Person Re-Identification: Literature Review

Author Details: Van Nam Pham
University of Economics - Technology for Industries, Vietnam
Correspondence: Van Nam Pham, 456 Minh Khai, Hai Ba Trung, Ha Noi

Abstract:
Person ReID is known as associating cross-view images of the same person when he/she moves in a non-overlapping camera network. In recent years, along with the development of surveillance camera systems, person re-identification (ReID) has increasingly attracted the attention of computer vision and pattern recognition communities because of its promising applications in many areas, such as public safety and security, human-robotic interaction, and person retrieval.

Keywords: *Person ReID, camera*

1. Introduction

Person ReID is known as associating cross-view images of the same person when he/she moves in a non-overlapping camera network [1]. In recent years, along with the development of surveillance camera systems, person re-identification (ReID) has increasingly attracted the attention of computer vision and pattern recognition com- munities because of its promising applications in many areas, such as public safety and security, human-robotic interaction, and person retrieval. In early years, person ReID was considered as the sub-task of Multi-Camera Tracking (MCT) [2]. The purpose of MCT is to generate tracklets in every single field of view (FoV) and then associate the tracklets that belong to the same pedestrian in different FoVs. In 2006, Gheissari et al [3] firstly considered person ReID as an independent task. On a certain aspect, person ReID and Multi-Target Multi-Camera Tracking (MTMCT) are close to each other. However, the two issues are fundamentally different from each other in terms of objective and evaluation metrics. While the objective of MTMCT is to determine the position of each pedestrian over time from video streams taken by different cameras. Person ReID tries to answer the question: "Which gallery images belong to a certain probe person?" and it returns a sorted list of the gallery persons in descending order of the similarities to the given query person. If MTMCT classifies a pair of images as co-identical or not, person ReID ranks the gallery persons corresponding to the given query person. Therefore, their performance is evaluated by different metrics: classifi- cation error rates for MTMCT and ranking performance for ReID. It is worth noting that in case of overlapping camera network, the corresponding images of the same per- son would be found out based on data association, and can be considered as person tracking problem, which is out of scope of this thesis. In the last decade, with the un- remitting efforts, person ReID has achieved numerous important milestones with many great results [4, 5, 6, 7, 8], however, it is still a challenging task and confronts various difficulties. These difficulties and challenges will be presented in the later section. First of all, the mathematical formulation of person REID is given as follows.

2. LITERATURE REVIEW

2.1. Person ReID classifications

Single-shot versus Multi-shot

According to the aforementioned definition of person ReID, in single-shot scenarios, the number of images for query person and per- son in gallery sets is one while in multi-shot scenarios, the value of ni and mj are greater than one. Early person ReID studies only focused on single-shot approach in which person matching is mainly relied on comparison between two images (one inprobe and another in gallery) [10, 11, 12, 13]. By contrast, in multi-shot person ReID, each pedestrian is described by multiple images or sequences. In 2010, the first studies on multi-shot person ReID were reported [14, 15]. On the one hand, although single- shot scenario is seem to be far from realistic applications, obtained results for this case would be a crucial brick for multi-shot.

Computational cost as well as memory stor- age requirements for single-shot problem are much lower than those for the multi-shot person ReID. On the other hand, multi-shot person ReID can provide richer and more useful information to improve the ReID accuracy. However, multi-shot person ReID has its own issues such as memory storage requirement, computation time. To solve this problem, several studies have introduced some solutions to extract key frames which contain sufficient information to describe a pedestrian. This approach not only helps to remove redundant information but also increases calculation speed and saves memory capacity. These key frames are extracted based on some cluster algorithms or distribution of motion energy.

2.2. Closed-set versus Open-set person ReID

In a broad view, person ReID can be defined as closed-set and open-set problems. Majority of the existing person ReID works belong to the former approach with the assumption that each individual appears in both probe and gallery sets. Also, this task can be understood as the matching problem which aims to seek the occurrences of a query person (probe) from a set of person candidates (gallery), called closed-set person ReID. However, the hypothesis that images of the same person are captured by different cameras is not always satisfied and limits the capability of practical ap- plications. Therefore, the open-set person ReID has become an inevitable trend and received more and more attention of the researchers all over the world. The open-set person ReID aims to answer the question: "Does a query person appear in the gallery set?". In this case, the query probe might appear in the gallery set or not and open-set person ReID is turned into person verification [16, 17, 18, 19, 20].

2.3. Supervised and unsupervised person ReID

Based on the availability of matched pairs used in the training phase, person ReID can be divided into supervised and unsupervised approaches. In general, training is the crucial phase in pattern recognition problems which helps improve performance of the recognition process. Accordingly, pedestrian's models on cross-view cameras are learned from the previously collected dataset and the matching pairs are labelled for the training phase. This requirement not only creates a burden for human labors but also limits the scalability of person ReID problem. Moreover, the data labeling process becomes even more difficult and infeasible when dealing with a large-scale dataset. In order to overcome this difficulty, some latest studies have followed the unsupervised approach which employs unlabelled data into consideration [21, 22, 23, 24, 25]. From this approach, person ReID is closer to a realistic context in which thousands of people appear on a camera view everyday in public spaces, such as airports, railway-stations, super markets. Since, labeling task is a self-taught process, without supervision. On the one hand, the unsupervised methods help to reduce the human labor and toward realistic systems. On the other hand, the matching rates of these methods are often much lower than those of the supervised ones. This is because of that without manually labelled matching pairs in cross-view cameras, existing unsupervised models cannot learn the appearance transformation of the same person from different camera views.

As the research focuses on the supervised person ReID approach, two person settings in which a certain pedestrian's model is previously trained or not are further discussed. In the former setting, person ReID is turned into person search [26, 27, 28], the identity of an interested person is determined based on the similarity between this person and each of trained models. This indicates that the identities of pedestrian in the testing set are the same ones in the training set. This setting is suitable and employed in several real situations, such as human management and monitoring, suspect/criminal search, etc. Obviously, when a pedestrian's model is previously learned, person ReID performance is greatly improved. By contrast, in the later one, the model of each individual in the testing phase is not learned in advance. Specially, the identities of the pedestrians for the test phase are different from ones for the training phase. This assumption is more likely to be closer to a realistic context and majority of existing person ReID studies support the second evaluated setting. Figure 1.4 shows the differences in two above evaluation settings for person ReID. Fig. 1.4a) used the same color to describe that the given query exists in the gallery set while Fig. 1.4b) illustrates that query person's appearance

models are not previously known, described by the colors of the query persons are different from those of the gallery persons. In the first setting, query persons exist in the gallery set. This means that the model of query persons is previously learned while in the second setting the query person's appearance models are not previously known.

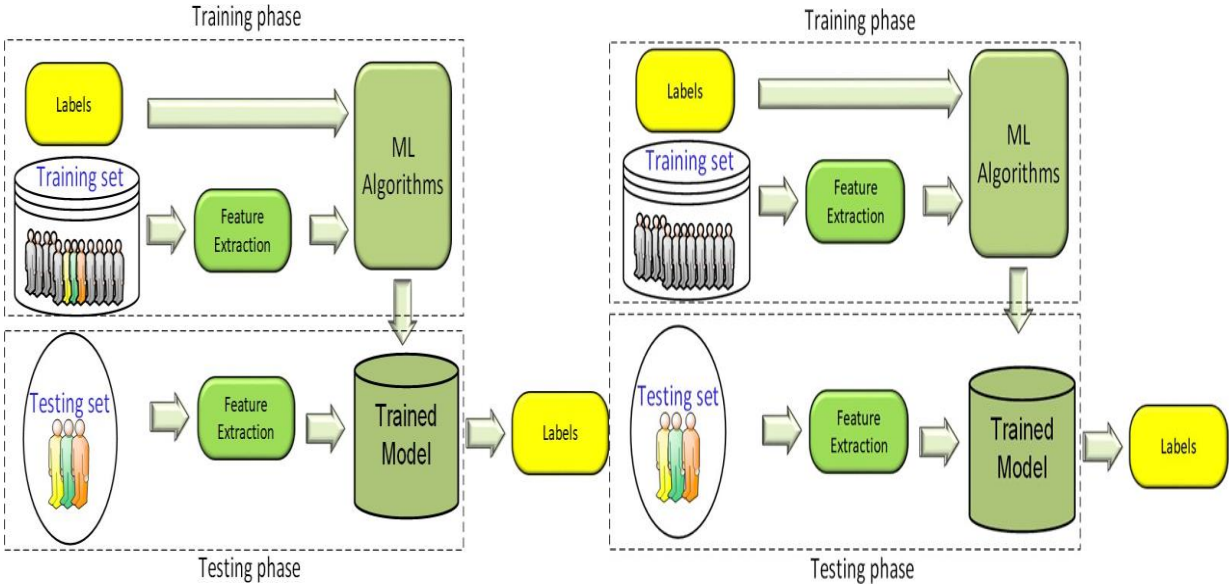
For more details, we will be analysis some prominent researches on person ReID to have a better understanding on the existing approaches on this problem. Similar to any pattern recognition problem, feature extraction and metric learning are the two indispensible components of person ReID problem. Most existing person ReID works have tried to exploit the effectiveness of one/both of them and toward the target of improving ReID performance.

3. Datasets and evaluation metrics

3.1. Datasets

In the literature, there are numerous datasets used for person ReID evaluation. Some

datasets are set up for either single- or multiple-shot approaches. However, the other datasets are set up for both scenarios. In single-shot approach, each person has sole one image in both probe and gallery sets. Meanwhile, each person is presented by multiple image in both probe and gallery sets in multi-shot approach. Additionally, two settings for person ReID will be mentioned in this thesis. For the first setting, a person in the testing set has appeared in the training set. Inversely, for the second setting, the testing and training sets are completely different. These concepts wil be described in more details in the Chapter 1. Five benchmark datasets used for performance evaluation of the proposed methods in this thesis will be indicated as follows.



(a) (b)

Figure 1. Two popular settings for person ReID problem:

a) The testing persons have appeared in the training set (represented by the same colors) b) Persons in the training and testing sets are absolutely different.

• Viewpoint Invariant Pedestrian Recognition (VIPeR) [29]. This is one of the most challenging datasets with strong variations in pose, illumination, view- point and occlusion. The dataset contains 1,264 images of 632 persons, with image resolution is 128 × 48. Each person has a pair of images captured by two different non-overlapping cameras. This dataset is used for the single-shot case.

- CAVIAR4REID [30]. This dataset contains 1,220 images of 72 pedestrians captured from two non-overlapping cameras in a shopping mall. However, there are only 50 persons appeared in both cameras. This dataset is generated to maximize the variation in illumination conditions, occlusions, image resolutions, and view-points. In this dataset, the image resolutions vary strongly, from 17 × 32 to 72 × 144. This also cause difficulty for person ReID.

- RAiD [31]. Comprising 6,920 images of 43 individuals appeared in four cameras (two indoor, two outdoor). Only 41 of the 43 total persons appear in all cameras. All images in this dataset are normalized to the same size of 64 × 128. The large illumination variations caused by collecting images from different scenarios is one of the challenges when working on this dataset. In this thesis, the gallery set is generated by selecting randomly 5 images for each person and the remaining images are used for the probe set.

- PRID-2011 (Person ReID) [32].

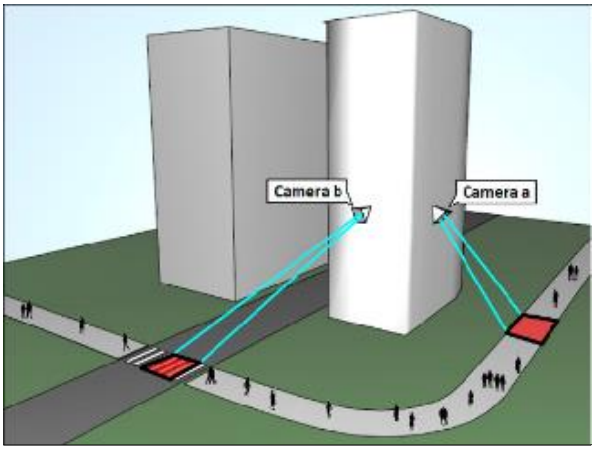


Figure 2: Camera layout for PRID-2011 dataset [33].

This dataset was created in the Austrian Institute of Technology (AIT) for ex- periments on person ReID. Images in this dataset are extracted from multiple pedestrian trajectories captured from two static non-overlapping cameras. These images suffer from large variations in illuminations, view-point, poses, etc. Figure

Figure 2 shows camera layout for PRID-2011 dataset, two cameras are installed in two sides of the building in AIT with different view-points. The original videos have 475 pedestrians from one view and 856 from the other view, in which 245 per- sons appear in both views. Some images are filtered out due to strong occlusions, sudden disappearance/appearance or number of reliable images for each person in each camera view less than five. After filtering, there are 385 persons in camera view A and 749 persons in camera view B. The first 200 persons appear on both views and are used in person ReID experiments. It is worth noting that

this thesis follows the experimental setting in [34], only 178 persons having more than 21 images in an image sequence were chosen for evaluation. The data is divided into two halves, one for the training and test phase. This random division is repeated 10 times and the reported result is the average value of these 10 splits.

- iLIDS-VID (Imagery Library for Intelligent Detection Systems) [35]. This dataset was recorded at an airport arrival hall under a multi-camera CCTV network. It consists of 300 pedestrians with 600 image sequences. The length of each sequence varies from 23 to 192 images, with an average number of 73. In person ReID evaluation, this dataset is also randomly split into two halves, one for training and the remaining for testing. This process is performed 10 times to achieve a fair comparison.

3.2. Evaluation metrics

In order to evaluate the proposed methods for person ReID, we used Cumulative Matching Characteristic (CMC) curves [37]. CMC shows a ranked list of retrieval person based on the similarity between a gallery and a query person. The value of the CMC curve at each rank is the rate of the true matching results and total number of queried persons. The matching rates at several important ranks (1, 5, 10, 20) are usually used for evaluating the effectiveness of a certain method.

3.3. Feature extraction

The first crucial component for any pattern recognition problem is feature extraction step. In order to describe a pedestrian image, biometric cues (eyes, iris, gait) or visual appearance is exploited. These are considered as the most useful information for person representation. However, because images/videos in person ReID are usually captured with low resolution, information extracted on eyes or iris is not sufficient for person representation. Besides, gait is a whole-body, behavioral bio-metric that is considered as a pedestrian's characteristic and has been studied for person ReID for a long time. However, it is not easy for extracting human gait because of the complexity of a realistic surveillance environment, such as airport, railway station, super market. Additionally, human gait usually strong depends on person mood and health condition. Consequently, majority of existing person ReID studies mainly focus on visual appearance of pedestrian [8]. In general, features are classified into two main categorizes: hand-designed and deep-learned features. In the early days, hand- designed features were proposed for image representation. These features are built on experiences and perceptions of researchers [28, 9]. Fortunately, in 2014, with the rapid development of Convolutional Neural Network (CNN), the first studies on deep-learned features was applied to person ReID. Since then, a lot of works have paid attention to exploiting the capability of numerous deep networks. Features are also categorized into three different abstract levels including low-, mid-, and high-level features [38]. While low-level features contain color, texture and shape information extracted from every pixel, more advanced low-level features are created by computing a covarian matrix from image derivatives [39] or seeking at local key points (SIFT [40]). Mid-, and high-level features are constructed by learning from pixel-level features, for example, Bag of Words (BOW) models [41] which encode low-level features into visual words are considered as mid-level features and deep-learned features extracted from the last layers of a CNN are high-level ones.

Hand-designed features

As widely recognized, color and texture are the most widely used features for person representation in person ReID. In comparison, color seems to be more important cue due to the low-resolution images captured in surveillance videos [9]. In 2006, Gheissari et al.[3] was the pioneer treating person ReID as an independent vision task. In this work, a novel segmentation method was proposed, in which spatial and temporal cues are exploited to generate invariant signatures for the change of pedestrians' appear- ance. Color and edge histograms are computed and combined for generating the unique signature for region representation. While color histogram is extracted on HSV color space, edge histogram encodes the dominant local boundary orientation and the RGB ratio on either sides of the edge. After that, these region-based feature vectors are incorporated for image representation. One of the most outstanding and comparable studies is proposed by Farenzena et al. [14]. In this work, by observing that human body is often located in the center region of an image and background might cause un- wanted noise for ReID process. The authors attempt to segment pedestrian foreground from background, human body is divided into three main parts composing head, torso, and legs based on as-symmetrical axes. Then, these parts are further divided by a symmetrical axis. Based on this body structure, three different kinds of features that are weighted color histogram (WH), the maximal stable color region (MSCR), and the recurrent high-structured patches (RHSP) are extracted on each part. For WH feature, it assigns larger weights to pixels near the symmetrical axis and constructs color histogram for each part. While MSCR searches stable color regions and extracts features containing information about color, area, and centroid of these regions. Be- sides, RHSP is a kind of texture features and captures recurrent texture patches. The combination of these features assists in building a robust and discriminative descriptor for person ReID problem. In [42], the authors try to construct a descriptor from information of a pixel such as the pixel co-

ordinates, its intensity as well as the first and the second-derivaties of this pixel. By this way, each pixel is presented by a 7-dimension vector and after that, these extracted feature vectors are turned into Fisher Vector, called Local Descriptors encoded by Fisher Vector (LDFV), for person representation. Zhao et al. [43] solved person ReID by finding saliency regions in an pedestrian image. In this work, saliency regions are defined as the outstanding and easily recognizable cues for distinguishing different persons. Dense patches 10 × 10 pixels are extracted with a step size of 5 pixels, and then, 32-dim LAB color histogram and 128-dim SIFT descriptor are computed on each patch. Adjacency constrained search is exploited to seek the best match for a query patch in horizontal stripes with the similar latitudes in gallery images. By this way, the higher scores are assigned to the more prominent patches. These scores are the fundamental for calculating the similarity between two persons.

With the consideration for the influence of illumination variations, in [44] the au- thors calculated color histograms in different color spaces and then, combined these histograms to take the more robust signature to variations of illumination. In this work, the authors claimed that performance of person representation is still not satisfactory if only relying on exploiting color histograms. Based on this analysis, they proposed a novel Salient Color Names based Color Descriptor (SCNCD). For SCNCD descriptor, 16 colors in RGB space including fuchsia, blue, aqua, lime, yellow, red, purple, navy, teal, green, olive, maroon, black, gray, silver, and white are used. Salient color names show that each color has a certain probability of being assigned to nearest color names, and the closer one owns a higher probability. In order to overcome the challenge caused by different viewpoints, Liao et al. [45] focused on horizontal occurrence of local fea- tures and maximize them to build a robust descriptor, named LOMO. This descriptor includes the color and Scale Invariant Local Ternary Pattern (SILTP) histograms [46]. Bins in the same horizontal stripe undergo max pooling and a three-scale pyramid model is built before a log transformation. LOMO descriptor is later is employed in several works [47, 48].

In [39], co-variance descriptor is employed for image representation in person ReID. In which, the target of this descriptor is to encode information on feature variances within a given image region, their correlation with each other as well as their spatial location. The effectiveness of this descriptor is based on the invariant property of co- variance matrices which help to overcome the changes in illumination and rotation. Additionally, one more advantage of the covariance descriptor is that it can be com- puted on any kind of images, such as intensity image, color image. This descriptor is also used in the study of Matsukawa et al. [49]. In this study, the authors claimed that mean and covariance are the most important cues for representing a person appearance. They proposed Gaussian of Gaussian (GOG) descriptor which inherits the advantage of covariance-based descriptor. This descriptor is constructed on three different levels including pixel-, patch-, and region-level. Gaussian distribution is applied twice at patch and region levels, therefore, named as Gaussian of Gaussian.

Different from directly exploiting low-level color and texture features in the above studies, some other works based on attribute-based features can be considered as mid- level representation. In comparison with low-level descriptors, attribute features are more invariant and robust to the change of person images captured from cross-view cameras. In [50], the authors introduced 15 binary attributes for describing a person image relating to attire and soft biometric, such as short, skirt, sandal, backpack, longhair, short-hair, gender and etc. Additionally, these attribute classifiers are trained by exploiting the low-level color and texture features. This information is combined with visual features extracted by SDALF method [14] to get a higher ReID performance.

In [51], the authors embed the binary semantic attributes of the same person in cross-view cameras into a continuous low-rank attribute space, so that the attribute vector is more discriminative for matching. Shi et al. [52] introduced a framework to learn a semantic attribute model from existing fashion photography datasets. These attributes help person ReID to achieve competitive results. Different from previous works, the authors take a generative modelling approach relied on the Indian Buffet Process (IBP). This model has several

advantages including: simultaneously learning of all attributes; the ability to naturally exploit training data in both supervised and unsupervised manners. With a great effort, Li et al. [53] built a large-scale dataset for pedestrian attribute recognition, namely Richly Annotated Pedestrian (RAP). This dataset is generated from real multi-camera surveillance scenarios with long term collection consisting 41,585 pedestrian samples with 72 attributes. In particular, environ- mental and contextual factors are also considered in this dataset.

From the above analysis, we can give some remarkable points about hand-designed features as follows. Most of the existing studies have tried to build a descriptor in- cluding useful information about color, texture or shape for person representation. To overcome difficulties caused by the variations in illuminations, view-points, poses,...the above information might be extracted on different scales/levels to form a descriptor is not only robust but also discriminative. Besides, attribute-based features are also considered and exploited as complementary information to appearance. In fact, hand- designed features are usually built based on knowledge and experiments of researches. Therefore, a descriptor might be only effective on several datasets but not be effective on the others. This is one of the disadvantage of hand-designed features. Addition- ally, to extract more and more information at different scales/levels, hand-designed features always have a large dimension, for example approximately 26,960 dimensions for LOMO or 27,622 dimensions for GOG descriptor. Large-dimensional vectors also bring difficulties to storage and computation in the person matching step.

Deep-learned features

Recent years have witnessed the impressive results of Convolutional Neural Networks (CNNs) on pattern recognition tasks. For person ReID, CNN and its variants are employed to achieve a higher performance. In 2014, the first two studies on person ReID exploiting the capability of the deep-learned network are introduced [54, 55]. In [54], the authors proposed a framework using Siamese deep neural network, in which appearance-based features (color, texture) and similarity metric are simultaneously learned. Siamese network has a symmetrical structure including two sub-networks that are connected by a cosine layer. Each sub-network composes two convolutional layers and a fully connected layer. By this way, similarity metric is learned directly through features extracted on image pixels. An input image is divided into three overlapping horizontal parts, and then, these parts are forwarded to the sub-networks and the output of the Siamese network is the similarity of image pairs based on cosine distance.

Different from the above work, Li et al. [55] designed a new deep network with a patch matching layer, namely Filter Pairing Neural Network (FPNN). With this ar- chitecture, the authors expected that several main challenges in person ReID, such as photo-metric and geometric transforms, occlusion, and background clutter, are jointly solved. Moreover, instead of employing hand-designed features this deep network at- tempts to learn not only an optimal feature but also photo-metric and geometric transforms for person ReID task. Besides, compared to the previous studies in which across- view transforms are assumed to be uni-modal, FPNN works with a mixture of complex transforms. Cheng et al.[56] introduced a multi-channel CNN model (MCCNN) con- sisting of three sub-networks (triplet) with the shared parameter set. This target of this network is to pull the instances of the same person closer and to push those of the different person farther from each other in the learned feature space. Additional, each sub-network composes of multiple channels to learn both the global full-body and local body-parts features. Based on this structure, both local and global signatures are employed to generate an effective descriptor for person representation. McLaughlin et al. [57] proposed a novel Recurrent Neural Network (RNN) for video-based person re-identification. In such framework feature vectors are extracted on both image and sequence levels by combining a convolutional neural network and a recurrent layer. The proposed network also has a symmetric structure based on Siamese network to learn the similarity between two image sequences. For this purpose, both appearance and motion information are exploited to generate the unique signature for each sequence. The obtained results are more impressive than those of the previous works, however, three network architectures used in this framework lead to increase processing time as well as computation complexity significantly. An invariant of RNN, Long-Short Term Memory (LSTM) is

also proposed to apply in person re-identification problem in [34]. The aim of LSTM network is to learn image-level features over time steps and create the final sequence-level feature. The authors claimed that although using a low-level feature as LBP-Color, the performance of proposed method is better than that of some existing ones thanks to the effectiveness of LSTM network. In order to reduce the com- putation time and memory requirement, only sub-sequence images are used for person representation. However, this might lead to insufficient information for person repre- sentation when images in a sub-sequence are slightly different from each other. In [58], the authors also proposed to use LSTM network for person ReID framework. In this framework, GOG descriptor is employed for feature extraction at image-level and then, these extracted features are pushed into LSTM network to generate the final signature for a given sequence of images. By exploiting the effectiveness of GOG descriptor and LSTM network, this work has obtained comparable results. In [59], the authors in- tegrated LSTM modules into a Siamese network. By this way, LSTM decides which information is remembered or forgotten and helps the network to be able to process information sequences for a long-time. Deep-learned features extracted from this model are more robust and discriminative for person representation in person ReID problem. Additionally, Varior et al. [60] introduced a new framework by adding a gating function after each convolutional layer to capture effective subtle patterns when working on a pair of testing images. This model can obtain state-of-the-art performance on several benchmark datasets, however, its only drawback is time-consuming due to the pairwise comparison. Similar to [60], Liu et al. [61] incorporated a soft attention based model in a Siamese network to adaptively pay attention to the important local parts of an input image pair, however, this method is also limited by computational inefficiency.

For deep-learned features, the two most strongest features for person ReID problem used in this thesis are extracted from GoogleNet and ResNet networks. These two deep-learned networks are detailed below.

GoogLeNet. GoogLeNet is the winner of the ImageNet Large Scale Visual Recognition Competition 2014 (ILSVRC), an image classification competition [62]. This network has relatively lower error rate compared with the VGGNet (1st runner-up in 2014). This network significantly outperforms ZFNet (The winner in 2013) and AlexNet (The Winner in 2014). As its name, it is widely perceived that it is from Google, and the word "LeNet" is understood for paying tribute to Prof.Yan LeCun. One interesting point is that GoogLeNet is evaluated to be very close to human level performance. GoogLeNet architecture includes 22 deep CNN layers with 4 million parameters. It is usually trained on a subset of ImageNet which is a well-known large- scale dataset consisting of over 15 millions labeled high-resolution images with around 22,000 categories. The subset of ImageNet is used for training GoogLeNet containing around 1,000 images for each of 1,000 categories.

ResNet Residual Neural Network (ResNet) [63], The winner of ILSVRC 2015 with error rate 3.57%, was developed by Microsoft with a deeper architecture. This network has a structure with multiple layers stacked and makes the model be deeper. From the first deep network structure of AlexNet with only 5 convolutional neural layers, the architecture of a CNN is deeper. For example, VGG and GoogLeNet have 1 9 and 22 layers, respectively. Therefore, deep architecture is an inevitable trend for CNNs in the research community. However, increasing the network depth does not simply stack multiple layers together. It is difficult to train a deep network because of vanishing gradient problem. This problem is caused by back propagation and repeated multipli- cation, which make the slope extremely small and the accuracy be saturated or even reduced. With ResNet structure, this issue is solved thanks to these skip connections defined as gated units or gated recurrent units connect the current layer with the pre- vious layer. These skip connections assist on training hundreds even thousands Neural Network layers. Figure 1.11 show a residual learning block with a skip connection and an example for ResNet-50 with 50 deep layers. In additional, batch normalization helps the model to converge more easily. Based on above advantages, ResNet is applied in different pattern recognition problems, such as image classification, object detection, face recognition, and etc.

Although obtaining numerous impressive results, deep-learned network still has its own drawbacks. It requires training on a large-scale dataset. A large number of labeled images is required to achieve a better image

presentation. Some current work proved that deep-learned features are not effective on a small dataset. In this case, person ReID accuracy when using deep-learned features is even lower than that of employing hand-designed features. Therefore, a combination between hand-designed and deep- learned features is inevitable trend for person ReID problem. Fusion schemes try to take advantage of each kind of features and help to improve person ReID performance. Feature fusion strategy might cause a significant burden on memory storage as well as computation time, especially when dealing with large-scale dataset. Fortunately, this can be solved through representative frame selection, instead of working with all frames of each person, only several key frames are chosen for feature extraction and matching. These are the two main contribution of this thesis and will be presented in the later Chapters.

3.4. Metric learning and person matching

Metric learning is one of the two most important steps for person ReID problem.

The main target of metric learning is to find a suitable and effective distance for person matching. It tries to minimize the distances between cross-view images corresponding the same person and maximize those for different persons. Metric learning is also known as learning a sub-space on which projected feature vectors are satisfied above- mentioned conditions.

At present, there are many kinds of metric learning techniques are proposed for person ReID.

First of all, we take a look at study of Weinberger et al. [65] in which the authors suggested the way how to learn a Mahalanobis distance metric for k-nearest neighbor (kNN) classification. The authors indicated that although kNN classifier is one of the most effective method for pattern classification, this method depends strongly on the metric utilized for computing distance between two different objects. For example, if kNN is used for classifying images of faces by age and gender, it is really hard to optimize the same metric distance for these two tasks (age and gender classifications). And, there have been numerous studies proposed different methods for improving the performance of kNN algorithm by learning a suitable and robust metric distance from labeled data [71, 72, 73], called distance metric learning. These studies showed that even a simple linear transformation of the original feature vectors can result in out- standing improvement for kNN classification. In [69], the authors claimed that by using a non-Euclidean distance even less distinctive features are sufficient for achieving a good result on the matching task.

Inspired by these work, Weinberger et al. [65] introduced a novel scheme to gain strength of kNN with the help of learning a Mahalanobis distance, which tried to minimize a loss function consisting of two terms: (1) The first term penalizes large distances between intra-class objects and (2) the second term penalizes small distances between extra-class objects. These distances are shown in the following equation. Denote that ˙xi, ˙xj are the original feature vectors and ˙xJi ,˙xJj are the projected ones, respectively.

Based on the above LMNN algorithm, Dikmen et al. [66] proposed a rejection framework which return no matches if all neighbors are exceed a given distance. For this framework, the authors introduced a novel cost similar to the LMNN scheme, namely Large Margin Nearest Neighbor with Rejection (LMNN-R). Additionally, person ReID is treated as classification problem and optimized by applying a traditional support vector machine (SVM). In order to solve person ReID problem, the authors adopted a universal threshold (τ) for maximize allowed distance for matching image pairs. And, this scheme could be extended to K nearest neighbors case, in which a label is assigned through majority voting of P nearest neighbors within τ . In this study, the authors claimed that although using only color histogram for person representation, LMNN- R scheme still achieved some significant results compared to the previous works on benchmark datasets. However, the main drawback of these methods is computing complexity if working with large scale samples.

In [13], the authors prefer learning effective metrics for person matching. For this work, two metric learning techniques that are Metric Learning to Rank (MLR) and Classification-based Maximally Collapsing Metric Learning (MCML) are combined to take advantage of their complementary, and called Coupled Metric Learning. The extracted feature vectors are projected into a sub-space through MCML technique and then, the ranking loss is minimized based on MLR. By applying fusion scheme on metric learning, person ReID performance is increased significantly. Although with a great effort, the rate of correct matching on single-shot scenario is still much lower than that on multiple-shot scenario and far from realistic requirement. Another effective metric learning technique is RankSVM [67], an improved version of Suport Vector Machine (SVM) technique, which reformulates the person ReID problem as a ranking problem. In addition, this algorithm also learns a metric space in which the highest rank is assigned for potential true matching rather than use a direct metric distance. Although RankSVM algorithm based on structural SVM to learn an effective sub-space, it ignores the role of loss function in its optimization framework.

In [45], the authors suggested a new sub-space on which metric distance and cross- view data are learned simultaneously, called Cross-view Quadratic Discriminant Anal- ysis (XQDA). XQDA is considered as the extended version of Bayesian [74] and Keep It Simple and Straightforward Metric (KISSME) [75] approaches to cross-view metric learning. For KISSME, it is evaluated as a simple but effective strategy for learning a distance metric from equivalence constraints, based on statistical inference perspective. KISSME is different from existing methods which are usually based on complex opti- mization problem and requires expensive computation. In this work, two independent generation processes for observed commonalities of similar and dissimilar pairs are con- sidered. The dissimilarity is defined as the rationality of belonging to one or the other one class. From this, the optimal statistical decision whether a pair (i, j) is dissimilar or not can be computed by a likelihood ration test. In [74], two intra-personal ΩI and extra-personal ΩE variations are defined and the multi-class classification problem is turned into distinguishing these two classes. A model for each of two classes with multi- variate Gaussian distribution is introduced, which concerns a Quadratic Discriminative Analysis (QDA) model with the above variations. This algorithm is firstly applied for face recognition problem. Inheriting the obtained results of these studies, Liao et al.[45] proposed an effective metric learning technique. In 2019, Matsukawa et al. [70] have introduced the improved version of XQDA with standard kernel trick [76], called Kernelized Cross-view Quadratic Discriminant Analysis (KXQDA). The target of ker- nel trick is to project input samples from the original space to Reproducing Kernel Hilbert Space (RKHS) H where the inner-product of sample xi and xj is defined by a kernel function $k(xi, xj) = \varphi(xi)^T \varphi(xi)$ [77]. Relying on this kernel, KXQDA can achieve a higher performance compared to other metric learning techniques.

Person matching

Person matching is the last step before ranking gallery persons corresponding to a given query person. The target of person matching is to provide evaluated scores for the ranking problem. As presented in the above subsection, to calculate the similarity between two persons (one in probe and one in gallery), the distance is can be computed directly or learned through a previously trained model. Metric learning techniques are introduced in more details in the above subsection, and in this subsection, we discuss more about different manners for computing directly distance between two persons in person ReID problem. In the single-shot approach, this distance is can be simply considered as the difference between two extracted feature vectors. Euclidean and Cosine distances are the two most popular methods for estimating the distances between objects in pattern recognition in general.

In the multi-shot approach, this distance would be estimated in different ways. There are two crucial ways to solve this problem: (1) turning multi-shot problem into single-shot one or (2) calculating a unique distance metric for two sets of feature vectors. For the first way, some above-mentioned works have tried to use a pooling layer. The purpose of the pooling layer is to get the unique signature for representing a sequence of images following a prior rule, such as min-, max-, or average pooling [78, 57]. Some other works have learned

features through consecutive frames to generate the final signature for each individual [34]. By this way, each person is represented by a unique feature vector, and multi-shot problem is turned into single-shot one.

For the second way, on one hand, a similar distance is calculated based on two sets of feature vectors to determine whether they represent the same person or not. This distance can be defined as the average [79, 51] or minimum value [37, 42, 80, 81] of all calculated pair-wise distances. However, it takes much time to calculate all pair- wise distances. On the other hand, these sets of feature points are modeled, and the comparison between two sets is turned into structural model comparison [82, 83, 84, 85]. In [84, 85], each set is modeled as a manifold and the target of these works is to maximize manifold margin through learning an embedding space. In [86], the authors modeled each set as a covariance matrix and the distance between two objects is computed as the distance between two matrices.

4. Fusion schemes for person ReID

To leverage the robustness of different features for person representation, some ef- forts have tried to combine several features. This strategy, named feature fusion, has been widely applied in various research fields. Feature fusion are divided into feature- level (early fusion) [87, 88, 89, 90, 91, 92] and score-level fusions (late fusion) [93, 94, 95]. In the early fusion approach, features are concatenated to form a large-dimension vector for representing an image while methods belonging to the second approach combine the weights/scores, obtained from the matching processes for the corresponding features, in a similarity function to get the final score.

Following the early fusion strategy, Gao et al. [87] combined a high-dimensional low-level feature, called as Weighted Histograms of Overlapping Stripes (WHOS) and low-dimensional mid-level feature, namely color name descriptor, for representing a person image. WHOS descriptor is generated by incorporating color histograms and Histogram of Oriented Gradients (HOG). Besides, color name descriptor is built to map RGB values to 11 pre-defined colors consisting of black, blue, brown, grey, green, orange, pink, purple, red, white and yellow. Based on combining these features, useful information assists on creating a robust and discriminative descriptor for person repre- sentation. The authors have proved that the recognition rates are improved when using the combined features with experiments conducted on different datasets. In [89], the authors proposed an early fusion scheme in which both hand-designed and deep-learned features are exploited for image representation. Firstly, human body parts are extracted by employing a Pose Prediction Network (PPN) [96] which will be used for guiding the deep network process for each sub-region. By applying Convolutional Pose Machines (CPM) [97], body joint response maps is created in a coarse-to-fine manner. For each stage, a convolutional neural network is utilized for extracting deep-learned features and then, concatenated with LOMO descriptor [45], employed as hand-designed fea- tures, to build the final signature. One more work in early fusion approach is the study of Xin et al. [98], in which the authors proposed a way to integrate the distribution of color to a descriptor including texture and edge information to get an improvement in person re-identification problem.

For the late fusion strategy, scores for all of each features are used to formulate the final score to decide the rank of each gallery person respect to a given query person. There are different manners to combine matching scores for ranking. Some works have extracted features at different abstract levels, for examples low-, mid-, and high-levels [99, 38]. Inversely, some other works have been interested in extracting features on different parts of images or whole image. And then, late fusion is performed after matching process for individual feature [27, 38]. When working with CNN networks, features extracted on different layers can be incorporated for the higher performance [92, 88]. Next, we will discuss several remarkable studies belonging to this approach in more details.

Eisenbach et al. [93] proved that late fusion methods can achieve better results than early fusion ones. Both early and late fusion schemes are surveyed in this work in different manners. In this work, nine different features are employed for image rep- resentation. After that, their matching scores are fused to create the final

score for ranking problem. This late fusion scheme includes three steps that are score normal- ization, weighting calculation, and fused score generation based on weighted sum. In comparison with early fusion scheme, all these steps consist of only few and simple calculation, therefore, late fusion scheme can be performed much faster. Li et al. [27] improved the performance of person re-identification by fusing scores which are ob- tained in both patch- and image-level. Wu et al. [99] leveraged advantage of different features at different abstract levels (low-, mid-, and high-level). In this work, late fusion strategy is presented in two different ways by exploiting either scores or rank aggregation. Lejbolle et al. [38] computed scores corresponding to parts of a person image (head, torso, and legs) and these scores are combined in a late fusion strategy. For this study, the authors exploited the representing ability of LOMO [45], a cross- view projective dictionary learning (CVPDL) [27], and Feature Fusion Network (FNN) [99]. Additionally, late fusion strategy can be performed in two different ways by using either scores or rank aggregation of the identities. In [92], the authors built an effective descriptor for person by combining feature vectors extracted from different layers in a deep learning network that used ResNet-50 [63]. This method achieves state-of-the- art performance. Nevertheless, a drawback of this strategy is that it has to process on a large dimensional feature vectors. In order to deal with this problem, Liu et al. [88] used a fully connected layer for reducing dimension of GOG features before con- catenating deep learning features extracted from ResNet. Unlike the previous studies, in which features are combined with constant weights, Zheng et al. [95] proposed a method to adaptively fuse features in both image search and person ReID, the feature weighting is changed over query images. Taking into account that features are not equally important for all queries, in [95], weights for different features are determined based on the content of a query image. The obtained results of this technique for the image-based person re-identification are very promising.

Inspired by the research of Zheng el al. [95], this thesis proposes a fusion scheme for multi-shot person ReID. Both hand-designed and deep-learned features are considered to take advantage of them. Beside equal weights assigned to evaluated features, adap- tive weights are examined. The fusion scheme is proposed for both settings of person ReID. In the first setting, person ReID is treated as information retrieval in which the identity corresponding of the given query is determined through the probability of his/her image belonging to each of trained appearance models. In the second setting, a combination of metric learning with fusion scheme is exploited for person ReID.

4.1. Representative frame selection

One more remarkable issue in multi-shot person ReID is to choose the frames used for person representation. Using all frames will make great pressure in computation process as well as storage memory. Multiple images for describing a person are either consecutive or discrete frames. There have been different strategies are proposed to solve this issue for each case. To avoid this problem, some previous studies select representative frames rather than using all frames in a sequences for representing an individual [78, 100, 101].

For the first case, when images for each individual are consecutive frames, some studies take a sub-sequence instead of whole sequence of images. For this target, in [100, 102, 34, 78], walking cycles are extracted based on either analyzing motion energy or tracking super pixels corresponding to the lowest portions of a human body. One of the first studies on walking cycle extraction was proposed by Wang et al. [102]. In this work, a sequence of images is broken down based on motion energy intensity. After that, these video fragments are employed to learn a model which can automatically chose the most discriminative fragments for person representation. Furthermore, to significantly reduce the processing time as well as memory storage requirement, extraction of four key frame is also considered [101].

Different from the above approach, Gao et al. [78] suggested that walking cycle can be extracted by tracking super pixels corresponding to the lowest portions of a human body (feet, or ankles, or legs near ankles). According to this work, the lowest portions of human have most significant and stable motion. Superpixels on the lowest portions of human are extracted in the first frame, and tracked through the consecutive frames to

draw curves corresponding to a human movement. And then, the best cycle is chosen for person representation. Moreover, the authors pointed the two advantages of this work compared to the above works. Firstly, walking cycle is extracted based on superpixel which are more robust and reliable than that based on individual pixels exploited in FEP method. Secondly, the best cycle is chosen in the unsupervised manner, an inevitable trend for any pattern recognition problem. Yan et al. [34] proposed to use Long-Short Term Memory (LSTM) unit, a variant of RNN, to create a unified signature for describing an individual over a sequence of images. By using LSTM units, the important information of each image sequence is retained. In this work, only a sub-sequence consisting of 10 images are used for LSTM network.

In the second case, when images for each person have no temporal linking, temporal information is not exploited to describe a person image. To overcome this difficulty, several studies base on key frame selection to get important information from a group of images [103, 104, 105]. In [103], the authors introduced a framework for key frame selection based on modeling the body's contour variations during the tracking. HOG features are used for describing person appearance and Bhattacharyya distance is em- ployed to calculate the similarity between two consecutive images. Based on these distances, a curve presenting the contour shape variation during a pedestrian's trajec- tory is drawn. After that, the trajectory is automatically segmented by detecting the posture transition and key frames are chosen for each fragment. The goal of the study of Hassen et al. [104, 105] is to keep only informative frames for person representa- tion. Redundant frame and useless information are removed to speed up execution processing. For this, Mean shift clustering algorithm is used to generate groups of sim- ilar images. Additionally, a small cluster which has small number of images/samples also removed because it might contain non-frequent or useless information. From that, only informative frames are kept and then, key frames are chosen as the center of each cluster. These key frames are representative for each person and used for ReID process. In summary, the above-mentioned works focus on extracting representative key frames for person representation for video-based person ReID, On one hand, this approach helps to reduce complexity and computation time. However, on the other hand because only several representative frames are selected to represent each pedestrian, it results in loss of information, which is a drawback of this strategy. Therefore, accuracy and consuming time are the two sides of the one issue in person ReID, and should be considered.

4.2. Fully automated person ReID systems

As we known in advance, a fully automated system contains not only person ReID phase but also other crucial phases of human detection and tracking. In fact, there are a few works focusing on this in contrast to a wide range of other ones for the only phase of person ReID. In [106] , the authors are interested in evaluating the entire system of person ReID on surveillance data captured from multiple cameras. These cameras have non-overlapping FOVs covering a wide area of moving paths. In this proposal, some advanced techniques of auto human detection, tracking and ReID are applied, such as DPM (Deformable Part-Based Model) and HOG (Histogram of Oriented Gradients) for detection, Tracking-by-Detection for tracking and gait feature for ReID. Zheng et al [107] introduced comprehensive baselines for end-to-end person ReID in raw video frames. A novel dataset is provided in this work, called Person Re-identification in the Wild (PRW), and extensive experiments are conducted by combining various de- tectors and recognizers to improve the overall person re-identification performance. In [108], the authors claimed that if background removal is performed directly by ap- plying binary masks which might cause loss of information and results in a slightly worse performance compared to case of using the original images. Consequently, back- ground should be removed in the feature-level extraction. It is called as Mask-Guided Contrastive Attention Model (MGCAM) with a binary mask considered as an addi- tional input which is accompanied by an RGB image to enhance feature learning. The effectiveness of this method is proven by impressive results on several public datasets. In this thesis, towards to a complete person ReID system, the author consider other crucial phases that are person detection and segmentation. The effect of person de- tection and segmentation on person ReID performance is carefully evaluated on both single shot and multi-shot scenarios. Three state-of-the-art person detection meth- ods that are Aggregate Channel

Features (ACF) [109], You Only Look One (YOLO) [110], and Mask R-CNN [111] are exploited. Besides, Pedparsing [112] method is also used for person segmentation to eliminate the affection of background information on performance of person ReID.

5. Conclusion

In Vietnam, several studies for person ReID have been introduced over the last decade. Research groups are at Hanoi University of Science and Technology (HUST); University of Information Technology (UIT), Vietnam National University - Ho Chi Minh City (VNU-HCM); Da Nang University of Science and Technology have reported some remarkable results. First of all, we take a look several studies of a group at HUST. With the observation that background of an image might cause noise and re- duce person ReID accuracy, Nguyen et al. in [113] proposed a framework in which several background removal methods are used. Moreover, inspired by work of Zhao et al. [43] the authors exploited saliency information for person representation step. In this study, background removal is performed manually via Interactive Segmenta- tion Tool [114] or automatically by using an elliptical/local salience-based binary mask on each person image. By applying these background removal methods, the matching rates at rank-1 are 27.18%, 24.81%, 23.80% compared to 20.00% in the original method [43]. Another work of this group relates to evaluating performance of several Recurrent Neural Network (RNN) variants in both terms of accuracy and number of parameters of each network [115]. Three Long-Short Term Memory (LSTM) variants and Gated Recurrent Unit (GRU) on Caffe deep learning framework are implemented. From the obtained results in this work, Le et al. suggest that GRU is the best choice because this network can achieve the highest accuracy with fewer trained parameters. In this case, the matching rates at rank-1 GRU are 59.2% and 48.4% which are higher by 5.4% and 2.2% compared to the second best results (LSTMC) on PRID-2011 and iLIDS-VID datasets, respectively. Another research group also belong to HUST, in International Research Institute MICA. Pham et al. have achieved some remarkable results on per- son ReID [116, 26]. In [116], the authors provided an improved version of Kernel Descriptor (KDES), which was first introduced by Bo et al. [117], for person appear- ance representation. KDES is evaluated as one of the most robust and discriminative descriptor for image representation. Pixel information containing its gradient, color, and texture is exploited. And then, this information is summarized in three different pyramid scales. By this way, person ReID performance is improved significantly. Some experiments are conducted on CAVIAR4REID, iLIDS-VID, and their own datasets. Matching rates at rank-1 are increased by 6.05%, 7.00% and 4.42% compared to those when employing KDES descriptor on CAVIAR4REID, iLIDS, and their own datasets, respectively. Based on these results, Pham et al. proposed a fully-automated person ReID system [26] including human detection and person ReID steps. Beside taking advantage of the improved version KDES descriptor introduced in [116], the authors suggested to use an effective shadow removal method in the proposed framework to achieve a higher person ReID performance. Inspired by the obtained results in the studies of Pham et al., in this thesis, the author propose to use KDES descriptor in the feature extraction in a fusion scheme and extend the above framework for the first setting of person ReID.

The second research group that has been studying on person ReID for recent years is at UIT, VNU-HCM. Nguyen et al [118, 119] proposed a new method to solve person ReID based on the relationships between attributes of person images, called Attribute Re-Scoring (ARS). The authors defined a set of attribute relationships by observing on the evaluated dataset. For example: Male → No skirt, this means a person who is a man does not likely wear a skirt. In this work, the authors conducted experiments on two benchmark datasets: VIPeR and PRID-2011. For VIPeR dataset, 21 attributes [120] are used for learning attributes relationships. However, for PRID-2011, only 17 attributes are exploited. From this process of learning attribute relationships, the accuracy of attribute detection is improved and then, performance person ReID is higher. In 2017, with the assumption that the probe persons appear at the same time, this research group introduced a re-ranking algorithm on the ranked lists obtained after the matching steps [121]. In this work, two penalty functions are proposed to generate new score for each gallery person corresponding to the given probe person.

Based on these new scores, the original lists are sorted and help to improve person ReID accuracy. Another group is also at UIT, VNU-HCM, Nguyen et al. [122] proposed a framework in which Deformable Part Models is exploited for human body parts detection.After that, feature extraction step is performed on each human body part and then, extracted feature vectors on all parts are concatenated to generate the final signature for person representation. In this work, HSV histogram and MSCR are used in feature extraction step. The similarity between two images is defined as the linear function of the similarities between their corresponding features. Several experiments are conducted on ETHZ dataset to prove the effectiveness of the proposed framework. Following the metric learning approach, in [123] a non-linear manifold space is ex- ploited with the purpose of learning an optimal distance to distinguish different persons more easily. The remarkable point of this study is to build a complete neighborhood graph based on the distances between two patches belonging to two cross-view images. From this, a weight matrix is generated and used for learning a new sub-space, called manifold space. In order to show the effectiveness of the proposed framework, experi- ments are conducted on VIPeR dataset. The matching rate at rank-1 is 23.9% higher than those obtained in [29] and [14].

From the above analysis we realize that most of studies in Vietnam performed on single-shot approach (VIPeR, ETHZ datasets). Only work of Le et al. [115] pays attention on video-based approach. Besides, the above obtained results are much lower compared to the-state-of-the-art ones. These are the motivation for the author to improve person ReID performance in the PhD course.

References

i. *Gong S., Cristani M., Loy C.C., and Hospedales T.M. (2014). The re- identification challenge. In Person re-identification, pp. 1–20. Springer.*
ii. *Wang X. (2013). Intelligent multi-camera video surveillance: A review . Pattern recognition letters, 34(1):pp. 3–19.*
iii. *Gheissari N., Sebastian T.B., and Hartley R. (2006). Person reidentification using spatiotemporal appearance. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pp. 1528–1535. IEEE.*
iv. *Bedagkar-Gala A. and Shah S.K. (2014). A survey of approaches and trends in person re-identification. Image and Vision Computing , 32(4):pp. 270–286.*
v. *Vezzani R., Baltieri D., and Cucchiara R. (2013). People reidentification in surveillance and forensics: A survey . ACM Computing Surveys (CSUR), 46(2):p. 29.*
vi. *Satta R. (2013). Appearance descriptors for person re-identification: a compre- hensive review . arXiv preprint arXiv:1307.5748 .*
vii. *Gou M., Wu Z., Rates-Borras A., Camps O., Radke R.J., et al. (2018). A sys- tematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. IEEE transactions on pattern analysis and machine intelligence, 41(3):pp. 523–536.*
viii. *Leng Q., Ye M., and Tian Q. (2019). A survey of open-world person re- identification. IEEE Transactions on Circuits and Systems for Video Technology .*
ix. *Zheng L., Yang Y., and Hauptmann A.G. (2016). Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 .*
x. *Perronnin F. and Dance C. (2007). Fisher kernels on visual vocabularies for image categorization. In 2007 IEEE conference on computer vision and pattern recognition, pp. 1–8. IEEE.*
xi. *Chang Y.C., Chiang C.K., and Lai S.H. (2012). Single-shot person re- identification based on improved random-walk pedestrian segmentation. In In- telligent Signal Processing and Communications Systems (ISPACS), 2012 Inter- national Symposium on, pp. 1–6. IEEE.*
xii. *Wei Y.L. and Lin C.H. (2013). Single-shot person re-identification by gaussian mixture model of weighted color histograms. In Intelligent Signal Processing and Communications Systems (ISPACS), 2013 International Symposium on, pp. 47–50. IEEE.*

xiii. *Li W., Wu Y., Mukunoki M., and Minoh M. (2013). Coupled metric learning for single-shot versus single-shot person reidentification. Optical Engineering , 52(2):p. 027203.*

xiv. *Farenzena M., Bazzani L., Perina A., Murino V., and Cristani M. (2010). Person re-identification by symmetry-driven accumulation of local features. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 2360– 2367. IEEE.*

xv. *Bazzani L., Cristani M., Perina A., Farenzena M., and Murino V. (2010). Multiple-shot person re-identification by hpe signature. In Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 1413–1416. IEEE.*

xvi. *Zheng W.S., Gong S., and Xiang T. (2012). Transfer re-identification: From person to set-based verification. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2650–2657. IEEE.*

xvii. *Cancela B., Hospedales T.M., and Gong S. (2014). Open-world person re- identification by multi-label assignment inference.*

xviii. *Zheng W.S., Gong S., and Xiang T. (2015). Towards open-world person re- identification by one-shot group-based verification. IEEE transactions on pattern analysis and machine intelligence, 38(3):pp. 591–606.*

xix. *Liao S., Mo Z., Zhu J., Hu Y., and Li S.Z. (2014). Open-set person re- identification. arXiv preprint arXiv:1408.0872 .*

xx. *Wang H., Zhu X., Xiang T., and Gong S. (2016). Towards unsupervised open- set person re-identification. In 2016 IEEE International Conference on Image Processing (ICIP), pp. 769–773. IEEE.*

xxi. *Chen Y., Zhu X., and Gong S. (2018). Deep association learning for unsupervised video person re-identification. arXiv preprint arXiv:1808.07301 .*

xxii. *Ye M., Ma A.J., Zheng L., Li J., and Yuen P.C. (2017). Dynamic label graph matching for unsupervised video re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pp. 5142–5150.*

xxiii. *Ma X., Zhu X., Gong S., Xie X., Hu J., Lam K.M., and Zhong Y. (2017). Person re-identification by unsupervised video matching . Pattern Recognition, 65:pp. 197–210.*

xxiv. *Liu Z., Wang D., and Lu H. (2017). Stepwise metric promotion for unsuper- vised video person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2429–2438.*

xxv. *Peng P., Xiang T., Wang Y., Pontil M., Gong S., Huang T., and Tian Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In Pro- ceedings of the IEEE conference on computer vision and pattern recognition, pp. 1306–1315.*

xxvi. *Pham T.T.T., Le T.L., Vu H., Dao T.K., et al. (2017). Fully-automated person re- identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method . Image and Vision Computing , 59:pp. 44– 62.*

xxvii. *Li S., Shao M., and Fu Y. (2015). Cross-view projective dictionary learning for person re-identification. In Twenty-Fourth International Joint Conference on Artificial Intelligence.*

xxviii. *Karanam S., Gou M., Wu Z., Rates-Borras A., Camps O., and Radke R.J. (2018). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. IEEE Transactions on Pattern Analysis & Machine Intel- ligence, (1):pp. 1–1.*

xxix. *Gray D. and Tao H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In European conference on computer vision, pp. 262–275. Springer.*

xxx. *Cheng D.S., Cristani M., Stoppa M., Bazzani L., and Murino V. (2011). Custom pictorial structures for re-identification. In British Machine Vision Conference (BMVC 2011).*

xxxi. *Das A., Chakraborty A., and Roy-Chowdhury A.K. (2014). Consistent re- identification in a camera network . In European Conference on Computer Vision (2014), pp. 330–345. Springer.*

xxxii. Hirzer M., Beleznai C., Roth P.M., and Bischof H. (2011). Person re-identification by descriptive and discriminative classification. In Scandinavian conference on Image analysis (2011), pp. 91–102. Springer.