Utilizing Silhouette and Head Information for Improved Cloth-changing Person Re-Identification

Yuzhang Li
School of International Business, Henan University, Zhengzhou 450000, China.
yuzhang_lyz@163.com

Abstract. In recent years, significant progress has been made in person re-identification (ReID). However, as the time span increases, the phenomenon of changing clothes often occurs, and the performance of methods based on person re-identification may decrease in this case. In this article, we focus on solving the problem of cloth-changing person re-identification from a single RGB image, which is more challenging, and improving the accuracy and robustness of traditional methods for recognition. We propose a cloth-changing Feature Regularization learning framework by fusing Silhouette and Head information (SHFR), a two-stream framework which transfers the silhouette information and head information learned by auxiliary stream to the main stream and supplements features unrelated to clothing. Specifically, the main stream approach we adopt is the traditional cloth-changing person re-identification network. In the auxiliary stream, we use the DeepLabV3 semantic segmentation model to extract person silhouette features, and the FCHD fully convolutional head detection model to extract head information with advanced semantic information. We concatenate the head information and background information into the silhouette as input for BackBone to continue training the network. In order to utilize silhouette information and head information, we regularize the cloth-changing features on both main stream and auxiliary stream, allowing main stream model to pay more attention to features unrelated to clothing and improve the accuracy of model recognition. We conduct extensive experiments on our model on the PRCC dataset, and the results show that our method has highly competitive performance.

Keywords: Cloth-changing Person ReID; Silhouette and Head Information.

1. Introduction

The goal of person re-identification (ReID) is to match person with the same identity in different cameras. At present, many advanced methods have made significant progress in ReID through deep learning. Most jobs often unintentionally assume that person with the same identity have the same clothing in different cameras, which performs well on short-term datasets. However, in real life, with the increase of time span, there is often a situation of changing clothes. This can lead to different person wearing similar clothes being incorrectly matched, resulting in a significant decrease in model performance. Figure 1 shows that all the images are from the same person. From the figure, we can see that when there is a slight change in movement (i.e. the time span is very small), person wears the same clothing, and traditional person re-identification methods have higher performance. However, when person’s clothes changes (i.e. with a large time span, such as transitioning from summer to winter), person will add clothes appropriately based on temperature changes. This can lead to incorrect matching issues in traditional person re-identification methods.
Fig. 1 Example of cloth-changing for the same person over a long-time span. All the images are from the same person. We can see that over a long time span, person will change their clothes according to changes in weather, while the head information and silhouette information contain recognition clues unrelated to the clothes.

In order to solve this problem, many existing works have studied cloth-changing ReID problem [1][2][3][4]. Due to the fact that the shape, color, and clothes do not significantly improve the performance of ReID, researchers will focus on exploring features unrelated to clothing to identify person identity. Previously, some researchers have made effective attempts in the task of cloth-changing person re-identification. For example, Han et al. [1] proposed a new CCReID Changing Feature Enhancement (CCFA) model, which extended the feature distribution to implicitly integrate semantically meaningful clothing changes in the feature space, and then ensured the effectiveness of the enhanced features through adversarial learning. Li et al. [5] proposed a new Dual Attention Biometric Clothes Transfusion network (DualBCT Net), which can effectively extract biometric features from the original queried person images. They also proposed a clothing template based cloth-changing person re-identification (CTCC ReID) task, with the aim of extracting clothing features from a given clothing template. Subsequently, a dual attention fusion model was used to fuse biological and clothing features to retrieve target task images, in order to solve the problem of cloth-changing person re-identification. Shu et al. [40] proposed a semantic-guided pixel sampling method. By using a pre-trained human parsing model to obtain body parts and sampling pixels from other pedestrians to change clothes. This model was forced to automatically learn clues unrelated to clothing.

Although researchers have made efforts to reduce the impact of clothing on ReID, they have not fully explored and utilized more discriminative feature information such as silhouette and gait. Therefore, we consider utilizing person's silhouette as the identity-specific feature under the setting of cloth-changing. In addition, the fact that head information can be used to accurately determine a person's identity. In cloth-changing scenes, head information still does not change too much, so we consider introducing head information into the method to improve the final recognition effect.

Specifically, we explore an effective method of utilizing both silhouette information and head information simultaneously. Firstly, we use the DeepLabV3 [21] semantic segmentation model and the effective receptive field of the network FCHD [6] full convolutional head detection model to extract silhouette information and head information, respectively. After extracting these two pieces of information, we concatenate the background and head information into the silhouette to obtain the concatenated image. We use the spliced images as the new input images for BackBone to train the model, which serve as an auxiliary process for adjusting the main stream ReID network. Finally, we encourage cloth-changing feature regularization to make auxiliary stream regulation focus on features unrelated to clothing. Therefore, it can improve the accuracy and robustness of main stream model recognition.

We summarize our main contributions as follows:
We propose a framework named SHFR that solves the problem of cloth-changing person re-identification. It effectively integrates cloth-independent silhouette information and head information, better helping the model learn cloth-independent features.

We explore methods for using silhouette and head information. We use existing methods to extract human silhouette and head information, and merge the head information into the silhouette through masking. In this process, we do not discard the background information used to extract more identity specific features.

The framework PRCC [19] achieves comparable performance on cloth-changing person re-identification dataset, indicating the advantages of framework in handling such challenging scenarios.

2. Related Work

With the continuous advancement of national demands such as smart public security and smart cities, ReID, as a key technology, plays an important role in security, urban management, and smart transportation. ReID can identify and track specific individuals by analyzing the person images captured by surveillance cameras. It can provide important technical support for solving problems such as character recognition and tracking. Many researchers have begun to pay more attention to this topic and provided many effective methods. ReID can be roughly divided into general person re-identification and cloth-changing person re-identification.

2.1 General Person Re-identification

The main task of ReID is to use deep learning and computer vision technology to extract and match features of person images, so as to accurately match images of different person with the same identity in a given image dataset when the same person moves between different cameras. Early ReID work mainly focused on feature extraction [8][9][10][11] and metric learning [7][12].

At present, with the continuous development and improvement of deep learning technology, researchers have applied deep learning to ReID work, using convolutional neural networks to learn and discriminate features [10][13][14], so that ReID can achieve more accurate and robust recognition results in different scenarios. For example, Hu et al. [15] proposed a ReID method based on fine-tuning the ResNet50 network and optimized the model using transfer learning. By preprocessing the dataset during the training phase, they used data augmentation methods to avoid overfitting, and used cosine distance to calculate the spatial distance between image features and the similarity ranking between samples. Wang et al. [16] proposed a novel attention driven framework named AFC for unsupervised ReID and clustering optimization. In order to provide more accurate clustering results, a clustering consensus method was introduced to estimate the similarity of pseudo labels in continuous training. Time propagation and ensemble were used to improve the pseudo labels. Li et al. [17] designed a multi-scale context aware network (MSCAN) to learn powerful features of the whole body and body parts, and proposed three constraints on the learned transformation parameters by using a spatial transformation network (STN) to reduce background information interference. Guo et al.[42] proposed a Camera-style Separation and Uncertainty Estimation Model (CSUE). They addressed the issue of unsupervised domain adaptation person re-identification from two perspectives. Firstly, they used a Camera-aware Style Decoupling module to eliminate constraints on feature extraction, thereby capturing more non discriminatory features independent of the camera. This module alleviated the negative impact of cross camera variations. In addition, they also avoided inherent flaws in clustering by estimating the certainty, gradually refining pseudo labels. Wang et al. [43] proposed a multi instance attention learning framework that uses weakly labeled data to learn person re-identification models from videos. It is worth noting that they only need to set video level labels, which greatly reduced the annotation cost for each frame of the video. Firstly, they placed the person re-identification of video in a multi instance learning setting, allowing videos with similar labels to be used to identify persons. Secondly, they also introduced a co-person attention mechanism to explore the similarity and correlation between videos of person with the same identity.
However, these studies are based on persons with the same identity having the same clothing in different cameras, which performs well on short-term datasets. In real monitoring scenarios, as the time span increases, there is often a situation of changing equipment. This can lead to different persons wearing similar clothes being incorrectly matched, resulting in a significant decrease in model performance. Therefore, an increasing number of researchers pay more attention to the task of cloth-changing person re-identifying and have made beneficial attempts in this task.

2.2 Cloth-changing Person Re-identification

Cloth-changing ReID refers to a special task in the field of re-id, whose main purpose is to solve the recognition problem of persons wearing different clothing in different scenes and times. At present, there is relatively little research on the cloth-changing person re-identification [18][19][20], and its core idea is to learn by extracting features unrelated to clothing such as face and body shape. Zheng et al. [18] proposed a dual path model that extracted both human appearance and shape features. Eliminate clothing information in appearance features by encouraging learning the similarity between appearance and shape features. Yang et al. [19] introduced a learning based spatial polar coordinate transformation (SPT) layer to transform contour sketch images and extracted more robust discriminative features from contour sketch images. In the subsequent layers, they also applied specific angle extractors (ASE) to extract finer grained specific angle discriminative features. Qian et al. [20] introduced a shape embedding module and a clothes elimination shape extraction module to learn identity sensitive and clothes insensitive representations. They utilized the interrelationships between key points in the human body to extract biological structural features, and used attention mechanisms to separate identity related features from relevant clothing information. This method aimed to eliminate unreliable clothing appearance features and focused on body shape information. Wu et al. [41] proposed an effective Identity Sensitive Knowledge Propagation framework (DeSKPro), which consisted of a global cloth-irrelevant feature stream(global stream) and a facial feature enhancement stream(face stream). A Cloth-irrelevant Spatial Attention module(CSA) that was trained globally under the guidance of the knowledgeable human parsing network to eliminate interference from clothes. Face stream restored lost facial details through previous facial knowledge by training a teacher network. Zhao et al. [44] proposed a novel joint Identity aware Mixstyle and Graph enhanced Prototype method. On the one hand, they proposed the identity aware mix style (IMS) as a fine-grained style transition from a domain generalization perspective by changing clothes. This not only mixed the instance level feature statistics of identity samples, but also preserved the correspondence between the synthesized samples and the potential label space. On the other hand, they proposed the graph enhanced prototype constraint (GEP) module to further reduce the differences between features caused by clothing changes. This module explored the potential graph similarity structure of samples across memory to construct more informative prototypes, thereby learning better examples of clothing independent metric learning. These two modules were integrated into the joint learning framework to learn from each other and improve the effectiveness of the model.

However, existing methods in cloth-changing person re-identification (ReID) often face challenges such as estimation errors and limited exploration of human semantic information and person identity information. In this work, we are committed to overcoming these challenges by introducing silhouette extractor and head detector into the ReID model. By fully utilizing these human semantic information and person identity features, our model can better identify the person in images and extract more discriminative features from them. Our method helps to solve the common problem of clothing changes in cloth-changing person re-identification and improves the performance of the person re-identification model in changing environments.
3. Methodology

3.1 Framework

SHFR framework is shown in Figure 2. Specifically, SHFR framework consists of main stream and auxiliary stream, with the auxiliary stream serving as a regulator to optimize the main stream model. In the main stream, we use ResNet50 [24] to extract features from each person. In the auxiliary stream, we use the DeepLabV3 model to extract the silhouette information of person in the input graphics, and use the FCDH model to extract the head information of person. Then, by masking, the head information is concatenated and fused into the silhouette to obtain a highly robust and recognizable concatenated image. We do not discard the background information here. Specifically, when changing different backgrounds, the model tends to recognize that the background cannot serve as the key information for distinguishing person identification, and focuses more on feature information related on identify. Finally, we put the concatenated image into BackBone for training, and adjust the main stream and auxiliary stream through cloth-changing feature regularization to optimize the network.

Fig. 2 Framework of SHFR. It contains two streams which are main stream and auxiliary stream. The auxiliary stream plays the role of a regulator in the model, extracting silhouette and head information from the input RGB image. Then, the background and head information are concatenated into the silhouette and fed into BackBone for training. Then, the main stream and auxiliary stream are jointly trained through cloth-changing feature regularization, driving the main stream to learn cloth-independent feature from a single RGB image.

3.2 Auxiliary Stream Design

For the silhouette extractor, we use the DeepLabV3 learning model for semantic image segmentation on a single RGB image to extract silhouette features. The backbone network of DeepLabV3 is usually a pre-trained deep convolutional neural network for feature generation, such as ResNet and MobileNet.

For the head detector, we use the FCHD fully convolutional head detector for head information extraction. FCHD can perform multi-scale detection of targets at different scales and positions. At each sliding window position, the fully convolutional detection head simultaneously predicts the target category and regresses the bounding box, allowing FCHD to learn rich head features.
After obtaining person silhouette and head information through silhouette extractor and head detector, we concatenate the background and head information of the image into the silhouette to obtain an image with high semantic information. We do not discard the background information here. Specifically, when changing different backgrounds, the model tends to recognize that the background cannot serve as the key information for distinguishing person identification, and focuses more on feature information related on identify. We train the spliced images as input images to BackBone. In this paper, we use ResNet50 as the backbone network to extract image features.

### 3.3 Cloth-changing Feature Regularization

In order to comprehensively learn cloth-independent features from RGB images, we integrate the appearance knowledge of main stream and auxiliary stream, and transfer the silhouette and head knowledge learned by auxiliary stream to the main stream for supplementation. We utilize soft labels for knowledge transfer by using Kullback Leibler (KL) Divergence in feature regularization ($D_{KL}$) [22]. The expression for $D_{KL}$ is:

$$D_{KL}(p^m || p^a) = \sum_{i=0}^{n} p^m_i \log \frac{p^m_i}{p^a_i}$$

Where $p^m$ represents the probability of the main stream output class, and $p^a$ represents the output probability of the auxiliary stream. From (1), it can be concluded that the KL divergence loss of the main stream and auxiliary stream is calculated as $D_{KL}(p^m || p^a) , D_{KL}(p^a || p^m)$, and the KL divergence expression can be obtained as:

$$L_{KL}(\partial_m, \partial_a) = D_{KL}(p^m || p^a) + D_{KL}(p^a || p^m)$$

Among them, $\partial_m$ and $\partial_a$ are the network parameters of the main stream and auxiliary stream, respectively. In the semantic space, KL divergence helps us to bring the main stream and auxiliary streams closer together, so that the main stream can learn more features that are not related to clothing.

### 3.4 Training and Inference

To train the network, we use the widely used identification loss (ID loss), which is the cross entropy loss used for classification. We also use triplet loss[23] to learn more robust and discriminative features.

The overall loss function for training our model consists of identification loss, triplet loss and KL divergence loss. Firstly, We use identification loss (cross-entropy loss) [33][34] to measure the difference between predicted and true values, where the main stream identity loss is $L_{id1}$ and the auxiliary stream identity loss is $L_{id2}$. Given sample I, the identity label is Y. We utilize predicted identity $C_{r}^m$ and $C_{d}^a$ defines identification loss $L_{id1}$ and $L_{id2}$ with real labels Y, and trains the ReID models for main stream and auxiliary stream by minimizing $L_{id1}$ and $L_{id2}$.

We use triplet loss $L_T$ to learn a feature space where anchor samples of the same category are closer to positive anchors, and anchors of different categories are further away from negative anchors. The formula is:

$$L_T(a, p, n) = \max(\beta + d_{ap} - d_{an}, 0)$$

Among them, $a$ represents anchor sample, $p$ represents positive sample with the same ID as anchor sample. $n$ represents negative sample with a different ID than anchor sample, and $d_{ap}$ and $d_{an}$ represent the Euclidean distances between anchor sample and positive sample, negative sample, respectively. $\beta$ refers to the margin, a constant greater than 0. The ultimate goal of optimization is to pull in the distance between the anchor and positive, and to pull out the distance between the anchor and negative.

Finally, the two streams are trained end-to-end through optimization, and the total loss function expression is:

$$L = \lambda_T L_T + \lambda_{id1} L_{id1} + \lambda_{id2} L_{id2} + \lambda_{KL} L_{KL}$$
In the formula, $\lambda_T$, $\lambda_{id1}$ represent the triplet loss weight, identification loss weight in the main stream. $\lambda_{id2}$ represents identity loss weight in the auxiliary stream, and $\lambda_{KL}$ represents KL divergence weights.

It's worth noting that in the inference, we only use main stream feature to calculate similarity and retrieve the most similar to calculate hit rate.

4. Experiments

4.1 Dataset and Evaluation Protocols

For dataset, we evaluate the PRCC dataset[19]. PRCC, designed for person re-identification research under moderate cloth-changing, is known as the Personal Identity Dataset under Moderate Clothing Change. It includes 221 individuals captured from three camera perspectives, each with 2 sets of clothing. PRCC has a total of 33698 images.

For evaluation protocols, we use mean average precision (mAP) and rank-1 accuracy. We follow its evaluation plan:

(i) Standard setting, which means consistent clothing setting. In this case, we only use the ground truth consistent with the clothing to calculate accuracy. We considered the matching situation of person wearing the same clothes between different cameras. Through this evaluation setting, we can evaluate the performance of the model in traditional person identify scenarios;

(ii) Change of clothes setting, in which the accuracy is calculated using the ground truth of the change of clothes. We consider the situation where the same person wears different clothes at different times or in different occasions. This setting is more closely related to real-life situations, as person often change clothing according to seasonal or situational changes.

4.2 Implementation Details

We adopt ResNet50 as ReID Backbone in main stream, and also use the same backbone network in the auxiliary stream. In order to increase the richness of features, we remove the last down sampling operation of ResNet50. For the images in the PRCC dataset, following LT-reID [25], we adopt global average pooling and global maximum pooling methods to integrate the output features. Then we connect these features together and normalize the image features through Batch Normalization [26]. Finally, according to past research [20], the size of the input image is adjusted to 384 * 192. This design aims to fully utilize the feature representation ability of the network, and improve the performance and robustness of person re-identification model.

Each batch contains 8 people and 8 images for each person. Adam [27] trained the model at 60 epochs. We set the initialization learning rate to $3.5e^{-4}$, dividing every 20 epochs by 10. In the triplet loss formula, we set $\beta$ to 0.5. For the PRCC dataset, we set $\lambda_{KL}$ to 5. We set the weights of triplet loss $\lambda_T$ and two identity losses $\lambda_{id1}$ and $\lambda_{id2}$ to 1.

4.3 Comparison with existing methods

We compare SHFR with the currently advanced 15 methods, as shown in Table 1. It is worth noting that the bold data in Table 1 is the best performing, while the underlined data is the second best performing. We can see that our method has more competitive performance on the PRCC dataset. Specifically, the four methods of GI ReID, FSAM, 3DSL, and ISP all utilize multimodal information to avoid interference from clothing on the model, but this increases additional computational costs compared to our method. Compared to the FSAM method which only uses sketch information to assist mainstream learning of features unrelated to clothing, SHFR focuses on more human semantic information (silhouette and head information) to improve mainstream performance.

For the other 10 methods that are only in RGB modality, they only consider the information in the RGB image and ignore the importance of appearance in improving the performance of person re-identification. SHFR uses a silhouette extractor to extract person silhouette information,
supplementing person appearance information into the person re-identification model, thereby improving the recognition performance of the model. RGB images are also affected by lighting, shooting angle, and noise. Due to different lighting conditions and shooting angles, the color and texture in RGB images may change, and noise may interfere with the model's feature extraction and matching, which reduces the accuracy of feature extraction. However, SHFR utilizes silhouette information to avoid the impact of lighting, shooting angle, and noise on model performance.

In addition, except for the causal relationship based clothing ReID model AIM, which has the best performance compared to other methods, SHFR's performance is in the top 2 compared to other methods. This comparison can demonstrate the effectiveness of SHFR's performance.

Table 1: Performance evaluation and comparison of 15 person re-identification methods on the PRCC dataset. Specifically, the bolded data in the table has the best performance, while the underlined data ranks second in performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Modality</th>
<th>PRCC standard</th>
<th>PRCC cloth-changing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>rank1 mAP</td>
<td>rank1 mAP</td>
</tr>
<tr>
<td>IANet[28]</td>
<td>RGB</td>
<td>99.4 98.3</td>
<td>46.3 45.9</td>
</tr>
<tr>
<td>STN[29]</td>
<td>RGB</td>
<td>59.2 -</td>
<td>27.5 -</td>
</tr>
<tr>
<td>HPM[30]</td>
<td>RGB</td>
<td>99.4 96.9</td>
<td>40.4 37.2</td>
</tr>
<tr>
<td>MGN[31]</td>
<td>RGB</td>
<td>99.5 98.4</td>
<td>33.8 35.9</td>
</tr>
<tr>
<td>HACNN[32]</td>
<td>RGB</td>
<td>82.5 21.8</td>
<td>-</td>
</tr>
<tr>
<td>RGA-SC[33]</td>
<td>RGB</td>
<td>98.4 -</td>
<td>42.3 -</td>
</tr>
<tr>
<td>PCB[34]</td>
<td>RGB</td>
<td>99.8 97.0</td>
<td>41.8 38.7</td>
</tr>
<tr>
<td>CAL [2]</td>
<td>RGB</td>
<td>100 99.8</td>
<td>55.2 55.8</td>
</tr>
<tr>
<td>Gl-ReID[3]</td>
<td>Multi-modal</td>
<td>-</td>
<td>37.6 -</td>
</tr>
<tr>
<td>FSAM[35]</td>
<td>Multi-modal</td>
<td>-</td>
<td>54.5 -</td>
</tr>
<tr>
<td>RCSANet[36]</td>
<td>RGB</td>
<td>100 97.2</td>
<td>48.6 50.22</td>
</tr>
<tr>
<td>SPT+ASE[19]</td>
<td>sketch</td>
<td>64.2 -</td>
<td>34.4 -</td>
</tr>
<tr>
<td>3DSL[37]</td>
<td>Multi-modal</td>
<td>-</td>
<td>51.3 -</td>
</tr>
<tr>
<td>ISP[38]</td>
<td>Multi-modal</td>
<td>92.8 -</td>
<td>36.6 -</td>
</tr>
<tr>
<td>AIM[39]</td>
<td>RGB</td>
<td>100 99.9</td>
<td>57.9 58.3</td>
</tr>
<tr>
<td>SHFR</td>
<td>RGB + silhouette</td>
<td>99.9 99.8</td>
<td>58.5 57.6</td>
</tr>
</tbody>
</table>

4.4 Ablation Studies

In this section, we investigate the effectiveness of the key components about the proposed method. As shown in Table 2, our proposed method outperforms the baseline model which is ReID BackBone significantly on the PRCC dataset. When we only supplementing the silhouette information to the baseline, these two methods are jointly optimized, which is represented as"+silhouette"(256,668),(908,696) in Table 2. When the head information is supplemented with "+silhouette", the fused result is represented as "SHFR " in Table 2. When we analyze the method of supplementing silhouette information, the experimental results show that in the cloth-changing scene, the method of supplementing silhouette information has improved performance but is very limited. After further fusing the head information, it can be seen from the data in Table 2 that SHFR's performance indicators are significantly improved in both the standard setting and the changing setting. The utilization of silhouette information and head information in a complementary manner proves to enhance the accuracy and robustness of the model in identifying persons.

Table 2: Ablation study of our method

<table>
<thead>
<tr>
<th>Method</th>
<th>PRCC standard</th>
<th>PRCC cloth-changing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rank1 mAP</td>
<td>rank1 mAP</td>
</tr>
<tr>
<td>baseline</td>
<td>99.8 97.9</td>
<td>45.6 46.6</td>
</tr>
<tr>
<td>+silhouette</td>
<td>99.7 98.1</td>
<td>51.1 50.6</td>
</tr>
</tbody>
</table>
4.5 Limitation

The SHFR model only extracts silhouette information and cannot capture specific human pose and action details. When there are significant changes in human post such as standing sideways or curling up, the performance of the model will be significantly reduced. In addition, the thickness of person clothing has a significant impact on silhouette. When the time span is long, person may add or remove clothing based on temperature, which may result in inconsistent silhouettes of the same person. In these special cases, the accuracy and robustness of the SHFR model for person identification may decrease.

5. Conclusion

In this article, we propose a cloth-changing Feature Regularization learning framework by fusing Silhouette and Head information (SHFR), which utilizes image silhouette information and head information with feature regularization to solve the cloth-changing ReID problem in a single RGB image. We optimize the main stream by using auxiliary stream as a regulator to improve the recognition ability and robustness of the model. The main stream is the common ReID model. In the auxiliary stream, we use DeepLabV3 semantic segmentation model and FCHD fully convolutional head detection model to extract person’s silhouette information and head information. Finally, we use cloth-changing feature regularization to optimize main stream model. We conduct extensive experiments on the PRCC dataset to demonstrate the effectiveness of our method.

References


