semantic levels at different stages through pose estimation, and then merge features of different semantic levels to align the features of the human body in different images and enhance the ability to express local details. Some methods use the attention mechanism to make the model pay attention to more expressive features. For example, Li [6] designed an attention

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

convolutional neural network that can simultaneously learn pixel-level global attention features and hard regions within bounding boxes. Level attention features, and maximize the complementary information between attention selection and feature recognition by re-identifying feature representations. Finally, many methods start with finer-grained dense semantic alignment to extract more precise component features. For example, Zhang [12] proposed the use of fine-grained semantics to solve the problem of person semantic misalignment for the first time. He designed a deep learning framework based on dense semantic alignment, in which dense semantic estimators were used to supervise the alignment of semantic features. The ability of person re-ID network to learn semantic feature alignment. However, the effective content of the images generated by the dense semantic estimator is very sparse, and there are estimation errors, especially low-resolution images. Because reid itself does not have dense semantics of labeling, it uses a dense model trained on the COCO-DesenPose dataset. However, these data sets have gaps in resolution, image quality, and pose distribution. And because of the removal of the background, some different content has also been removed, such as a red backpack. The above methods are either too complex or too rough to train, and no good solutions are proposed for the extraction of semantic components.

Due to the problems of person image background clutter, pose diversity, and occlusion, such as PCB [5] through hardening points to extract the features of person components, there are often problems such as spatial dislocation and feature mismatch. Local feature matching problems, such as Zhao [3] extract features of different semantic levels at different stages through pose estimation, and then merge features of different semantic levels to align the features of the human body in different images and enhance the ability to express local details. Some methods use the attention mechanism to make the model pay attention to more expressive features. For example, Li [6] designed an attention convolutional neural network that can simultaneously learn pixel-level global attention features and hard regions within bounding boxes. Level attention features, and maximize the complementary information between attention selection and feature recognition by re-identifying feature representations. Finally, many methods start with finer-grained dense semantic alignment to extract more precise component features. For example, Zhang [12] proposed the use of fine-grained semantics to solve the problem of person semantic misalignment for the first time. He designed a deep learning framework based on dense semantic alignment, in which dense semantic estimators were used to supervise the alignment of semantic features. The ability of person re-ID network to learn semantic feature alignment. However, the effective content of the images generated by the dense semantic estimator is very sparse, and there are estimation errors, especially low-resolution images. Because reid itself does not have dense semantics of labeling, it uses a dense model trained on the COCO-DesenPose dataset. However, these data sets have gaps in resolution, image quality, and pose distribution. And because of the removal of the background, some different content has also been removed, such as a red backpack. The above methods are either too complex or too rough to train, and no good solutions are proposed for the extraction of semantic components.

2.2 Domain Adaptation

Our work is closely related to UDA, that is, the data of the target domain is not labeled during the training process. Some methods try to solve this problem by reducing the difference between the source domain and the target domain. For example, CORAL [13] learned a linear transformation to align the mean and covariance of the feature distribution between two domains. Sun [14] proposed deep CORAL to extend the original method to deep neural networks and perform nonlinear transformations. The purpose of some other methods is to learn a transformation through the adversarial learning method to generate samples similar to the target domain. Recently, some studies have solved this problem by mapping the source data and target data to the same feature space for domain-invariant representation. For example, Ganin et al. [15] proposed a gradient reversal layer (GRL) and integrated it into a standard deep neural network to minimize the classification loss and maximize the domain aliasing loss. However, most of the existing unsupervised domain adaptation methods are based on the assumption that the class labels are the same across domains, and the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

personal identities of different re-ID data sets are completely different. Therefore, the above method cannot be directly used for personnel re-ID tasks.

3. Proposed Method

3.1 Overview

We propose a cross-domain person re-ID framework based on semantic components. First, we pretrain a deep neural network F(/θ) on the source domain, where θ represents the current network parameters as our initialization model, and then the network transfers to the image of the target domain for learning. We use ResNet50 [16] to extract global features, and use semantic component classifiers to extract features such as the head, upper body, lower body, and background of persons, and then optimize us by performing cross-entropy loss calculations on the features of the source and target domains. In order to eliminate background interference, we do not perform cross-entropy loss calculation on background features.

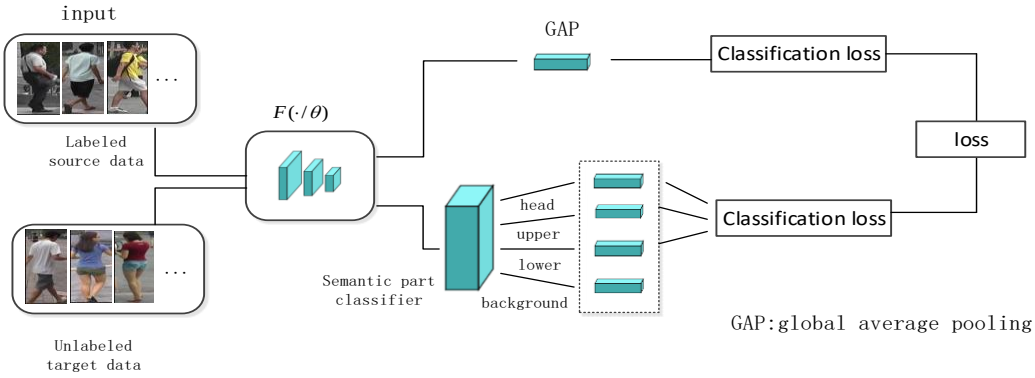

Fig. 1 Framework diagram of cross-domain person re-ID on semantic parts. The source domain and target domain images are input into the backbone network at the same time, and the global features are extracted separately, and the head, upper body, lower body and background features are extracted through the semantic component classifier, and the model is optimized by calculating the classification loss. The background feature does not calculate the classification loss.

3.2 Semantic Part Classifier

In order to extract fine semantic component features, we trained a semantic component classifier to divide the person image into four component confidence maps. Each confidence map represents a component feature. First, the global feature is output through the backbone mapping function (defined as f_θ) The mapping is as follows:

$$M_g^{c\times h\times w} = f_\theta(x_i) \tag{1}$$

Where θ is the backbone parameter, c, h, w are the channel, height and width.

Our component confidence map is defined as P, we usePh, Pu, Pl, Pb to represent the confidence of pixels (x, y) belonging to the head, upper body, lower body, and background parts, and then extract the feature maps of each part through the global feature map as follows:

$$M_k = P_k \circ M_g \tag{2}$$

Where ◦ means multiplying elements, and k refers to k parts.

3.3 Loss

We use cross-entropy loss to optimize our framework. We use the cross-entropy loss function to classify and calculate the global and semantic component features of the source and target domains. The formula for the cross-entropy loss function is as follows:

$$L_{id}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i) \tag{3}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

Where N refers to the number of person images in the source domain or target domain, and p(yi/xi) is the probability that the source domain or target domain image belongs to category y when calculating the classification loss.

The semantic component feature loss function includes the loss of head, upper body, and lower body features. The formula is as follows:

$$L_p = L_h + L_u + L_l \tag{4}$$

Where Lh is the head loss, Lu is the upper body loss, and Ll is the lower body loss.

The total loss function L is the source domain loss Ls and the target domain loss Lt, the formula is as follows:

$$L = L_s + L_t \tag{5}$$

4. Experiments

4.1 Implement Details

Market1501 [8] contains 32,668 pictures, 1,501 tagged people from six cameras. Specifically, 12936 face images of 751 identities detected by DPM [17] are used for training. For the test, a total of 19,732 portraits with 750 identities plus some distractors formed a gallery set, and 3,368 hand-cut portraits from 750 identities formed a query set.

DukeMTMC-reid [9] is a subset of the DukeMTMC data set. It contains 1812 identities captured by 8 cameras. There are 16,522 training images, 2228 query images, 17,661 gallery images, of which 1,404 identities appear in more than two cameras. In addition, similar to Market1501, the remaining 408 identities are considered interference factors.

In the experiment, we use cumulative matching characteristic curve (CMC) and average accuracy (mAP) to evaluate the performance of the re-ID algorithm. For Market-1501 and DukeMTMC-ReID, we use the evaluation packages provided by [8] and [9] respectively. In addition, for the sake of simplicity, all the results reported in this paper are under a single query setting, and no post-processing such as reordering [18] is applied.

We use ResNet50 [16] trained on ImageNet [19] as the backbone network. Given each labeled image and its label in the source data set, we keep the size of the input image and adjust it to 256 x 128. In order to enhance the data, we used random cropping, flipping and random erasure [20]. During the training process, we use Adam [21] with a weight decay of 0.0005 to optimize the parameters of 80 epochs. The initial learning rate is set to 3 x10-4. Our model is implemented on the Pytorch [22] platform, and all our experiments on different data sets follow the same settings mentioned above.

4.2 Ablation Study

In order to verify the effectiveness of the semantic component features obtained by the semantic component classifier, we conducted ablation experiments on the Market1501 and DukeMTMC-reID person re-ID data sets. As shown in Table 1, cross-domain on the Market1501 data set Experiments, compared with ECN [11], after adding semantic component features, R-1 increased by 2.5%, and mAP increased by 1.8%. Similarly, as shown in Table 2, on the DukeMTMC-reID data set, our proposed method R-1 increased by 2.4%, mAP increased by 1.9%, and the experimental results proved the effectiveness of our method.

Table 1. Ablation experiment on the effectiveness of semantic features on Market1501

Methods Market-1501	R-1/%	R-5/%	R-10/%	mAP/%
ECN[11]	75.1	87.6	91.6	43.0
Our	77.6	88.4	91.9	44.8

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

Table 2. Ablation experiment on the effectiveness of semantic features on DukeMTMC-reID

Methods DukeMTMC-reID	R-1/%	R-5/%	R-10/%	mAP/%
ECN[11]	63.3	75.8	80.4	40.4
Our	65.7	77.9	81.8	42.3

4.3 Comparision with State-of-arts

We will use the proposed method to compare and verify the performance of cross-domain person re-ID on the Market-1501 and DukeMTMC-reID dataset. Compared with the most advanced cross-domain adaptive methods, our proposed semantic component model is superior to existing methods and has significant advantages.

In the Market-1501 cross-data set person re-ID experiment, we compare with the state-of-arts cross-domain adaptive person re-ID methods, such as ECN [11]. The method in this paper is compared with the state-of-arts cross-domain adaptive person re-ID method ECN [11], its mAP is 1.8% higher, and R-1 is 2.5% higher. In the DukeMTMC-reID→Market-1501 cross-dataset person re-ID experiment, our method is 1.9% and 2.4% higher than the state-of-arts cross-domain person re-ID method ECN [11].

In summary, our proposed cross-domain person re-ID method based on semantic components effectively improves the performance of the cross-domain person re-ID model.

Table 3. Comparison with state-of-arts cross-domain methods on the Market-1501 dataset

Methods Market-1501	R-1/%	R-5/%	R-10/%	mAP/%
UMDL[23]	34.5	52.6	59.6	12.4
PUL[24]	45.5	60.7	66.7	20.5
SPGAN[10]	51.5	70.1	76.8	22.8
MMFA[25]	56.7	75.0	81.8	27.4
CamStyle[26]	58.8	78.2	84.3	27.4
HHL[27]	62.2	78.8	84.0	31.4
ECN[11]	75.1	87.6	91.6	43.0
Our	77.6	88.4	91.9	44.8

Table 4. Comparison with state-of-arts cross-domain methods on the DukeMTMC-reID dataset

Methods Market-1501	R-1/%	R-5/%	R-10/%	mAP/%
UMDL[23]	18.5	31.4	37.6	7.3
PUL[24]	30.0	43.4	48.5	16.4
SPGAN[10]	41.1	56.6	63.0	22.3
MMFA[25]	45.3	59.8	66.3	24.7
CamStyle[26]	48.4	62.5	68.9	25.1
HHL[27]	46.9	61.0	66.7	27.2
ECN[11]	63.3	75.8	80.4	40.4
Our	65.7	77.9	81.8	42.3

5. Conclusion

We propose a cross-domain person re-ID framework based on semantic components. The semantic component classifier is used to extract the local features of person, which avoids the problem of feature mismatch and eliminates the interference of background features, which improves the domain of cross-domain person re-ID. Adaptability, while also solving the problem of occlusion of person images to a certain extent.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

References

[1] . Xuan Z, Hao L, Xing F, et al. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification [J]. 2017.

[2] Zheng Z, Zheng L, Yang Y. Pedestrian Alignment Network for Large-scale Person Re-identification [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[3] Zhao H, Tian M, Sun S, et al. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017.

[4] Huang H, Yang W, Chen X, et al. EANet: Enhancing Alignment for Cross-Domain Person Re-identification [J]. 2018.

[5] Sun Y F, Zheng L, Yang Y, et al. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline)//Proceedings of 2018 IEEE European conference on computer vision. Munich, Germany: IEEE: 2018, p: 480-496.

[6] Li W, Zhu X, Gong S. Harmonious Attention Network for Person Re-Identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[7] Huang H, Yang W, Lin J, et al. Improve Person Re-Identification With Part Awareness Learning [J]. IEEE Transactions on Image Processing, 2020, p:1-1.

[8] Zheng L, Shen L, Tian L, et al. Scalable Person Re-identification: A Benchmark. IEEE, 2016.

[9] Zheng Z, Liang Z, Yi Y. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro[C]//IEEE International Conference on Computer Vision. IEEE Computer Society, 2017.

[10] Deng W, Zheng L, Ye Q , et al. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[11] Zhong Z, Zheng L, Luo Z, et al. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019.

[12] Zhang Z, Lan C, Zeng W, et al. Densely Semantically Aligned Person Re-Identification [J]. 2018. Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In AAAI, 2016.

[13] Sun B, Feng J, Saenko K. Return of Frustratingly Easy Domain Adaptation [J]. AAAI Press, 2015.

[14] Sun B, Saenko K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation [J]. Springer International Publishing, 2016. p: 443-450.

[15] Ganin Y, Lempitsky V. Unsupervised Domain Adaptation by Backpropagation[J]. 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE CVPR, 2016, p: 770-778.

[17] Pedro F F, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. IEEE TPAMI, 2010.

[18] Zhong Z, Zheng L, Cao D, et al. Re-ranking Person Re-identification with k-reciprocal Encoding [J].In IEEE CVPR, 2017, p: 3652-3661.

[19] J Deng. ImageNet : A Large-Scale Hierarchical Image Database[J]. In IEEE CVPR, 2009, p: 248-255.

[20] Zhong Z, Zheng L, Li S, et al. Random erasing data augmentation. 2017.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014.

[22] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. 2017.

[23] Peng P, Tao X, Wang Y, et al. Unsupervised Cross-Dataset Transfer Learning for Person Re-identification[C]//Computer Vision & Pattern Recognition. IEEE, 2016.

[24] Fan H H, Zheng L, Yan C J, et al. Unsupervised person re-identification: Clustering and fine-tuning. ACM TOMM, 2018.

[25] Shan L, Li H, Li C T, et al. Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification[C]// BMVC 2018.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0050

[26] Zhong Z, Zheng L, Li S, et al. CamStyle: A Novel Data Augmentation Method for Person Re-identification. [J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2018.

[27] Zhong Z, Zheng L, Li S, et al. Generalizing a Person Retrieval Model Hetero- and Homogeneously[C]// European Conference on Computer Vision. Springer, Cham, In ECCV, 2018, p: 172-188.