Ph.D. in Electronic and Computer Engineering

Dept. of Electrical and Electronic Engineering

University of Cagliari

Person Re-Identification Techniques for Intelligent Video Surveillance Systems

Bahram Lavi Sefidgari

Supervisor: Prof. Giorgio Fumera

Co-supervisor: Prof. Fabio Roli

Curriculum: ING-INF/05 - Sistemi di Elaborazione delle Informazioni

XXX Cycle

February 2018

Ph.D. in Electronic and Computer Engineering

Dept. of Electrical and Electronic Engineering

University of Cagliari

Person Re-Identification Techniques for Intelligent Video Surveillance Systems

Bahram Lavi Sefidgari

Supervisor: Prof. Giorgio Fumera

Co-supervisor: Prof. Fabio Roli

Curriculum: ING-INF/05 - Sistemi di Elaborazione delle Informazioni

XXX Cycle

February 2018

I dedicate this dissertation to the memory of my mother, a strong and gentle woman whom I still miss every day.

Acknowledgments

"We don't read and write poetry because it's cute. We read and write poetry because we are members of the human race. And the human race is filled with passion. And medicine, law, business, engineering, these are noble pursuits and necessary to sustain life. But poetry, beauty, romance, love, these are what we stay alive for."
Quote by Robin Williams, movie of *Dead Poets Society*.

First and foremost, I would like to thank my supervisor Prof. Giorgio Fumera for his general support, contribution, and funding to make my Ph.D. experience productive. Besides my supervisor, my sincere thanks goes to Prof. Fabio Roli as director of the *PRA Lab* and also as my co-supervisor, who provided me an opportunity to join his lab to pursue my Ph.D degree.

My special thanks goes to my closest friend Mehdi Fatan for the sleepless nights we were uninterrupted discussing and working together through his many interesting ideas in Computer Vision area; I also thank my friends Farshid Azizi, Temitope Asagunla, and Reza Farmani, Dragana Jadzic, and Sonia Aldana for their kind supports and encouragements who made it possible to keep moving on my little own way, and my fellow labmate Mansour Ahmadi for all the fun we have had in the last two years of my Ph.D. program. I need to thank so many other friends that for lack of space I cannot list them here.

Completion of this Ph.D. dissertation was possible with the support of several people, when I met with some difficulties during my studies.; and I therefore would like to express my sincere gratitude to all of them; Dr. Gabriela Hoff, Eleni Merkoury, and Serenella Putzu,

have all extended their support in a very special way, and I gained a lot from them, through their personal and scholarly interactions.

Last but not least, I would like to deeply thank my family for all their love and encouragement; for parents who raised me with a love of science and supported me in all my pursuits; my brother and sisters for supporting me spiritually throughout writing this thesis and my life in general.

Bahram Lavi
University of Cagliari
February 2018

Abstract

Nowadays, *intelligent video-surveillance* is one of the most active research fields in computer vision and machine learning techniques which provides useful tools for surveillance operators and forensic video investigators. Person re-identification is among these tools; it consists of recognizing whether an individual has already been observed over a network of cameras. This tool can also be employed in various possible applications, e.g., off-line retrieval of all the video-sequences showing an individual of interest whose image is given as query, or on-line pedestrian tracking over multiple cameras. For the off-line retrieval applications, one of the goals of person re-identification systems is to support video surveillance operators and forensic investigators to find an individual of interest in videos acquired by a network of non-overlapping cameras. This is attained by sorting images of previously observed individuals for decreasing values of their similarity with a given probe individual.

This task is typically achieved by exploiting the clothing appearance, in which a classical biometric methods like the face recognition is impeded to be practical in real-world video surveillance scenarios, because of low-quality of acquired images. Existing clothing appearance descriptors, together with their similarity measures, are mostly aimed at improving ranking quality. These methods usually are employed as part-based body model in order to extract image signature that might be independently treated in different body parts (e.g. torso and legs). Whereas, it is a must that a re-identification model to be robust and discriminate on individual of interest recognition, the issue of the processing time might also be crucial in terms of tackling this task in real-world scenarios. This issue can be also seen from two different point of views such as processing time to construct a model (aka *descriptor generation*); which usually can be done off-line, and processing time to find the correct individual from bunch of acquired video frames (aka *descriptor matching*); which is the real-time procedure of the re-identification systems.

This thesis addresses the issue of processing time for descriptor matching, instead of improving ranking quality, which is also relevant in practical applications involving interaction with human operators. It will be shown how a trade-off between processing time and ranking quality, for any given descriptor, can be achieved through a multi-stage ranking approach inspired by multi-stage approaches to classification problems presented in pattern recognition area, which it is further adapting to the re-identification task as a ranking problem. A discussion of design criteria is therefore presented as so-called multi-stage re-identification systems, and evaluation of the proposed approach carry out on three benchmark data sets, using four state-of-the-art descriptors. Additionally, by concerning to the issue of processing time, typical dimensional reduction methods are studied in terms of reducing the processing time of a descriptor where a high-dimensional feature space is generated by a specific person re-identification descriptor. An empirically experimental result is also presented in this case, and three well-known feature reduction methods are applied them on two state-of-the-art descriptors on two benchmark data sets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The importance of security and safety of people in crowd is continuously growing day by day in our society. Private/public companies, governments, public areas such as airports and malls, etc., are seriously going along this need which requires too much expenses and efforts. To accomplish this goal, video surveillance systems are playing a key role in this manner. These days, plenty of video cameras are growing everywhere which is an useful tool for addressing a different kind of security issues such as forensic investigations, crime preventing, and safeguard of the environments. Recording massive quantity of video frames from network camera per day is one the major critical problems of video surveillance systems. This due to by monitoring and analyzing the acquired videos from tens or hundreds of camera which are all needed to be done by surveillance operators at the same time.

Intelligent video surveillance systems aim to automate the issue of monitoring and analyzing the videos from camera networks to help the surveillance operators in handling and understanding the acquired videos by camera networks. This is one the most active and challenging research area in computer engineering and computer science in which computer visions and machine learning techniques are required. This field of research enables some various tools such as: recognizing a suspicious actions, on-line pedestrians tracking, off-line forensic investigations. In this manner, person re-identification has been proposed as a tool of intelligent video surveillance systems; which consists of recognizing an individual over a video surveillance camera network with non-overlapping fields of view [6, 56]. Figure 1.1 shows an example of a video surveillance camera network with non-overlapping fields of view.

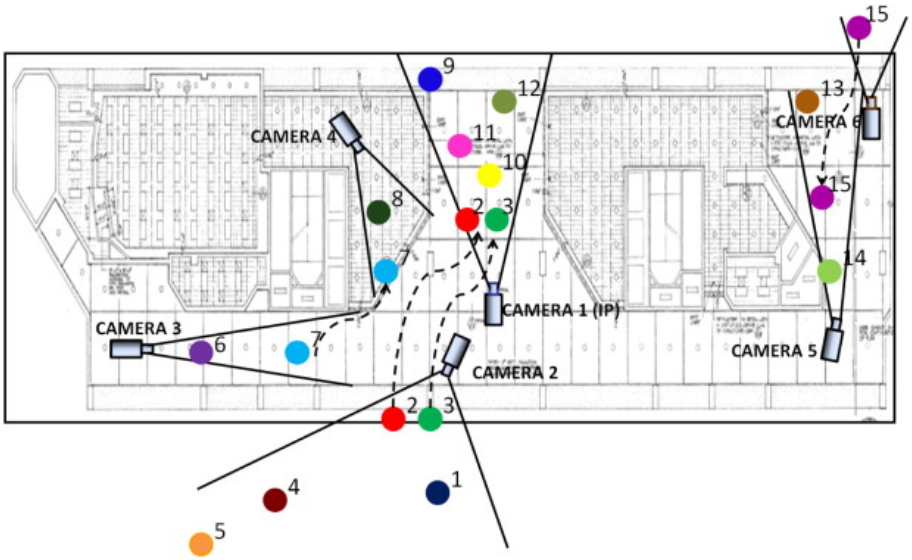

Figure 1.1: Example of multi-camera surveillance illustrated for person re-identification [5].

As pointed out, one of the applications of person re-identification is to support surveillance operators and forensic investigators in retrieving video frames showing an individual of interest, given an image as a query (aka *probe*). To this aim, the video frames or tracks of all the individuals (aka *template gallery*) recorded by the camera network are ranked in order of decreasing similarity to the probe, to allow the user to find out the individual of interest (if any) ideally in top positions. Figure 1.2 demonstrates a typical scheme of this re-identification scenario. This is a challenging task in a typical video surveillance system, due to low image resolution, unconstrained pose, illumination changes, and occlusions, which do not allow to exploit strong biometric techniques like face recognition (see Fig. 1.3). Clothing appearance is therefore the most widely used cue; other cues like gait and anthropometric measures have also been investigated.

Most of the existing techniques are based on defining a specific descriptor of clothing appearance (typically including color and texture), and a specific similarity measure between a pair of descriptors (evaluated as a *matching score*) which can be either manually defined or learnt from data [28, 21, 31, 6, 50]. On the other hand, with considering to the great sucess of deep learning in image classification [37], some authors have been attempting also to employ deep learning techniques on person re-identification task [70, 44].

1.1 Person re-identification scenario

Apart from the methodology of a specific person re-identification technique (i.e. appearance-, gait-, or anthropometric-based techniques), this task generally consist of three main steps (see Fig. 1.4):

i The individual of interest must be separated from other part of the image.

ii A significant image representation must be generated, and

iii Finally, the similarity scores of the generated image representations most be computed between the query image and the gallery set.

Among above-mentioned steps, since the first step is not the main challenge of person re-identification task, at this thesis work, only the last two steps of re-identification task will be considered and discussed.

Let D denote a descriptor for person re-identification, $m(\cdot,\cdot)$ the corresponding similarity measure between a pair of images, t the processing time for computing it, $\mathbf{T}$ and $\mathbf{P}$ the descriptors of a template and probe image, respectively, and $G = \{\mathbf{T}_1,\ldots,\mathbf{T}_n\}$ the template gallery. For a given probe $\mathbf{P}$, a standard re-identification system computes the matching scores $m(\mathbf{P},\mathbf{T}_i)$, $i = 1,\ldots,n$, and returns the list of template images ranked in order of decreasing values of their score. Ranking accuracy is widely evaluated using the cumulative matching characteristic (CMC) curve, defined as the probability that the correct identity is

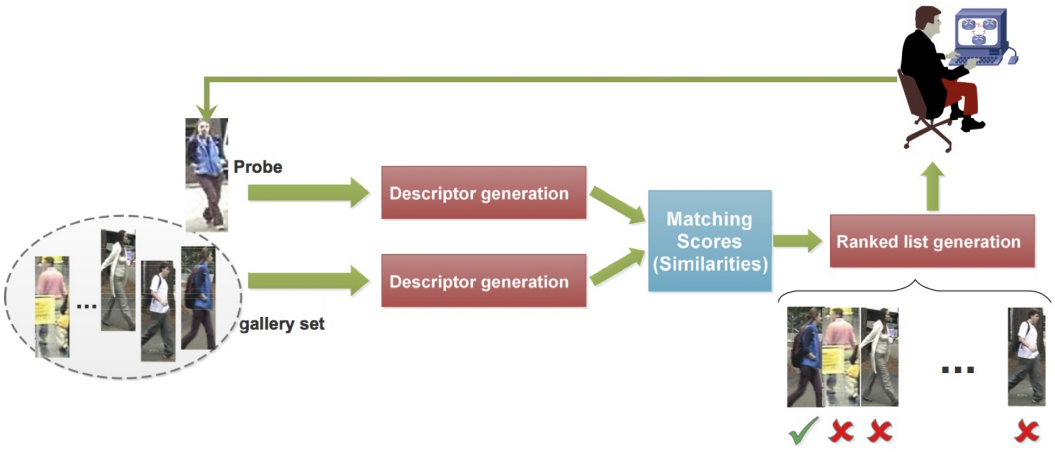

Figure 1.2: A standard re-identification system for the application of an off-line support of a human operator.

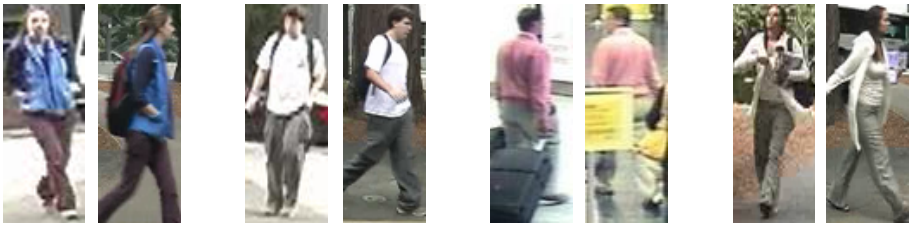

Figure 1.3: Sample images of a video-surveillance camera, taken from VIPeR [28] and i-LIDS [7] data sets: low image resolution, uncostrained poses, and occlusions.

within the first r ranks, for $r = 1,\ldots,n$. By definition, the CMC curve increases with r, and equals 1 for $r = n$.

The main focus of existing works in this field is to attain a high ranking accuracy. Processing time is an issue which has received much less attention so far, instead (to our knowledge, only in [57, 18, 36]), despite its relevance in practical applications involving interaction with human operators, like the ones mentioned above. Many of the existing similarity measures (e.g. standard or learnt from data) are indeed rather complex, and require a relatively high processing time, e.g., [21, 58, 50, 45]. On the other hand, in real-world applications the template gallery can be very large, and even if the processing time for a single matching score is low (e.g., the Euclidean distance between fixed-length feature vectors [50]), evaluating the matching scores for all the templates can be relatively time-consuming.

1.2 Motivation and scope

As mentioned above, the consequence of person re-identification becomes more challenging and difficult due to some image handling issues; although many researches have been

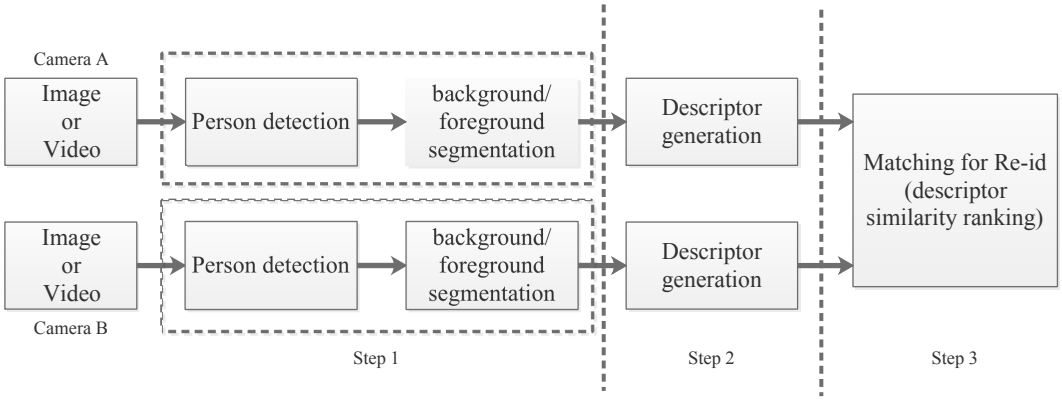

Figure 1.4: Diagram of general re-identification systems.

done in this field, several problems still not resolved/solved [56]. For instance, consider the re-identification in public areas such as shopping malls, airports,etc. which are different from one in a private place such a private company. Whereas, a re-identification system must be robust in handling of illumination changes, pose variation, etc., it must be also swift enough to tackle this task in a real-world application. This thesis work, instead, aims to consider this issue of processing time in person re-identification task, and investigate possible solutions to reduce the processing time of this problem within the online stages.

One possible solution to reduce processing time is to reduce the complexity of a given descriptor and/or of the associated similarity measure; however, this is likely to reduce ranking accuracy as well. A known approach in the pattern recognition field, in particular for supervised classification systems, to trade a lower classification accuracy for a lower processing time, is to use a multi-stage architecture (e.g., [62, 66]). Inspired by this approach, I investigate whether and how a multi-stage architecture can be seen as a ranking problem and exploited to attain an analogous trade-off between ranking accuracy also in person re-identification systems. In particular, I focus on attaining such a trade-off for any, *given* descriptor, without limit on the type of descriptor.

Since existing multi-stage solutions cannot be directly applied to person re-identification, which involves a *ranking* problem rather than a *classification* one, I first provide a formalization of multi-stage ranking systems: I develop an analytical model of their processing time, and discuss the behaviour of the corresponding ranking accuracy, measured through the CMC curve. Based on our model, I then discuss and propose practical design criteria for multi-stage person re-identification systems, considering applications requirements given in terms of strict constraints on the maximum allowed matching processing time. The main contribution of this thesis work is the extension of the multi-stage architecture used in pattern classification to being utilized on person re-identification task (using any *given* descriptor and similarity measure), by formalizing the underlying multi-stage approach as a ranking problem; and a practical design criteria is suggested to attain a significant trade-off between recognition time and processing time.

As an alternative solution, one can also apply a feature reduction method to reduce the dimensionality of a feature vector in order to reducing the cost of processing time. It is noted that feature reduction methods can only be applied for specific type of descriptors (i.e. a

descriptor that generates a fixed-size feature vector). The feature reduction methods nevertheless employ further in the context of the multi-stage approach at this thesis work, they (*any of them*) can be individually also used alone to reduce processing time of a given re-identification descriptor. It therefore leads this thesis work to investigate also on some typical feature reduction methods, by comparing their behaviors on person re-identification data sets.

1.3 Outline of the Thesis

This thesis is structured as follows. I first summarize and review some related works on standard person re-identification techniques in chapter 2, including a brief survey on deep learning techniques for person re-identification task.

In chapter 3 the proposed multi-stage ranking approach is formalized and then a design criteria is developed for multi-stage person re-identification systems. Along with technical details of the implementation I present also the experimental evaluations on three benchmark data sets, using four state-of-the-art descriptors.

In chapter 4, some feature reduction methods are discussed and compared their behaviour on fixed-size features for processing time reduction purposes.

Chapter 5 concludes the thesis with suggesting directions for future research.

1.4 List of Publications Related to the Thesis

1.4.1 Journal paper

- [40] B. Lavi, G. Fumera, and F. Roli, *A Multi-Stage Ranking Approach for Fast Person Re-Identification*. The journal of the Institution of Electrical Engineers Computer Vision (IET CV), 2017. (Relation to Chapter 3)

1.4.2 Conference papers

- [38] B. Lavi, G. Fumera, and F. Roli, *A Multi-Stage Approach for Fast Person Re-Identification* in *International Workshop on Structural and Syntactic Pattern Recognition (SSPR 2016) and Sta-*

tistical Techniques in Pattern Recognition (SPR 2016) Merida-Mexico, 30th of September 2016, http://www.s-sspr.org (Relation to Chapter 3)

- [41] B. Lavi, M. Fatan Serj, and D. Puig Valls, *Comparative Study of the Behaviour of Feature Reduction Methods in Person Re-identification Task.* The International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2018. (Relation to Chapter 4)

1.4.3 In press. paper

- B. Lavi, *Deep Learning Techniques on Person Re-Identification Task: A Survey*, (Relation to Chapter 2)

Chapter 2

Literature review

In this chapter, many of the person re-identification techniques are presented and discussed. Although, some steps are essential in this area before applying a re-identification method such as human-body detection, background/foreground segmentation, shadow elimination, etc., these steps of person re-identification task must be done off-line as a pre-processing step. Thus, they will not be considered at the rest of this thesis. At the following, I discuss on major issues which is concerned by most of the researchers in this this field; e.g. *descriptor generation* and *similarity computation*. As pointed out, many descriptors on person re-identification task rely on generating an image signature based on clothing appearance of individuals. Therefore, first a summary of literature works is given based on existing person re-identification methods, by discussing their advantages and disadvantages; Then, I take a short journey through the existing deep learning techniques for person re-identification task, in which existing neural network models is discussed (e.g. classification, Siamese model). Next a brief categorization of the person re-identification techniques is presented in terms of the efficiency of matching scores which has been considered on their works (a comprehensive discussion will be presented at chapter 3). And finally, some well-known benchmark data sets are discussed at the end of this chapter.

2.1 Standard techniques for person re-identification

Person re-identification consists of matching individuals from different camera network, possibly non overlapping views. It can provide some useful applications such as off-line retrieval of video sequences to find out an individual given as a query, and on-line pedestrians tracking. Clothing appearance is one the most widely cue among the researchers in this area, and therefore many methods have

been proposed based on the people's appearance in order to generating an image signature. The proposed appearance-based methods can be subdivided into four main categories: color-histograms-, interest-points-, convariance-, and textural-based descriptors. At the following of this section, the appearance-based methods for generating individual's signatures will be discussed.

2.1.1 Descriptor generation

At the following, some existing approaches for generating clothing appearance descriptors are presented.

Color-histogram-based descriptors

Color histogram is a popular way to describe an image in terms of the occurrence frequency of colors. Some methods employed this approach to describe an individual of interest. But, for better performance of this technique in person re-identification task, that arises to compute the histogram within the segmented part of image (i.e. detected human body). The color-histogram-based descriptors are typically defined in different color spaces; RGB [75, 28], HSV [34, 31, 58, 21], and LAB [31, 13] color histogram based descriptors. Among these color spaces, HSV color histogram is robustness because of its promising results in person re-identification task. In [58], the proposed descriptor subdivides body into torso and legs, and extracts some randomly positioned image patches from each part. Each patch is represented by HSV histogram. Artificial patches are also generated to improve robustness to illumination changes, by changing the brightness and contrast of the original patches in the RGB color channel.

Additionally, the proposed works in [29, 49] suggest to divide the image into some horizontal stripes where the color histograms can be extracted from each strips, separately, and a simple concatenation among all the histograms can be represented as final image signature. They believe this leads the color histogram to make the descriptor more discriminant.

Interest-point-based descriptors

The idea is to find out large amount of interest points (*aka key points*) including high information contents about color and structural information around regions. However, descriptors based on interest point are applicable in terms of pose variations and illumination changes, but, the redundancy of interest points as well as sensitively on the edges are not desirable, which must be taken into account. Martinel et al. [51] employed scale-invariant feature transform (SIFT) [48] in which the in-

terest points as the centers of circular regions, and a Gaussian function employed for constructing a weighted color histogram from the interest points. The work propsoed in [14], used speed up robust feature SURF [4] to determine and locate the interest points which also contains the HSV histogram information of each point.

Convariance-based descriptors

Covariance descriptors have been employed in person re-identification task for handling of noise and in-variance to be a proportional shifting of color [12, 2, 30, 19]. In [12], the spatial convariance regions (SCRs) descriptor has been proposed in which the location of the point by concerning to RGB color values, orientations, and gradient's magnitudes combined to generate the final image representation. This method is robust to handle the illumination change, and pose variation. Hirzer et al. [30] proposed a methodology to generate the covariance descriptor by subdividing the image into horizontal patches. For the pose handling, their final feature vector contains y position, LAB color channels, and vertical/horizontal derivation of the luminance channel. The convariance descriptors are robust in illumination changes, pose variations, and dense representation from overlapped regions. For the convariance-based descriptors, it is worth to point out that the generated features are not meant in Euclidean space, since it does not contain a special structure, and therefore they suggest to use the mean of covariance of each regions to compare two covariance descriptors, instead of whole descriptor at the same time[53].

Textural-based descriptors

Typically, this kind of descriptors are used as a complementary feature vector to construct appearance-based descriptors. The extracted textural features usually combined with color features to improve the recognition accuracy in person re-identification task, which robust on pose variation and rotation change. For instance, Farenzena et al. [21] utilized recurrent high-structured patches (RHSP) descriptor on this purpose. The descriptor selects some patches from the segmented foreground , and transform their invariance through geometric variations. Also, Gabor [24] and Schmid [59] filters have been applied for this kind of descriptors (e.g. Ma et al. [50]). These filters are also robust on pose variation and rotation change.

2.1.2 Similarity computation

A standard re-identification system computes the matching scores between a query image and all the images in gallery set either by manually defining a similarity measurement or leaning from data. Many of the existing similarity measures relied on applying common distance metrics and nearest-neighbor approach to compute the similarity scores [27, 65]. The Euclidean and the Bhattacharyya distance measures are usually employed depending on the type of specific descriptor [28, 13]. Farenzena et al. [21] applied a simple linear combination to merge the matching scores of three different descriptors by taking into account of a suitable weights for each descriptor. Satta et al. [58] utilized the Hausdorff distance as distance measurement to avoid sensitivity of outlying elements by adopting the k-th Hausdorff distances in which takes the k-th ranked distance rather than the maximum value. This requires a high computational cost because of the use of the Hausdorff distance as similarity measure, which makes the processing time proportional to the *square* of the number of patches.

On the other hand, learning a good metric from data recently becomes in attention of many researchers in this area. A metric learning algorithm usually helps in boosting re-identification performance. It is worth to point out that all these kind of learning methods require a supervised (i.e. labelled data) training set; for instance the template gallery requires a fixed sample set size; templates cannot be added during system procedure. However, this strategy might be too comprehensive enough to tackle with real-world application scenarios. In [16], Large Margin Nearest Neighbor (LMNN) proposed to obtain an omptimized metric for nearest neighbor classification in which support vector machine (SVM) was employed. Inspired by [16], Hirzer et al. [31] utilized the relaxed pairwise metric learning (RPML) method in which a distance matrix M is automatically estimated from a training set and further used it in the matching steps. This significantly takes into account of the body parts with the highest priority are chosen by taking higher weights, through the matrix M. Zheng et al. [75] proposed the novel probabilistic relative distance comparison (PRDC) model for triplet images aiming to minimize the distance of a pair of correct matches and maximize it with a wrong match pair. In [34], the score-level fusion proposed which use linear logistic regression (LLR) and the likelihood ratio between positive and negative samples for high generalization capability.

2.2 Deep learning techniques for person re-identification

This section gives a taxonomy of some recent works which concern deep learning techniques for person re-identification task. The bulk of interesting deep learning works proposed to improve the performance of person re-identification which have done either by modifying the existing deep learn-

ing architecture of the network, or proposing a new one. Generally speaking, two types of deep learning models have been employed in this research task: (i) a classification model for person re-identification problem, and (ii) the Siamese models based on pairwise or triplet comparisons. This has been started by employing a Siamese network for pairwise comparison of input pair of images for person re-identification task.

The deep learning methods in person re-identification are still suffering from the lack of training data. Some of the person re-identification data sets provide only two images for each individual (i.e. VIPeR data set [28]). Probably, the Siamese models have been mostly chosen because of the lack of training samples within the existing person re-identification data sets[74].

2.2.1 Classification models

Xiao et al. [69] proposed learning deep features representations from multiple data sets by using CNNs to discover effective neurons for each training data set. They first produced a strong baseline model that works on multiple data sets simultaneously by combining the data and labels from several re-identification data sets together and trained the CNN with a softmax loss. Next, for each data set, they perform the forward pass on all its samples and compute for each neuron its average impact on the objective function. Then, they replaced the standard Dropout with the deterministic Domain Guided Dropout in order to discarding useless neurons for each data set, and continue to train the CNN model for several more epochs. Some neurons are effective only for a specific data set which might be useless for another one, this caused by data set biases. For instance, the i-LIDS is the only dataset that contains pedestrians with luggage, thus the neurons that capture luggage features will be useless to recognize people from the other data sets.

2.2.2 Siamese models

As pointed out, the Siamese network models have been widely employed in person re-identification task. Siamese neural network is a type of neural network architectures which contains two or more identical sub-networks; this means these sub-networks have the same network architecture with the same parameters and weights (*aka* shared weight parameters, indicated by w between the sub-networks). A Siamese network can be typically employed as pairwise: with two sub-networks, or triplet: with three sub-networks. The output of Siamese model leads to be a similarity score at the top of the network. An objective function is used to train the network models, which makes the distance between the matched pairs less than the mismatched pairs in the learning feature space. In

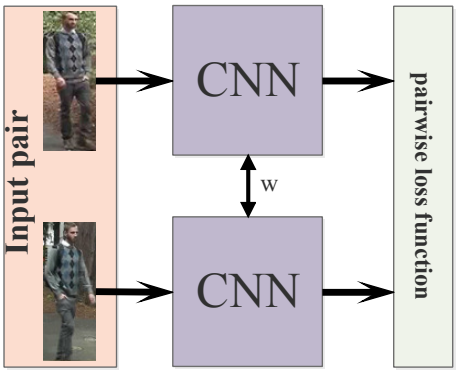

Figure 2.1: An example of a standard convolutional Siamese network based on input pair of images.

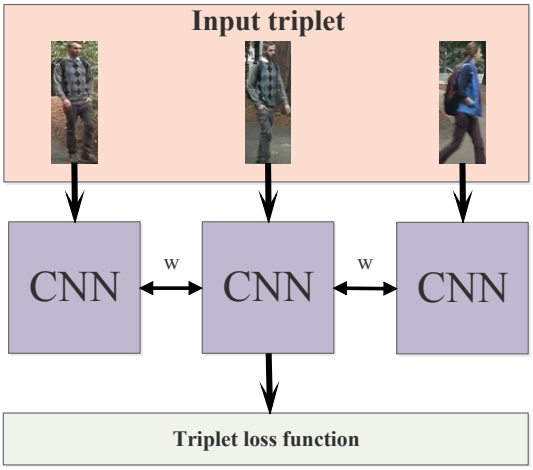

Figure 2.2: An example of a standard convolutional Siamese network based on input triplet of images.

order to output of a Siamese model, a softmax layer is employed at the top of the network on both distance outputs. Figures 2.1 and 2.2 present an example of Siamese models for pairwise and triplet comparison, respectively.

Pair-based models

In [71], the Siamese pair-based model takes two images as the input of the two sub-networks which are locally connected to the first convolutional layer. They employed a linear SVM at the top the network instead of using the Softmax activation function in order to measure similarity of input images pair as the output of the network. In [70], a Siamese neural network has been constructed to learn pairwise similarity. Each input image of pair first partitioned into three overlapping horizontal parts. The part pairs are matched through three independent Siamese networks, and finally are fused at the score level. Li et al. [44] proposed a deep filter pairing neural network to encode photo-metric

transformation across camera views. A patch matching layer is added to their network to multiply the convolution feature maps of pair images in different horizontal stripes. Later, Ahmed et al. [1] improved the pair-based Siamese model in which the network takes pair of images as the input, and outputs the probability of whether two images in the pair are of the same person or different people. The model begins with two layers of convolution by passing input pair of images. The generated feature maps are passed through a max-pooling kernel to the another convolution layer and followed another max-pooling layer to decrease the size of feature map. Then a cross-input neighborhood layer computes the differences of the features in neighboring locations of the other image.

Liu et al. [47] utilized a deep learning model to integrate a soft attention based model in a Siamese network. This model focus on the important local parts input images under pair-based Siamese model. Chen et al. [8] proposed a deep ranking framework to jointly learn representation and similarities for comparing pair of images. They aim to learn a deep CNN that assigns a higher similarity score to the positive pair than any negative pairs in each ranking unit by utilizing the logistic loss function. They first stitched the pair of images of persons horizontally to form an image which used as the input of the network, and then, the network returns a similarity score as its output. Franco et al. [25] proposed a coarse-to-fine approach to achieve a generic-to-specific knowledge through a transfer learning. The approach is followed by three steps: first a hybrid network is train to recognize a person, then another hybrid network employed to discriminate the gender of the person, and finally the output of two networks are passed through the coarse-to-fine transfer learning method to a pairwise Siamese network to accomplish the final person re-identification in order to measure the similarity between those two features. Later, the same authors proposed a novel type of features based on convolutional covariance descriptor (CCF) in [26]. They intend to obtain a set of local covariance matrices over the feature maps extracted by the hybrid network under the strategy of above-mentioned framework.

Wang et al. [67] proposed to employ a metric learning method to learn spatio-temporal features under pairwise Siamese model. The network takes a pair of images in order to obtain CNN features, and outputs whether two images reports a same person or different person by employing the quadratic discriminant analysis method. In [61], a Siamese network takes a CNN learning feature pair, and outputs the similarity value between them by applying the cosine/Euclidean distance function. A CNN framework employed to obtain deep features of each input image pair, and then, each image is split into three overlapping color patches. The deep network built in three different branches and each branch takes a single patch as its input. Finally, the three branches are concluded by a fully-connected layer.

Triplet-based models

Each triplet unit contains three images with pair of images from the same person and one from a different person. Cheng et al. [9] proposed a triplet loss function in which the network takes the triplet images as input. The network enables jointly learning of the global full-body and local body-parts features from a given person's image, and the fusion of these two types of features as the output of the network. The CNN model begins with a convolution layer, and afterward, it is divided into four equal parts, and each part forms the first layer of an independent body-part channel that aims to learn features from the respective body part. The four body-part channels together with the full-body channel constitute five independent channels that are trained separately from each other. At the top of the network, the outputs from five separate channels are concatenated into a single vector which is passed through a final fully-connected layer. Su et al. [63] proposed a semi-supervised three-stage learning in which the network first trained on an independent data set to predict the attributes, and then the attributes are trained the triplet loss on data sets with individuals labels.

2.2.3 Loss function

In most of the statistical areas such as machine learning, computational neuro-science, etc., a loss function (*aka cost function*), is a function that aims to map intuitively some values into a one single real number; this typically represents a cost which associated to those values. The techniques like Neural Networks (NNs) are in the same way to optimally minimize that loss function. When, a loss function is used for a Siamese model, it depends on the type of model which is going to be chosen (i.e. pairwise or triplet model). At the following, I discuss some of loss functions which commonly used on pair- and triplet-based models, particularly applied on person re-identification task.

Pairwise loss function

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be a set of person images and corresponding label for each, respectively, and to distinguish from the positive and negative pairs

$$I_s(x_i, x_j) = \begin{cases} positive & if\, y_i = y_j, \\ negative & if\, y_i = y_j \end{cases} \tag{2.1}$$

Pairwise hinge loss: the features from the positive pairs are geometrically close in the Euclidean distance, while the negative pairs not below a certain margin of the Manhattan distance to each other.

$$I(x_1, x_2, y) = \begin{cases} \|x_1 - x_2\| & if\, y = 1 \\ max(0, m - \|x_1 - x_2\|) & if\, y = -1 \end{cases} \tag{2.2}$$

Cosine similarity loss: this similarity loss function maximizing the cosine value for positive pairs to reduce the angle between them, and at the same time, minimizing the cosine value for the negative pairs when the value is less than margin (denoted by m).

$$I(x_1, x_2, y) = \begin{cases} max(0, cos(x_1, x_2) - m) & if\ y = 1 \\ 1 - cos(x_1, x_2) & if\ y = -1 \end{cases} \tag{2.3}$$

The total loss of two above-mentioned pairwise loss functions is computed as:

$$L(X_1, X_2, Y) = -\frac{1}{n}\sum_{i=1}^{n} I(x_i^1, x_i^2, y_i) \tag{2.4}$$

Triplet loss function

This loss function typically creates a margin between distance metric of positive pair and distance metric of negative pair. At the following, we discuss a few of common loss functions employed at existing deep learning works for person re-identification task. The triplet loss function is used to train the network models, which makes the distance between the matched pairs less than the mismatched pairs in the learning feature space. Let $O = \{(I_i, I_i^+, I_i^-)\}_{i=1}^N$ be a set of triplet images, in which I_i and I_i^+ are referred to images of the same person, and I_i and I_i^- present the different persons.

Typically, Euclidean distance is common to be used as the distance metric of this function. The loss function under *L2* distance metric has been employed in some of the triplet-based models such as [17, 47, 9, 63], and is denoted as $d(W, O_i)$; where $W = W_i$ is the network parameters, and $F_w(I)$ represents the network output of image I, in which the difference in the distance is computed between the matched pair and the mismatched pair of a single triplet unit O_i:

$$d(W, O_i) = \|F_w(I_i) - F_w(I_i^+)\|^2 - \|F_w(I_i) - F_w(I_i^-)\|^2 \tag{2.5}$$

and relatively the loss function is computed as:

$$f_w(O_i) = \sum_{O_i} max\{d(W, O_i), C\} \tag{2.6}$$

where C is a constant margin parameter. The total loss function over all the triplets can be calculated as:

$$L(I_i, I_i^+, I_i^-) = \sum_{i=1}^{N} loss(f_w(O_i), f_w(O_i^+), f_w(O_i^-)) \tag{2.7}$$

An improved triplet loss function employed in [8] as follows:

$$L(I_i, I_i^+, I_i^-, w) = \frac{1}{N}\sum (max\{d^n(I_i, I_i^+, I_i^-, w), \delta_1\} + \beta max\{d^p(I_i, I_i^+, I_i^-), \delta_2\}), \tag{2.8}$$

where N is the number of triplet training examples, β is a weight to balance the inter-class and intra-class constraints. In this implementation, the distance function $d(.,.)$ is defined as the L2-norm distance as explained above.

The hinge loss function aims to minimize the squared hinge loss of the linear SVM which is equivalent in order to finding the max margin according to the true person match and false person match over training step. This loss function is a convex approximation in range of 0-1 ranking error loss, which approximate the model's violation of the ranking order specified for a triplet unit as follows,

$$L(I_i, I_i^+, I_i^-) = max(0, C + D(I_i, I_i^+) - D(I_i, I_i^-)) \tag{2.9}$$

where C is a margin parameter which regularizes the margin between the distance of the two image pairs: (I_i, I_i^+) and (I_i, I_i^-), and D is the euclidean distance between the two euclidean points.

2.3 Processing time in person re-identification

To our knowledge, the issue of processing time has been explicitly addressed so far in the context of person re-identification only in [18, 57, 36]. To tackle person re-identification in a real-time application, the issue of processing time is one of the critical problem to be faced. However, the phase of generating the descriptors can be constructed off-line, but computing the matching score between a query image and the images in the gallery set can be reasonably high, even the similarity measurement is fast.

The authors in [18] only proposed a solution as a multi-stage framework in terms of the efficiency of the processing time: the first stage selects a subset of templates using a descriptor which is built upon a bag-of-words feature representation and an indexing scheme based on inverted lists, and requires a low processing time for computing matching scores; the second stage ranks only the selected templates using a different, more complex descriptor based on mean Riemann covariance. In [18] only two stages are considered, and only a subset of templates is ranked by the whole system, possibly losing the correct identity. Moreover, a different, specific descriptor is used in each stage, whereas our approach can be applied to any descriptor, and uses different versions of the *same* descriptor at each stage.

In [57], they proposed a dissimilarity-based approach to design descriptors made up of bags of local features, possibly extracted from different body parts. It consists in finding a set of M representative local features (called prototypes) from all individuals of the template gallery, and in representing each template and probe image as a vector of M dissimilarity values between the corresponding

Table 2.1: Summary of benchmark person Re-ID datasets.

Dataset	Multiple images	Multiple camera	Illumination variations	Pose variations	Occlusions	Scale variations
VIPeR		✓	✓	✓	✓	
i-LIDS	✓	✓	✓	✓	✓	✓
ETHZ	✓		✓		✓	✓
CUHK01	✓	✓	✓	✓	✓	
CUHK02	✓	✓	✓	✓	✓	
CUHK03	✓	✓	✓	✓	✓	
Market-1501	✓	✓	✓	✓	✓	✓
PRID		✓	✓	✓	✓	
CAVIAR	✓	✓	✓	✓	✓	✓

bag of local features and the templates. This allows the matching score to be computed as a distance between feature vectors, rather than using a more complex similarity measure between bags of local features. Contrary to the multi-stage approach proposed in this paper, the one of [57] can be applied to descriptors made up of bags of local features. The method of [36] reduces processing time in the specific multi-shot setting (when several images per individual are available), and for specific descriptors based on local feature matching, e.g., interest points. It first filters out irrelevant interest points, then it builds a sparse representation of the remaining ones.

2.4 Data sets in person re-identification

To evaluate the performance of a person re-identification method, some factors must be taken into account to reach a reliable recognition rate on this task; it makes this task challenging due to difficulty of the available bench-mark data sets such as occlusion (e.g. this is obvious on i-LIDS data set), and illumination variation. On the other hand, background and foreground segmentation in order to distinguish person's body in challenging, while some of the data sets perfectly provide this segmentation to subtract the person's body (e.g. VIPeR, ETHZ, and CAVIAR datasets). There are several available data sets that have been employed to measure the performance of re-identification methods[1]. Among these datasets, VIPeR, CUHK01, and CUHK03 are mostly interested by researchers of this field of research to evaluate the deep learning techniques. However, VIPeR is most commonly used for re-identification evaluations due to it is challenging on individuals images. At the following, I give a brief description on a few of bench-mark data sets on person re-identification task, and additionally table 2.1 provides a summary of them.

VIPeR [28] is a challenging data set for person re-identification; it is made up of two images of 632 individuals from two camera views, with pose and illumination changes. This is one of the most

[1]For comprehensive information about the available person re-identification data sets, check out the following link: http://robustsystems.coe.neu.edu/sites/robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html

Figure 2.3: Example of images from VIPeR data set [5]. Images on the same column represent the same person.

Figure 2.4: Example of images from i-LIDS data set [5]. Images on the same column represent the same person.

challenging data sets yet for person re-identification task. The images are cropped and scaled to be 128×48 pixels. Fig. 2.3 shows some example images from this data set.

i-LIDS [7] was acquired in crowded public spaces which contains 476 images of 119 pedestrians taken at an airport hall from non-overlapping cameras, with pose and lightning variations and strong occlusions. A minimum of 2 images and on an average there are 4 images of each pedestrian. Fig. 2.4 shows some example images from this data set.

ETHZ [20] contains three video sequences of a crowded street from two moving cameras; images exhibit considerable illumination changes, scale variations, and occlusions. The images are of different sizes. The data set provides three sequences of multiple images of an individual from each sequence. Sequences 1, 2 and 3 have 83, 35, and 28 pedestrians respectively. Fig. 2.5 shows some example images from this data set.

As a recent well-known dataset provided by Chinese University of Hong Kong (CUHK), which particularly gathered persons images for person re-identification task, and includes three different

Figure 2.5: Example of images from ETHZ data set. The first, second, and third row present the images from sequences 1,2, and 3, receptively [5].

partitions with specific set up for each; *CUHK01* [43] includes 1,942 images of 971 pedestrians; it has only two images captured in two disjoint camera views, and camera B mainly includes images of the frontal view and the back view, and camera A has more variations of viewpoints and poses. Fig. 2.6 presents some samples of this data set; *CUHK02* [42] contains 1,816 individuals constructed by five pairs of camera views (P1-P5 with ten camera views). Each pair includes 971, 306, 107,193 and 239 individuals respectively. Each individual has two images in each camera view. This dataset is employed to evaluate the performance when camera views in test are different than those in training. Fig. 2.7 presents some samples of this dataset; and *CUHK03* [44] includes 13,164 images of 1,360 pedestrians. This data set has been captured with six surveillance cameras. Each identity is observed by two disjoint camera views and has an average of 4.8 images in each view; all manually cropped pedestrian images exhibit illumination changes, misalignment, occlusions and body part missing. Fig. 2.8 presents some samples of this dataset.

The **Market-1501** [73] is the largest person re-identification data set up to date, and contains 32,643 fully annotated boxes of 1501 pedestrians. Each perseon is captured by maximum six cameras and boxes of person are cropped by employed a state- of-the-art detector, the Deformable Part Model (DPM) [22]. Fig. 2.9 shows some example images from this data set.

The **PRID** [30] data set is specially designed for person ReID in single shot. It contains two image

Figure 2.6: Example of images from CUHK01. Images on the same column represent the same person.

Figure 2.7: Example of images from CUHK02 provided in [42], which has five pairs of camera views denoted with P1-P5, with two images per person are shown for each of pairs.

sets containing 385 and 749 persons captured by camera A and camera B, respectively. These two data sets share 200 persons in common.

CAVIAR [10] contains 72 persons and two views in which 50 of persons appear in both views while 22 persons appear only in one view. Each person has 5 images per view, with different appearance variations due to resolution changes, light conditions, occlusions, and different poses. Fig. 2.10 shows some example images from this data set.

Figure 2.8: Example of images from CUHK03. Images on the same column represent the same person.

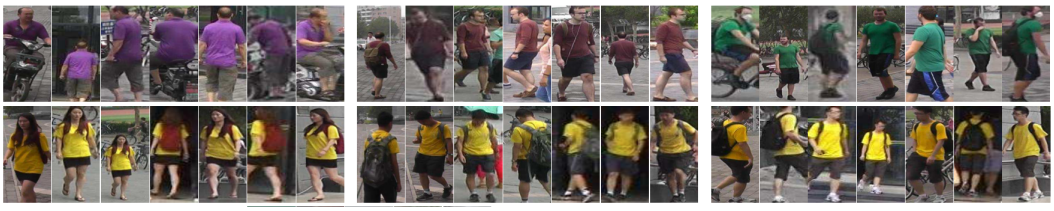

Figure 2.9: Example of images from MARKET-1501 data set provided in [73].

Figure 2.10: Example of images from CAVIAR data set [5]. Images on the same column represent the same person.

Chapter 3

Multi-stage person re-identification systems

As pointed out at chapter 1, existing multi-stage approaches to classification problems, aimed at trading classification accuracy for the processing time, cannot be directly applied to person re-identification, which involves a ranking problem. As the main contribution of this part of my thesis work, a specific formulation of multi-stage ranking problems is proposed and developed in order to trade-off between ranking accuracy and processing time; in which only focuses on trading ranking accuracy for processing time as spend for matching phase of person re-identification problem, for any given descriptor and similarity measure. In particular, I first develop an analytical model of processing time and discuss the behaviour of the corresponding ranking accuracy measured using the CMC curve. Based on these results I also propose a practical design criteria. This work extends my preliminary work in [39] by considering in the analytical model of the behaviour of multi-stage re-identification systems, in the design criteria, and in a wider empirical investigation. The former approach focused on practical application scenarios characterized by a very large template gallery to be ranked in response to a query by a human operator, and/or by a similarity measure exhibiting a high processing time. As the main drawback of this approach, in practice, it could also be difficult to accurately estimate the corresponding optimal values of the number of templates to be ranked by each stage (but the first one), as they depend on the size of template gallery. I, further, investigate on a design criteria which focused instead on strict requirements characterized by the maximum allowed processing time.

At the following of this chapter, I first discuss on some proposed multi-stage systems in classification problems. Next, I summarize existing multi-stage methods which specifically proposed for person re-identification task. I then describe my proposed multi-stage ranking approach on person

re-identification problem by formulating a possible criterion to attain a proper trade-off between ranking accuracy and processing time. At the end, I present some experimental evaluations of multi-stage ranking approach; carried out by using four state-of-the-art descriptors on three benchmark data sets.

3.1 Overview on multi-stage approaches

Although, the idea of multi-stage system is not that novel and has been employed in many applications, it still keeps its novelty on person re-identification systems in which *ranking problem* always takes into account. At this section, a brief taxonomy of multi-stage system is presented for its applications in pattern recognition, and then more specifically discussion of application of multi-stage system in person re-identification task.

3.1.1 Multi-stage classification approaches

The multi-stage approach is used since a long time in pattern classification systems. For instance, in [55] a cascade of classifiers was proposed to attain a trade-off between classification accuracy and the cost of feature acquisition, e.g., for medical diagnostics applications: each classifier uses features that are more discriminant, but also more costly [55] than previous classifiers. The goal is to assign an input instance (e.g., a medical image) to one of the classes (e.g., the outcome of a diagnosis) with a predefined level of confidence, using features (e.g., medical exams) with the lowest possible cost; if a classifier but the last one does not reach the desired confidence level, it *rejects* the input instance (i.e., withholds making a decision), and sends it to the next stage. This approach has later been exploited to attain a trade-off between classification accuracy and processing time, e.g., in handwritten digit classification [35, 62, 64]. A similar approach is used in the well-known algorithm of [66] for designing fast object detectors: it consists in detecting and discarding background regions of the input image as quickly as possible, using classifiers based on features fast to compute; this allows focusing the attention on regions more likely to contain the object of interest, using classifiers based on more discriminant features that also require a higher processing time.

3.1.2 Multi-stage re-identification approaches

Multi-stage re-identification systems have already been proposed by some authors. Their aim is however to improve ranking accuracy, without taking into account processing time [30, 52, 46, 68, 33].

In [30] the first stage uses returns the operator the 50 top-ranked templates; if the probe identity is not among them, a classifier is trained to discriminate the probe image from other identities, and is used to re-rank the remaining templates. In [52] person re-identification is addressed as a content-based image retrieval task with relevance feedback, for settings where several instances of a probe can be present in the template gallery; accordingly, the aim is to increase recall. In each stage (i.e., iteration of relevance feedback) only the top-ranked templates are shown to the operator, then his feedback is exploited to adapt the similarity measure for the probe at hand, and the remaining templates are re-ranked. A similar multi-stage strategy was proposed in [46] for reducing the operator's effort in analyzing the template images: in each stage only the top-ranked templates are presented to the operator, who is asked to select a "strong negative" (i.e., a different individual whose appearance is most dissimilar to the probe), and optionally a few "weak negatives"; a post-rank function is then learnt based on this feedback and on the probe image, and the remaining templates are re-ranked in the next stage. A similar, two-stage approach was proposed in [68]: the operator is asked to label some pairs of locally similar and dissimilar horizontal image regions in the top-ranked templates, and this feedback is exploited to re-rank all templates. Another two-stage approach was proposed in [33], to improve the ranking provided by a given first-stage descriptor: a small subset of the top-ranked templates is re-ranked by the second stage, by a different descriptor that uses a manifold-based method with three specific low-level features.

Accordingly, as the main contribution of this work, in this chapter we develop a specific formulation of multi-stage ranking problems focused on trading ranking accuracy for processing time in person re-identification systems, for any given descriptor and similarity measure. In particular, we first develop an analytical model of processing time and discuss the behaviour of the corresponding ranking accuracy measured using the CMC curve. Then I define practical design criteria for multi-stage person re-identification systems, based on my analytical model of their behaviour.

At this work I consider application scenarios characterized by strict requirements on the processing time for obtaining the ranked list of templates, e.g., due to real-time constraints. In particular, I consider requirements expressed by the constraint $t \leq t_{max}$, where t_{max} is an application-specific value. Many existing appearance descriptors attain a high recognition rate at the expense of a high complexity, which results in a relatively high value of t, e.g., [21, 58, 50, 45]. Moreover, even if t is relatively low, when the gallery set size is very large an even lower t_{max} value may be required. Focusing on the case when a *given* descriptor D exhibits a satisfactory ranking accuracy, but does not meet the constraint $t \leq t_{max}$, in the next section I propose a multi-stage ranking approach capable of trading a lower ranking accuracy for a lower processing time.

3.2 A multi-stage ranking approach for person re-identification

Let me first discuss the case of a two-stage ranking system. Consider a given descriptor, that I denote as D_2, and assume that it exhibits a satisfactory ranking accuracy (CMC curve) but a too high processing time, $t_2 > t_{max}$, as explained above. My approach is based on modifying D_2, by changing its parameters, into a descriptor D_1 that exhibits a lower processing time $t_1 < t_{max}$. Usually this can be attained only at the expense of a lower accuracy, i.e., the CMC curve of D_1 (denoted as CMC_1) lies below that of D_2 (CMC_2). If CMC_1 is not satisfactory for the application at hand, D_1 and D_2 can be combined into a two-stage system to meet the constraint on processing time, attaining at the same time a CMC curve better than CMC_1. To this aim, for a given probe, first all n templates are ranked using D_1, then the n_2 top-ranked ones are re-ranked using D_2, for a given n_2, with $1 < n_2 < n$. The resulting average processing time per probe, t_{1-2}, is given by:

$$t_{1-2} = \frac{1}{n} t_{D_1} + t_1 + \frac{n_2}{n} t_2 \,, \tag{3.1}$$

where also the time t_{D_1} for computing the descriptor D_1 of the probe is taken into account (the same descriptor can be computed offline for templates, and is therefore not considered). Note that the impact of such an overhead time reduces as the overall number of templates to be ranked increases. From Eq. (3.1), the constraint $t_{1-2} \leq t_{max}$ translates into:

$$n_2 \leq n \frac{(t_{max} - t_1)}{t_2} - \frac{t_{D_1}}{t_2} \,. \tag{3.2}$$

Consider now the resulting CMC curve, denoted by CMC_{1-2}. To make an analytical derivation of its behaviour possible, at least to some extent, I disregard the general case when CMC_1 and CMC_2 cross in one or more points, and consider only the case when $CMC_1(r) < CMC_2(r)$ for ranks $r \leq r^*$, and $CMC_1(r) = CMC_2(r)$ for $r > r^*$, for a given rank $r^* \leq n$, as in the example of Fig. 3.3. In other words, when D_2 gives a rank between 1 and r^* to the template of the correct identity, the rank given by D_1 is on average lower; when D_2 gives a rank between r^* and n, instead, the rank given by D_1 is on average the same (see the example in Fig. 3.1). In the limit cases of $n_2 = 1$ and $n_2 = n$, it is easy to see that $CMC_{1-2} = CMC_1$ and $CMC_{1-2} = CMC_2$, respectively. For $1 < n_2 < n$, the above assumption implies that CMC_{1-2} lies between CMC_1 and CMC_2, and approaches CMC_2 as n_2 increases. This can be proven as follows. First, $CMC_{1-2}(r) = CMC_1(r)$ for all $r \geq n_2$, since for any $r \geq n_2$ the correct identity is among the r top ranks of the two-stage system, if and only if it is among the r top ranks of D_1. Second, since the n_2 top-ranked templates by D_1 are re-ranked by the more accurate D_2, it

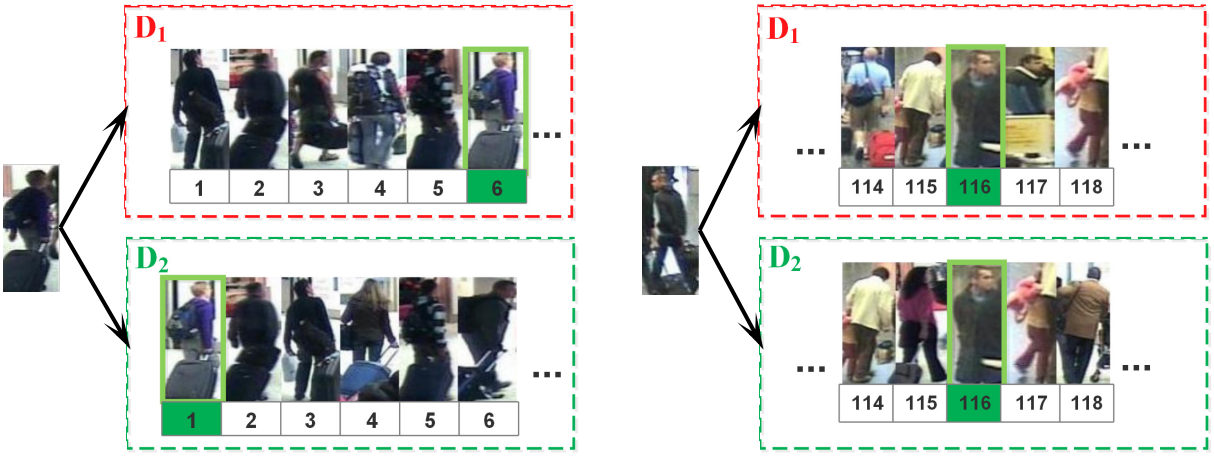

Figure 3.1: Two examples of the ranked list of templates produced by a descriptor D_2 and by a less accurate version of it, D_1, for a given probe (the correct identity is marked in green). Left: the correct identity is in the top ranks, and is ranked *higher* by D_2. Right: the correct identity has a low rank, and is ranked *identically* by both descriptors.

follows that $CMC_{1-2}(r) \geq CMC_1(r)$ for $r < n_2$. An example of this behaviour is reported in Fig. 3.3 for two different values of n_2.

To sum up, for two-stage systems a trade-off between processing time and ranking accuracy can be attained by values of n_2 that satisfy constraint (3.2): the higher n_2, the higher the resulting processing time and ranking accuracy.

The above results can be generalized to multi-stage systems with $N > 2$, using the original descriptor in the last stage as D_N, and different versions of D in the previous stages as $D_1,\ldots,D_{N-1}$, characterized by increasing ranking accuracy and increasing processing time, $t_1 < t_2 < \ldots < t_{D_N}$, with $t_1 < t_{max}$ (see Fig. 3.2). Denoting by n_i the number of matching scores computed by the i-th stage, under the constraint:

$$n_1 = n > n_2 > \ldots > n_N > 1\,, \tag{3.3}$$

the corresponding average processing time t_{1-N} is:

$$t_{1-N} = \frac{1}{n}\sum_{i=1}^{N-1} t_{D_i} + t_1 + \sum_{i=2}^{N}\frac{n_i}{n}t_i\,. \tag{3.4}$$

Accordingly, the constraint $t_{1-N} \leq t_{max}$ can be rewritten as:

$$\sum_{i=2}^{N} n_i t_i \leq n(t_{max} - t_1) - \sum_{i=1}^{N-1} t_{D_i}\,. \tag{3.5}$$

Note that constraint (3.2) is a particular case of (3.5) for $N = 2$.

Assuming that the CMC curves of any pair of adjacent stages, CMC_i and CMC_{i+1}, exhibit the same behaviour considered above (see Fig. 3.3), by the same arguments above it follows that the CMC curve of the multi-stage system, CMC_{1-N}, lies between CMC_1 and CMC_N. In particular, in the limit

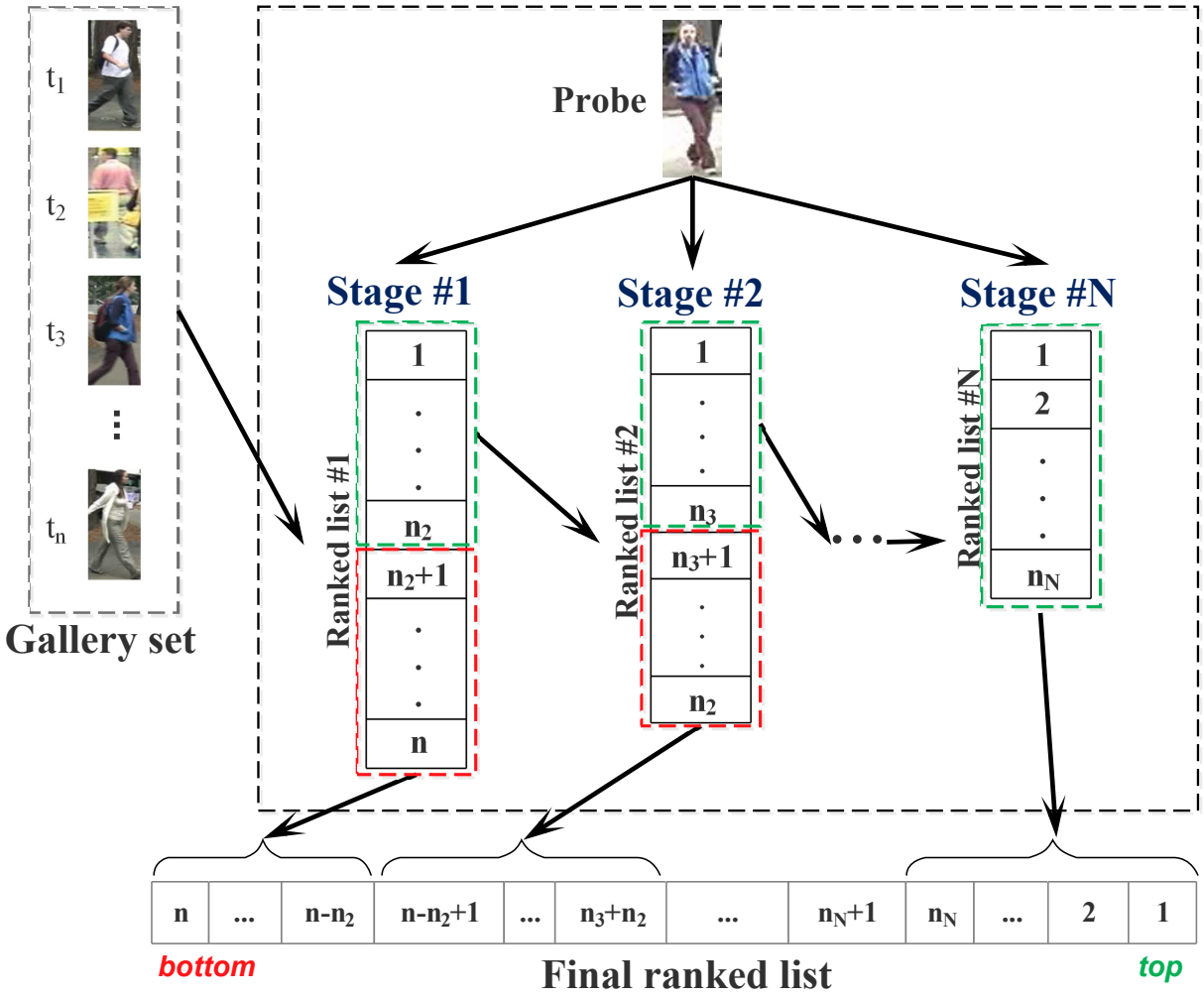

Figure 3.2: Scheme of the proposed multi-stage ranking approach.

cases when $n_i = 1$, and $n_i = n$ for every $i > 1$, we obtain $CMC_{1-N} = CMC_1$, and $CMC_{1-N} = CMC_N$, respectively. Moreover, $CMC_{1-N}(r) = CMC_1(r)$ for $r \geq n_2$. In general, for increasing values of $n_2, \ldots, n_N$, CMC_{1-N} gets closer to CMC_N.

Accordingly, for a generic multi-stage system a trade-off between processing time and accuracy can be attained when the n_i's satisfy constraints (3.3) and (3.5); the higher $n_2, \ldots, n_N$, the higher the resulting processing time and ranking accuracy.

3.3 Design criteria

Designing a multi-stage re-identification system according to the above approach requires to choose the number N of stages, the descriptors $D_1, \ldots, D_{N-1}$, and the number of templates $n_2 > \ldots > n_N$ to be re-ranked at each stage, under constraints (3.3) and (3.5). The best solution, among the ones that

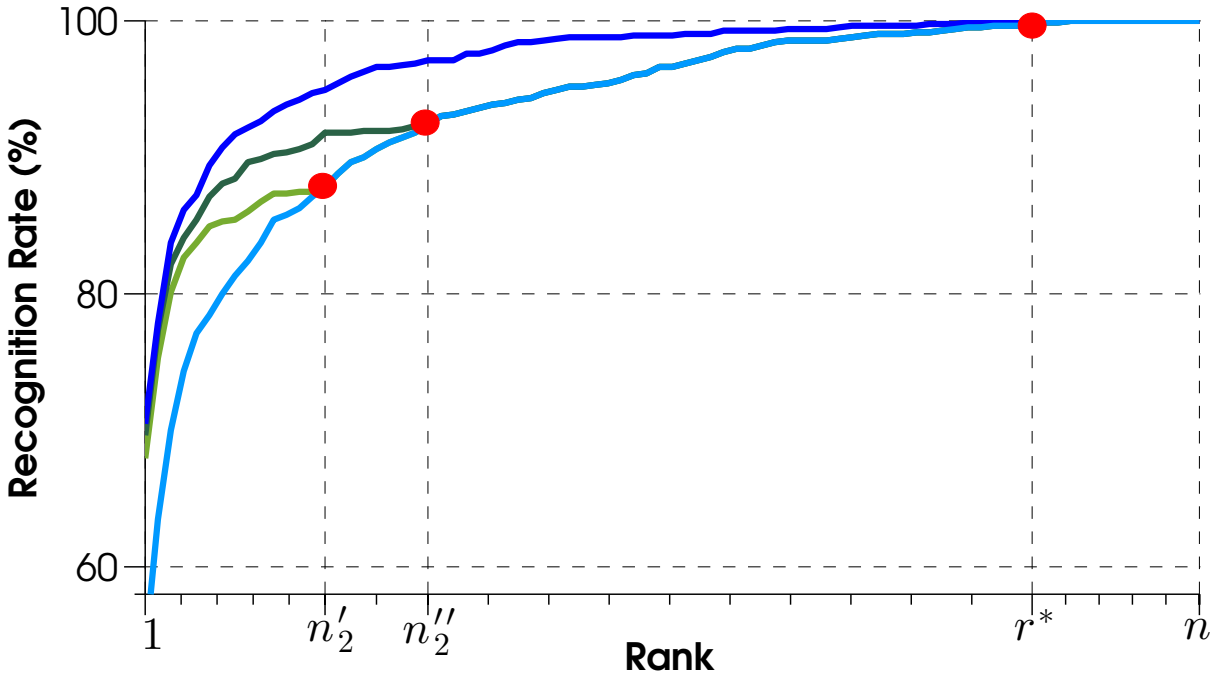

Figure 3.3: Example of CMC curves of two-stage systems. Light blue: first-stage; dark blue: second-stage; r^* is the rank from which their CMC curves become identical; light and dark green: two-stage systems corresponding to different values of n_2.

satisfy such constraints, is the one that maximizes ranking accuracy. However it cannot be analytically found, and finding it empirically by evaluating all possible choices is clearly impractical, since the three choices above are interrelated and many possible solutions may even exist. In the following I discuss each choice separately, and suggest practical, though suboptimal, design criteria.

Descriptors. Consider first the problem of developing different versions $D_{N-1}, \ldots, D_1$ of a given descriptor D, exhibiting a decreasing ranking accuracy and a decreasing processing time. This can be attained by suitably modifying the parameters of D. However existing descriptors can be very complex and contain several parameters. Moreover, only an empirical evaluation is usually possible of the impact of any parameter on ranking accuracy; for instance, the relative behaviour of the CMC curves of any two descriptors depends on the data at hand: see, e.g., the CMC curves of the original and of the first-stage SDALF descriptor on the VIPeR and ETHZ1 data sets, in Fig. 3.4. To define a practical design criterion I propose to subdivide descriptors into two main categories: fixed-size feature vectors (e.g., [50, 45]), and descriptor with variable size (e.g., [21]). For fixed-size feature vectors, an unsupervised feature reduction technique like PCA can be used. The suitability of PCA to person re-identification tasks is witnessed to its use in the pre-processing step of gBiCov [50]. For descriptors with variable size I suggest to modify the parameter that has the highest impact on processing time; for instance, in SDALF descriptor [21] such a parameter is the number of "blobs" of its MSCR com-

ponent (see Sect. 3.4.1). Once a *single* parameter has been chosen (either the feature set size for the former category of descriptor, or a descriptor-specific parameter for the latter category), its value for each stage (but the last one) can be set according to the corresponding processing time, which has to be empirically evaluated. The only constraint is that the left-hand side of inequality (3.5) is positive, which amounts to:

$$t_1 < t_{max} - \frac{1}{n} \sum_{i=1}^{N-1} t_{\mathrm{D}_i} \,. \tag{3.6}$$

As a simple guideline, one should set t_1 to be no more than half the above upper bound.

Number of templates to be re-ranked at each stage. Assuming that N and D_{N-1}, ..., D_1 have already been chosen, the choice of $n_2, \ldots, n_N$ can be discussed separately for $N = 2$ and $N > 2$. For two-stage systems, the single value of n_2 has to be chosen under constraint (3.2). In this case the best trade-off between processing time and ranking accuracy can be identified a priori: it is attained when the second stage re-ranks the highest possible number of templates, which leads to:

$$n_2 = \left\lfloor n \frac{(t_{max} - t_1)}{t_2} - \frac{t_{\mathrm{D}_1}}{t_2} \right\rfloor \,. \tag{3.7}$$

In the case $N > 2$, constraints (3.3) and (3.5) define a convex polyhedron in the $N-1$-dimensional space, and the feasible solutions are all the points $\mathbf{n} = (n_2, \ldots, n_N)$ with integer coordinates belonging to such a polyhedron. However, among these solutions it is not possible to identify a priori the one that maximizes ranking accuracy. One can only discard the *dominated* solutions: if a solution $\mathbf{n}' = (n'_2, \ldots, n'_N)$ is dominated by a different solution $\mathbf{n}'' = (n''_2, \ldots, n''_N)$, i.e., $n'_2 \leq n''_2, \ldots, n'_N \leq n''_N$, then $\mathbf{n}'$ can be discarded, since each of its stages (but the first one) re-ranks a lower or identical number of templates than the corresponding stage of $\mathbf{n}''$, and consequently its ranking accuracy will be lower. Instead, for any pair of non-dominated solutions $\mathbf{n}'$ and $\mathbf{n}''$, if $n'_i < n''_i$ for some i, then some j exists such that $n'_j > n''_j$; this means that their relative ranking accuracy can be evaluated only empirically, which is impractical if the number of non-dominated solutions is high.

To avoid such problems, I consider a simpler, though potentially suboptimal criterion for multi-stage systems with $N > 2$: I consider values of $n_2, \ldots, n_N$ such that, beside satisfying constraints (3.3) and (3.5), the number of templates between two consecutive stages is reduced by a same amount $\alpha < 1$, i.e.:

$$n_i = \lfloor \alpha n_{i-1} \rfloor, \; i = 2, \ldots, N \,. \tag{3.8}$$

It is now easy to see that ranking accuracy is maximized by choosing the maximum value of α that satisfies constraints (3.3) and (3.5), which can be found by a simple line search.

Number of stages. Taking into account the design criteria suggested above, I suggest to limit the choice of the number of stages to two or three, to avoid a time-consuming empirical evaluation

of more alternatives. In practice, for a two-stage system one can set the parameter of D_1 such that $t_1 < \frac{1}{2}\left(t_{max} - \frac{1}{n} t_{D_1}\right)$ (see above); for a three-stage system one can set the parameter of D_1 and D_2 such that $t_1 < \frac{1}{2}\left[t_{max} - \frac{1}{n}\left(t_{D_1} + t_{D_2}\right)\right]$, and t_2 about twice t_1. Then the choice between a two- and a three-stage system can be made based on an empirical comparison of the corresponding ranking accuracy.

3.4 Experimental evaluation

I evaluated the proposed approach on three benchmark data sets (VIPeR, i-LIDS and ETHZ data sets), and four state-of-the-art appearance descriptors, using two- and three-stage systems.[1] Take into consideration that, I used only the first sequence "SEQ. #1" (ETHZ1) which contains the largest number of pedestrians (83), and 4,857 images in total. I also rescaled the images of i-LIDS and ETHZ1 to the same size of 128 × 48 pixels as in VIPeR, to get a similar processing time.

3.4.1 Descriptors

I used the SDALF, gBiCov, LOMO and MCM descriptors. SDALF and MCM are not fixed size descriptors: I chose ad hoc parameters to modify as described below. Although, gBiCov and LOMO are fixed-size descriptors, instead, and according to my suggested design criteria I obtained faster and less accurate versions of each of them by using PCA, I additionally obtained different version of the each of them with ad hoc parameters for the sake of comparison. Since my aim was not to fine-tune these descriptors to maximize their performance on each data set, I chose the parameter values by preliminary experiments, and used the same versions of each descriptor for all data sets.

SDALF[2] [21] subdivides body into four parts: left and right, torso and legs. Three kinds of features are extracted from each part: maximally stable color regions (MSCR), i.e., elliptical regions (blobs) exhibiting distinct color patterns (their number depends on the specific image), with a minimum size of 15 pixels; a 16 × 16 × 4-bins weighted HSV color histogram (wHSV); and recurrent high-structured patches (RHSP) that characterize texture. A specific similarity measure is defined for each feature; the matching score is computed as their linear combination. In my experiments I did not use RHSP, due to its relatively lower performance. I obtained faster and less accurate versions of SDALF by increasing the minimum MSCR blob size to 65 and to 45 for the first and second stage, respectively (which reduces the number of blobs), and by reducing the corresponding number of bins of the wHSV histogram to 3 × 3 × 2 and to 8 × 8 × 3.

[1] The source code of the experiments is available at https://github.com/bahramlavi/MultiStageRanking
[2] Source code: http://www.lorisbazzani.info/sdalf.html

gBiCov[3] [50] is based on biologically-inspired features (BIF) obtained by Gabor filters with different scales over the HSV color channels. The resulting images are subdivided into overlapping regions of 16×16 pixels; each region is represented by a covariance descriptor that encodes shape, location and color information. BIF and covariance descriptors are concatenated, and PCA is used to reduce its dimension. I obtained different versions of gBiCov by increasing the region size to 32×64 and 16×32 pixels for the first and second stage, respectively. I also obtained different versions of gBiCov by reducing its dimension to 5 for two-stage systems, and to 2 and 5 for three-stage systems.

LOMO[4] [45] extracts an $8 \times 8 \times 8$-bins HSV histogram and two scales of the Scale Invariant Local Ternary Pattern histogram (characterizing texture) from overlapping windows of 10×10 pixels; it then retains one only histogram from all windows at the same horizontal location, obtained as the maximum value among all the corresponding bins. These histograms are concatenated with the ones computed on a down-sampled image. A metric learning method is used to define the similarity measure. I obtained different versions of LOMO by increasing the window size to 20×20 and 15×15 for the first and second stage, respectively, and by decreasing the corresponding number of bins of the HSV color histogram to $3 \times 3 \times 2$ and $4 \times 4 \times 3$. Additionally, I used PCA to reduce the dimension of the LOMO descriptor to 20 for two-stage systems, and to 5 and 20 for three-stage systems.

MCM[5] [58] subdivides body into torso and legs, and extracts 80 randomly positioned image patches from each part. Each patch is represented by a $24 \times 12 \times 4$-bins HSV histogram. Artificial patches are also generated to improve robustness to illumination changes, by changing the brightness and contrast of the original patches in the RGB color channel. The similarity measure is the average k-th Hausdorff distance between the set of patches of each pair of corresponding body parts, where k was set to 10 in [58]. I obtained different versions of MCM by reducing the number of patches to 10 and to 20 for the first and second stage, respectively, and the corresponding number of bins of the HSV histogram to $3 \times 3 \times 2$ and $12 \times 6 \times 2$.

3.4.2 Experimental setup

For each descriptor D, I designed two- and three-stage systems; for the sake of simplicity I used the same version of D to implement D_1 in two-stage and D_2 in three-stage systems. As in [21], for each data set I repeated my experiments on ten different subsets of individuals, using one image of each individual as template and one as probe, and reported the average CMC curve over the ten runs. I used an Intel Core i5 2.6 GHz CPU. I considered three different values of t_{max} defined as a fraction of

[3] source code: http://vipl.ict.ac.cn/members/bpma
[4] source code: http://www.cbsr.ia.ac.cn/users/scliao/projects/lomo_xqda/
[5] source code is available upon request to the authors.

Table 3.1: Number of templates processed at each stage for each descriptor and data set, and for the different values of β.

Descriptor	Data set	Two-stage systems				Three-stage systems					
			β=0.3	β=0.4	β=0.5	β=0.3		β=0.4		β=0.5	
		n	n_2	n_2	n_2	n_2	n_3	n_2	n_3	n_2	n_3
SDALF	VIPeR	316	25	57	88	84	22	120	45	150	71
	i-LIDS	119	9	21	33	31	8	45	17	56	26
	ETHZ1	83	7	15	23	22	5	31	11	39	18
gBiCov	VIPeR	316	50	81	113	140	62	169	90	193	118
	i-LIDS	119	19	31	43	53	23	63	33	72	44
	ETHZ1	83	13	21	30	37	16	44	23	50	30
gBiCov+PCA	VIPeR	316	92	124	156	170	91	197	123	221	154
	i-LIDS	119	35	47	59	64	34	74	46	83	58
	ETHZ1	83	24	33	41	44	23	51	31	58	40
LOMO	VIPeR	316	57	107	138	149	70	178	100	203	130
	i-LIDS	119	28	42	52	56	26	67	37	76	48
	ETHZ1	83	20	28	36	39	18	46	25	53	34
LOMO+PCA	VIPeR	316	88	120	151	167	88	194	119	218	150
	i-LIDS	119	33	45	57	63	33	73	44	82	56
	ETHZ1	83	23	31	40	43	22	51	31	57	39
MCM	VIPeR	316	94	126	157	172	93	199	125	222	156
	i-LIDS	119	35	47	59	64	34	74	46	83	58
	ETHZ1	83	25	33	41	45	24	52	32	58	40

the processing time of the original descriptor used in the last stage, $t_{max} = \beta t_N$, for $\beta = 0.3, 0.4, 0.5$; as can be seen from Table 3.2, all the resulting values satisfied the constraint $t_1 < t_{max}$ (see Sect. 3.2).

3.4.3 Results

The average processing time for computing one matching score at each stage, evaluated on VIPeR, is reported in Table 3.2. Similar processing times were observed in the other data sets, due to the use of the same image size. Note that processing time of MCM cannot be compared to the one of the other descriptors, since MCM was implemented in C# and the other descriptors in Matlab. Note also that the original MCM descriptor has a much higher processing time than its versions used in the first and (for three-stage systems) second stage, with respect to the other descriptors: this is due to the use of the Hausdorff distance as similarity measure, which makes the processing time proportional to the *square* of the number of patches (see Sect. 3.4.1).

The number of templates processed at each stage, chosen according to the proposed design criterion, is reported in Table 3.1. The average CMC curves are shown in Figs. 3.4 and 3.5, respectively for two- and three-stage systems.

Note first that, since I did not fine-tune the different versions of each descriptor to each data set, in some cases the first and last stages turned out to exhibit very similar CMC curves, and therefore the CMC curve of the correponding multi-stage systems is similar to both. For instance, this is the case of SDALF and MCM on VIPER, and of gBiCov on i-LIDS, in two-stage systems (Fig. 3.4).

As another alternative guidelines for the fixed-size descriptors, as pointed out in Sec. 3.3, I obtained the fast versions of a given descriptor by using PCA method. The average CMC curves are shown in Figs. 3.6 and 3.7, respectively for two- and three-stage systems.

Table 3.2: Average processing time t_i (in msec.) for computing one matching score in the i-th stage, for each of the four descriptors. Note that the original descriptor is used in the last stage.

		SDALF	*gBiCov*	*gBiCov+PCA*	*LOMO*	*LOMO+PCA*	*MCM*
two-stage systems	t_1	2.08	0.0057	0.0003	0.0023	0.0008	0.060
three-stage systems	t_1	1.60	0.0015	0.0002	0.0017	0.0003	0.051
	t_2	2.08	0.0057	0.0003	0.0023	0.0008	0.060
last stage	t	9.44	0.0400	0.0400	0.0370	0.0370	27.400

In all the other cases the trade-off between the ranking accuracy and the processing time (given by t_{max}) of multi-stage systems clearly emerges; see, e.g., the CMC curves of SDALF on ETHZ1, both in two- and in three-stage systems. In particular, note that in the top ranks the CMC curve of these multi-stage systems is almost identical to the one of the corresponding original descriptor; it then decreases, starting from a rank that depends on the specific data set and descriptor, up to becoming identical since rank n_2 to the CMC curve of the first stage. Moreover, for a given data set and descriptor, the CMC curve of the corresponding multi-stage system worsens as t_{max} decreases, i.e., as n_2 (and, for three-stage systems, n_3) increases. As point out that this behaviour agrees with the one that has been derived analytically in Sect. 3.2, and then exploited in Sect. 3.3 to define the proposed design criterion. Accordingly, this provides evidence that my design criterion, albeit suboptimal for systems made up of more than two stages, allows one to attain an effective trade-off between processing time and ranking accuracy.

3.5 Conclusions

I proposed a multi-stage ranking approach for person re-identification, aimed at trading a lower processing time for a lower ranking accuracy for any *given* appearance descriptor. My approach is inspired by the well-known multi-stage classification architecture used in pattern recognition systems, which I adapted to ranking problems by developing an ad hoc analytical model of the trade-off be-

tween their ranking accuracy and processing time. I also suggested practical design criteria based on my analytical model, and carried out a first empirical investigation on benchmark data sets and state-of-the-art descriptors. Multi-stage re-identification systems can be useful in practical applications that involve interaction with human operators and are characterized by very large template galleries and/or complex descriptors, requiring strict constraints on processing time. They can be useful also in application scenarios when the operator cannot or does not want to scan all the ranked template images (e.g., in real-time settings): in this case, only the subset of templates ranked by the last stage can be returned to the operator. If needed, the attainable trade-off between processing time and ranking accuracy can be improved, with respect to my suggested design criteria, by fine-tuning the different system parameters discussed in Sect. 3.3, at the expense of an additional effort to empirically evaluate the different alternatives.

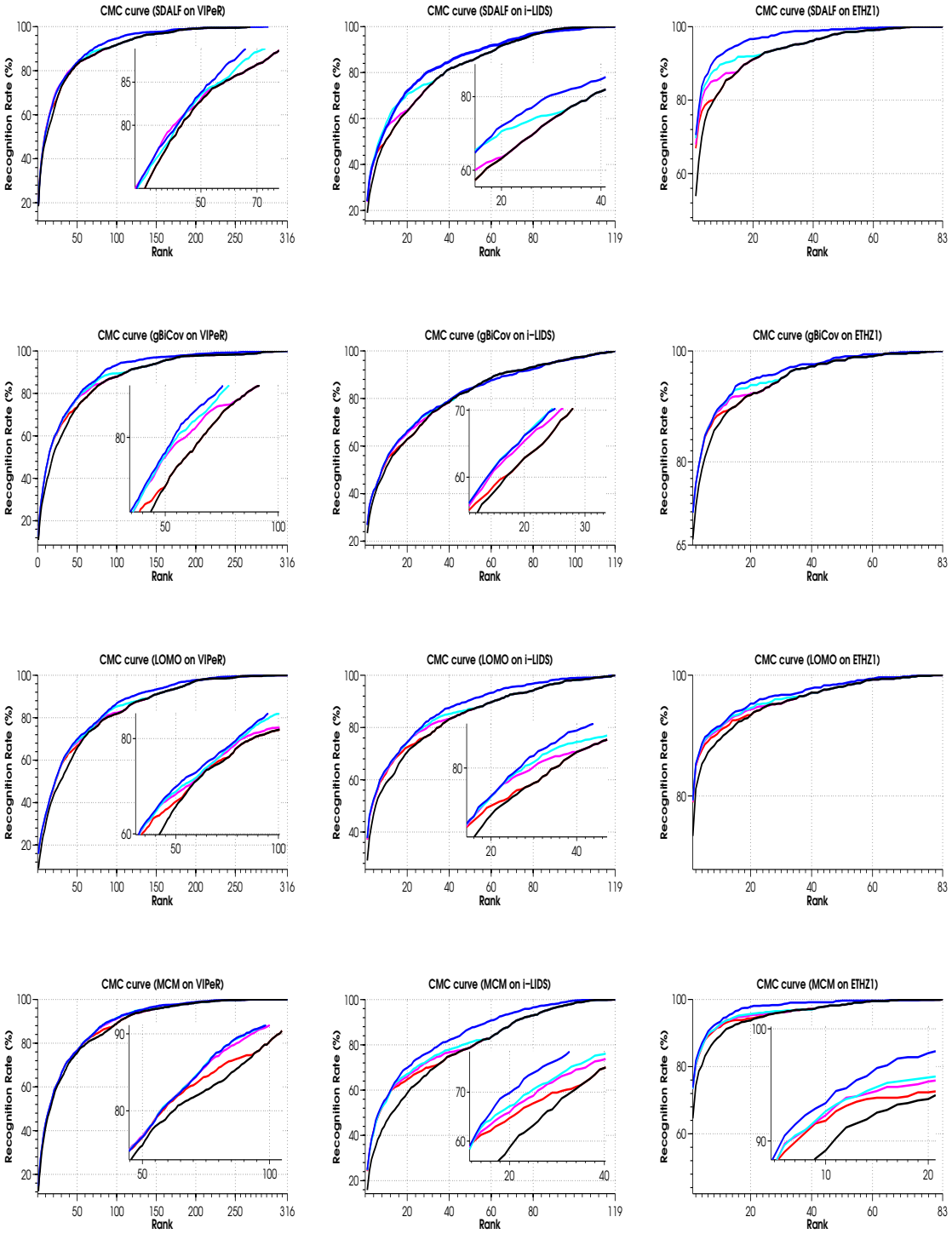

Figure 3.4: CMC curves of two-stage systems where the different versions of descriptors obtained by ad hoc parameters modification. Black: first stage; blue: second stage (original descriptor); red, pink, and cyan: two-stage systems with $\beta = 0.3, 0.4, 0.5$, respectively. Enlarged version of plots with very close CMC curves are shown for better visualization.

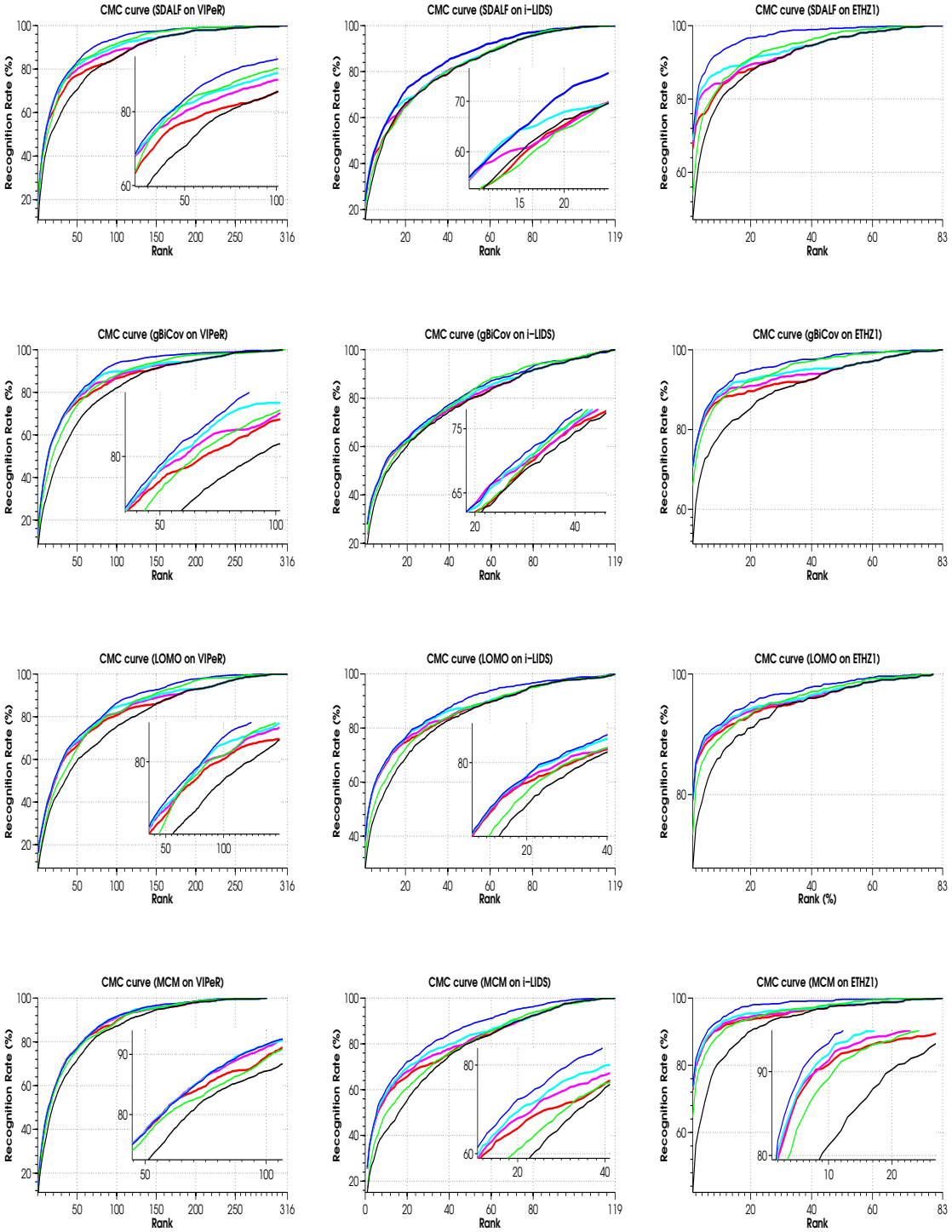

Figure 3.5: CMC curves of three-stage systems where the different versions descriptors obtained ad hoc parameters modification. Black: first stage; green: second stage; blue: third stage (original descriptor); red, pink, and cyan: three-stage systems with $\beta = 0.3, 0.4, 0.5$, respectively. Enlarged version of plots with very close CMC curves are shown for better visualization.

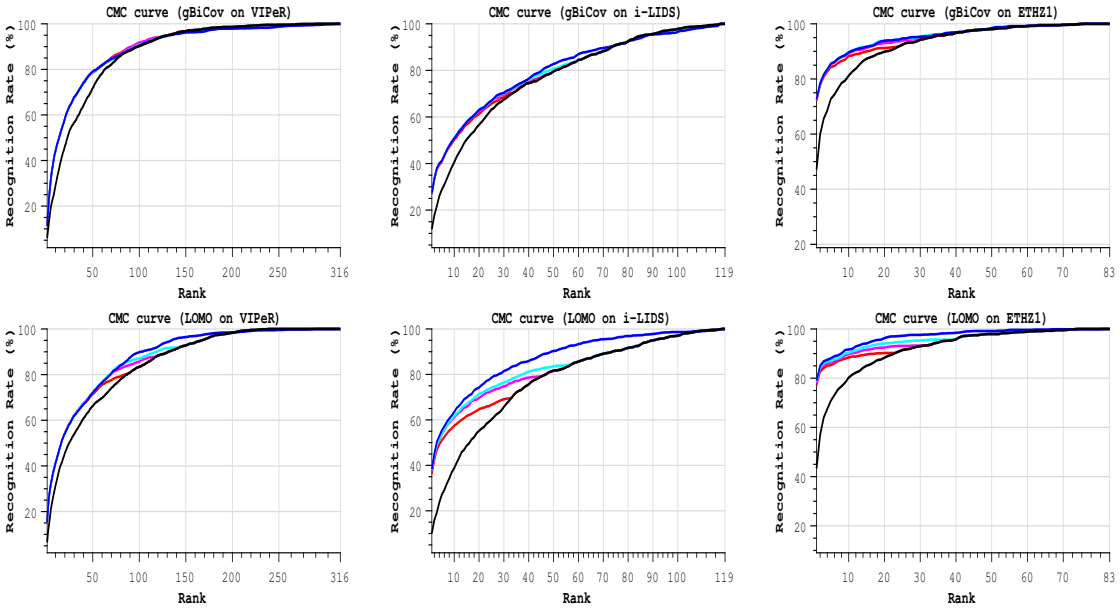

Figure 3.6: CMC curves of two-stage systems where the different versions descriptors obtained by PCA feature reduction method. Black: first stage; blue: second stage (original descriptor); red, pink, and cyan: two-stage systems with $\beta = 0.3, 0.4, 0.5$, respectively. Enlarged version of plots with very close CMC curves are shown for better visualization.

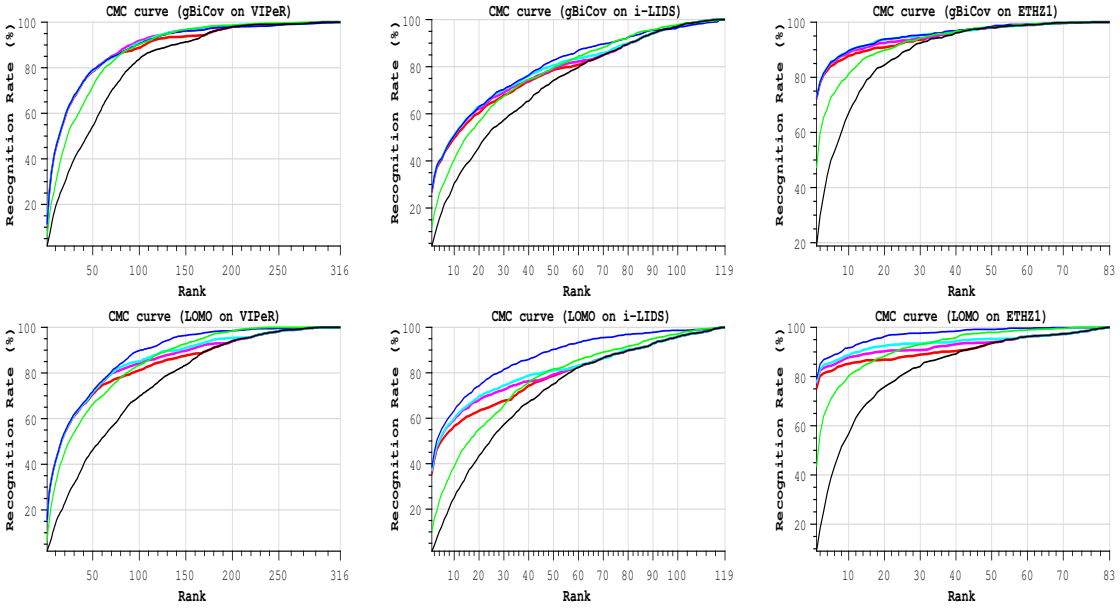

Figure 3.7: CMC curves of three-stage systems where the different versions descriptors obtained by PCA feature reduction method. Black: first stage; green: second stage; blue: third stage (original descriptor); red, pink, and cyan: three-stage systems with $\beta = 0.3, 0.4, 0.5$, respectively. Enlarged version of plots with very close CMC curves are shown for better visualization.

Chapter 4

Comparative Study of the Behavior of Feature Reduction Methods in Person Re-identification Task

Dimensional reduction is an essential pre-processing step in machine learning techniques (i.e. classification). Generally speaking, reducing the high-dimensional feature space into a low-dimensional feature space can be achieved by a dimensional reduction method, in which new, low-dimensional features are derived from the original feature space. It is desirable to achieve a low-dimensionality features as low as possible, not only to reduce the computational load, but also to make the system robust [11].

For a given $\mathbf{X}_P$, a standard re-identification system computes the matching scores $m(\mathbf{X}_P, \mathbf{X}_{T_i})$, $i = 1, \ldots, n$, and returns the list of template images ranked for decreasing values of the score. Ranking accuracy is typically evaluated using the cumulative matching characteristic (CMC) curve, i.e., the probability (recognition rate) that the correct identity is within the first ranks. Hereinafter I consider only the generated *fixed-size* feature vector (e.g. X) by a specific descriptor. Figure 4.1 presents the whole scheme of the strategy; aiming to employing a feature reduction method on person re-identification task

Apparently, some redundancies of patterns can be occurred within a feature vector, which are intuitively effected on processing time on real-time applications. It is worth to remind the readers that the issue of processing time in person re-identification can be categorized from two point of views: the processing time of constructing descriptor (*aka descriptor generation*); which can be done off-line for the gallery set, and the processing time of computing matching score between pair of descriptors

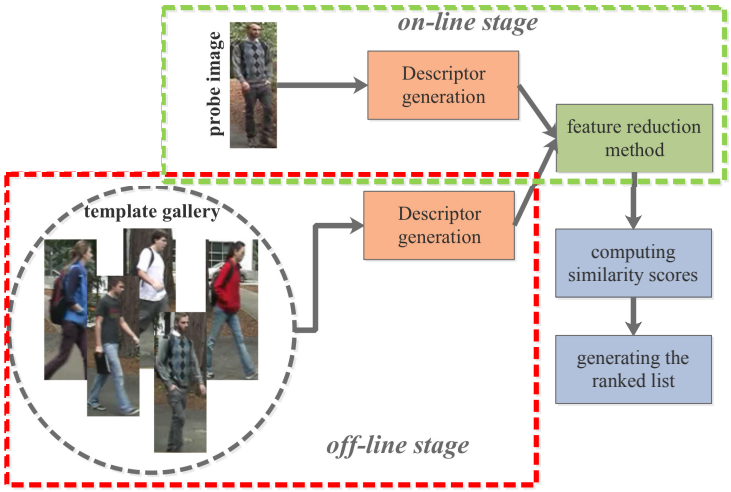

Figure 4.1: Application of a feature reduction method in person re-identification.

(*aka descriptors matching*); which has to be done on-line for investigating an individual of interest (i.e. probe image) as the procedure of the real-time application. At this thesis work, whereas a feature reduction method needs a training phase to project the proper patterns into the low-dimensional space as same as a step of the re-identification system which needs to construct the descriptor for each individual. I consider, instead, the issue of the matching processing time of a single probe image and a template image. I therefore study on the feature vectors generated by the descriptors, and investigate an empirical procedure to attain a significant trade-off between processing time and ranking quality in person re-identification task.

Moreover, having redundant and irrelevant patterns from the feature vectors might be caused in overfitting problem. Removing these irrelevant pattern from the feature space before tacking them in real-world application scenarios is know as prepare data step in machine learning processes. To sum up, there might be three key advantages of feature reduction methods:

1. decrease the risk of overfitting; which allows the algorithm to make a decision in less redundant data.

2. improve the recognition accuracy; which avoid the algorithm by occurrence of misleading those irrelevant data.

3. decrease the processing time; which leads the method to be faster.

4.1 Feature reduction methods

Usually, reduction in high-dimensional feature is achieved by subspace projection. There are many existing linear projection methods as well as their non-linear version. PCA [32] is a well-known method in terms of compressing data pattern which consists of calculating the Eigenvectors of the covariance matrix of the original feature space, and describing the variation of a set of variables in terms of a reduced set of uncorrelated linear space of such variables with maximum variance (*aka principal components (PCs)*). KPCA [60] is the nonlinear version of PCA in which the original feature space is mapped to a higher-dimensional features space using a kernel function, and then PCA is calculated. Isomap (*aka* isometric feature mapping) is popular in terms of computing quasi-isometric from a high-dimensional feature space to a low-dimensional feature space . Isomap is highly efficient and applicable to a wide range of data points and dimensions [3]. The isometric feature space can be supposed as a kernel function and so this method can also be known as a type of KPCA technique. At this section, briefly explaination these methods are presented.

4.1.1 Principal Component Analysis (PCA)

PCA is pretty a well-known method for linear dimensional reduction. This method leads to identify important patterns within a data, and express the data in low-dimensional patterns by keeping the nature of the data at the same time(*aka* compressing data). PCA, typically, employs Singular Value Decomposition (*SVD*) of the features to project a feature vector into a lower dimensional space. SVD can be more fundamental in the concept of feature reduction method, since not only it provides direct approach to compute the principle components(PCs), but also simultaneously helpful to obtain row and column spaces [72]. At the following, I go through some brief introductions of the mathematical point of views of SVD.

Let X = $\{x_1, x_2, \ldots, x_N\}$ be the given feature vector of size N to be compressed. While X is denoted as $X \in R^N$. The SVD of X is defined as

$$X = USV^H \tag{4.1}$$

where $U \in R^{N\times N}$ and $V \in R^{N\times N}$ are unitary matrices, and $S \in R^{N\times N}$ as a diagonal matrix, $S = diag\{\alpha_1, \ldots, \alpha_r, 0, \ldots, 0\}$. The singular values are ordered in decreasing order, $\alpha_1 \geq \ldots \geq \alpha_r \geq 0$. Accordingly, in many applications, it can be useful to approximate X by considering whole matrix as

$$X = \begin{bmatrix} U_r & U_{n-r} \end{bmatrix} \begin{bmatrix} S_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^H \\ V_{n-r}^H \end{bmatrix} \tag{4.2}$$

where $U_r = (u_1, u_2, \ldots, u_r)$, $V_r = (v_1, v_2, \ldots, v_r)$, and $S_r = diag\{s_1, s_2, \ldots, s_r\}$ which are used to project the original feature space into the low-rank feature space of rank r. Therefore, the approximate of matrix X with low-rank matrix is computed as

$$\bar{X} = U_r S_r V_r^H \tag{4.3}$$

where $\bar{X}$ is the final projected feature space in $rank = r$.

4.1.2 Kernel Principal Component Analysis (KPCA)

Kernel-PCA (KPCA) is an improved theory of traditional linear PCA in a high-dimensional space which is constructed by employing a kernel function. On the other words, KPCA is a non-linear dimensional reduction for the features to project a lower dimensional space using the kernel method, and then compute PCA on the high-dimensional feature space. However, KPCA can be applied based on a specific kernel of k, this leads the application to choose a proper kernel function [60]. Given a data set X of input samples $\{x_1, x_2, \ldots, x_N\}$, a kernel is defined as follows

$$\begin{aligned} k : \mathrm{X} \times X &\rightarrow \mathbb{R} \\ (x_i, x_j) &\mapsto k(x_i, x_j), \end{aligned} \tag{4.4}$$

where the kernel $k(.,.)$ gives a scalar that describes the similarity of the samples x_i and x_j. In this work, the Gaussian kernel (RBF) is employed as follow

$$k(x_i, x_j) = exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{4.5}$$

Using the obtained kernel, the originally linear operations of PCA are performed (as explained in 4.1.1) to reduce the dimensionality of the kernel feature space. The Gaussian kernel (RBF) is chosen in this studying because the linear kernel function gives the same performance as applying the normal PCA in the original feature space.

4.1.3 Isomap

Isomap is a non-linear dimensional reduction through isometric feature mapping which consists of calculating quasi-isometric to obtain low-dimensional embedding of a set of high-dimensional data points. The algorithm is based on estimating of geometry features of data distribution, and then mapping data to a new space. Isomap is quite straightforward technique that resolves feature reduction problem by computing geodesic distances between each data point. Geodesic distance basically computes between two points over the manifold. In order to compute the distances between data

points x_i, $i = 1,2,\ldots,N$, a neighborhood graph of G is constructed in which each data point x_i is connected with its k nearest neighbors x_{ij}, $j = 1,2,\ldots,k$, within a data set, and forming an estimation of the geodesic distance between two points by taking into account the shortest path between them. One can compute the shortest-path between all the data points by employing Dijkstras [15] or Floyds algorithms [23], and form into a pairwise geodesic distance matrix.

The whole procedure of the isomap feature reduction method can be summurized in three steps as follow:

1. Constructing the neighborhood graph by assigning to each data point with its neighbors.
2. Measuring shortest paths on the neighborhood graph and create a graph distances matrix.
3. Constructing an embedding of the data in R^d from the graph distances matrix.

The computational complexity of this method, only the second step of it, is the most time consuming and it is performed in $O(N^3)$ operations.

4.1.4 Reconstruction Error

The performance of a re-identification system is typically measured using the CMC curve, defined as the probability that the correct identity is within the first rnk ranks, for rnk = $1,\ldots,n$. By definition, the CMC curve increases with rnk, and equals 1 for rnk = n. Whereas, this work aims to reduce the feature space of the original descriptor, I employ reconstruction error to estimate the variances the projected feature space. However, this can be simply identified from the behaviour corresponding CMC curve, but for sake of comparison of different reduction methods, the reconstruction error is computed between the original feature space and the projected feature space. In order to estimate the reconstruction error of the projected feature vector, I employed Frobenius norm. To this aim, by recalling the projected feature space ($\bar{X}$) and the original feature vector(X), the reconstruction error is estimated as

$$E = \frac{\left\| X - \bar{X} \right\|_F^2}{\|X\|} \tag{4.6}$$

4.2 Experimental evaluation

The comparison has been carried out with two well-known descriptors in person re-identification problem: gBiCov and LOMO, on two benchmark data sets: VIPeR and i-LIDS. More details of the used descriptors as well as the data sets have been discuss in Sect. 3.4.1 and Sect. 2.4, respectively.

These descriptors are gBiCov and LOMO; which have been chosen because of their fixed-size image representation. The generated original image representation by gBiCov contains ≈ 6000 elements, while it is ≈ 27000 for LOMO descriptor.

4.2.1 Experimental setup

One image for each person was randomly selected to build the template gallery; the other images formed the probe gallery. As in [21], for each data set the experiment is repeated on ten different subsets of individuals, using one image of each individual as template and one as probe, and reported the average CMC curve over the ten runs, used an Intel Core i5 2.6 GHz CPU. The above-mentioned feature reduction methods are applied on two well-known descriptors on VIPeR data set in person re-identification task.

The original feature generated by a *given* descriptor (X), are reduced for different sizes in $r = \{2, 5, 20, 50, 80, 100, 130, 150, 200, 300, 500, 800, 1000, 1200, 2000, 2500, 3500\}$. For KPCA, and the Gaussian kernel is chosen because of its good performances. For Isomap, I set the number of neighborhoods to $k = 5$.

4.2.2 Experimental results

Figures 4.2 and 4.3 present the corresponding CMC curves obtained by using different descriptors as well as different feature reduction methods on VIPeR, and i-LIDS data sets, respectively. The CMC curves are presented only in the first ranks for better visualization, however, the first ranks are typically taken into consideration for the determination of the power of a re-identification system for real-time application. PCA as a standard technique, which the new techniques are still unable to outperform it. KPCA also have a very similar behaviour in terms of the recognition accuracy in person re-identification. Also, in both techniques, the recognition accuracy outperformed the original CMC curve on LOMO descriptor when PCA and KPCA are applied for the original feature vector, and in the same way, with slightly better performance by using gBiCov on VIPeR and i-LIDS data sets. In contrast, KPCA is very time consuming because of the computational complexity when the feature vector is relatively larger in comparison to the other methods. As pointed out in Sect. 4.1.4, for the sake of comparison between the original descriptor and the projected descriptor into the new feature space, the measure of reconstruction error can be simply computed between two feature spaces. Figures 4.4 and 4.5, therefore, demonstrate the estimated errors among different feature reduction methods for different values of r. Apparently from the presented figures, the error leads to be zero at the certain

value of $r > 1000$ on VIPeR, and $r > 500$ for i-LIDS data set where LOMO and gBicov descriptors with PCA and KPCA are employed. However, Isomap achieves better performances only on i-LIDS data set with respect to its behaviour on VIPeR data set, and this is also obvious from the computed error estimations in which the error leads to be zero for $r > 500$.

The average processing time for computing one matching score, evaluated on VIPeR and i-LIDS, are reported in Fig. 4.6. Similar processing times were observed in all data sets, due to the use of the same image size.

4.3 Conclusions

At this chapter, the performances of most popular fundamental dimensional reduction approaches were compared between PCA, KPCA, and Isomap feature reduction methods on person re-identification data sets. The comparison is done through the experiments conducted by using two descriptors on two benchmark data sets. The experimental results evidenced that generated features by these descriptors might be not well-optimum. PCA and KPCA outperformed the original CMC curve on LOMO and gBiCov descriptors on VIPeR and i-LIDS data sets. This was apparent also from their error estimation of projected feature space using two descriptors on two data sets. Both these reduction methods achieve better performances rather than Isomap method in person re-identification task. The reason relies in the fact that PCA and KPCA can explore higher order information of the original inputs than Isomap. It is worth to point out that, at this work, PCA was better than others in terms of the computational cost, while KPCA was more time consuming with respect to the other two reduction methods. It therefore can be stated that PCA achieved promising performance for handling of optimization of raw data and projection of it to low-dimensional feature space. This has only been studied for the descriptors with fixed-size feature vector. Finally, I point out that the optimization of the dimensional reduction methods analyzed in this paper is computationally and numerically practical in real-time applications. As the future work, I aim at carefully study the behavior of these feature reduction methods by concerning on some analytical terms, and visualize the projected data on the actual feature space to get better prospective on those behaviours.

4.4 Acknowledgment

I gratefully acknowledge Prof. Giuseppe Rodriguez of Dipartimento di Matematica e Informatica at University of Cagliari, for his initial analysis on the data and experimental evaluations of this chapter.

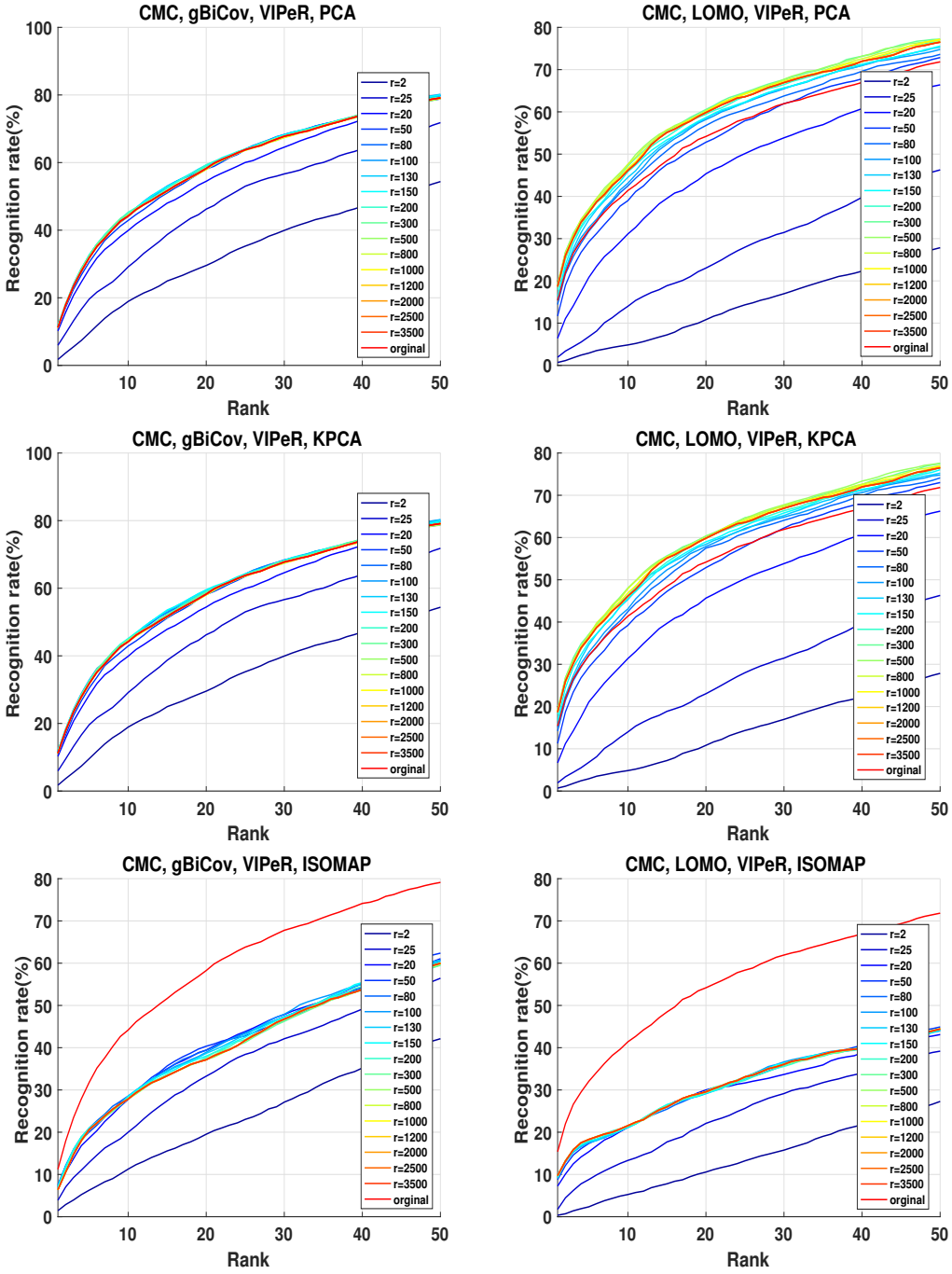

Figure 4.2: CMC curves obtained by gBiCov and LOMO descriptors on VIPeR data set in which the feature reduction methods have been employed.

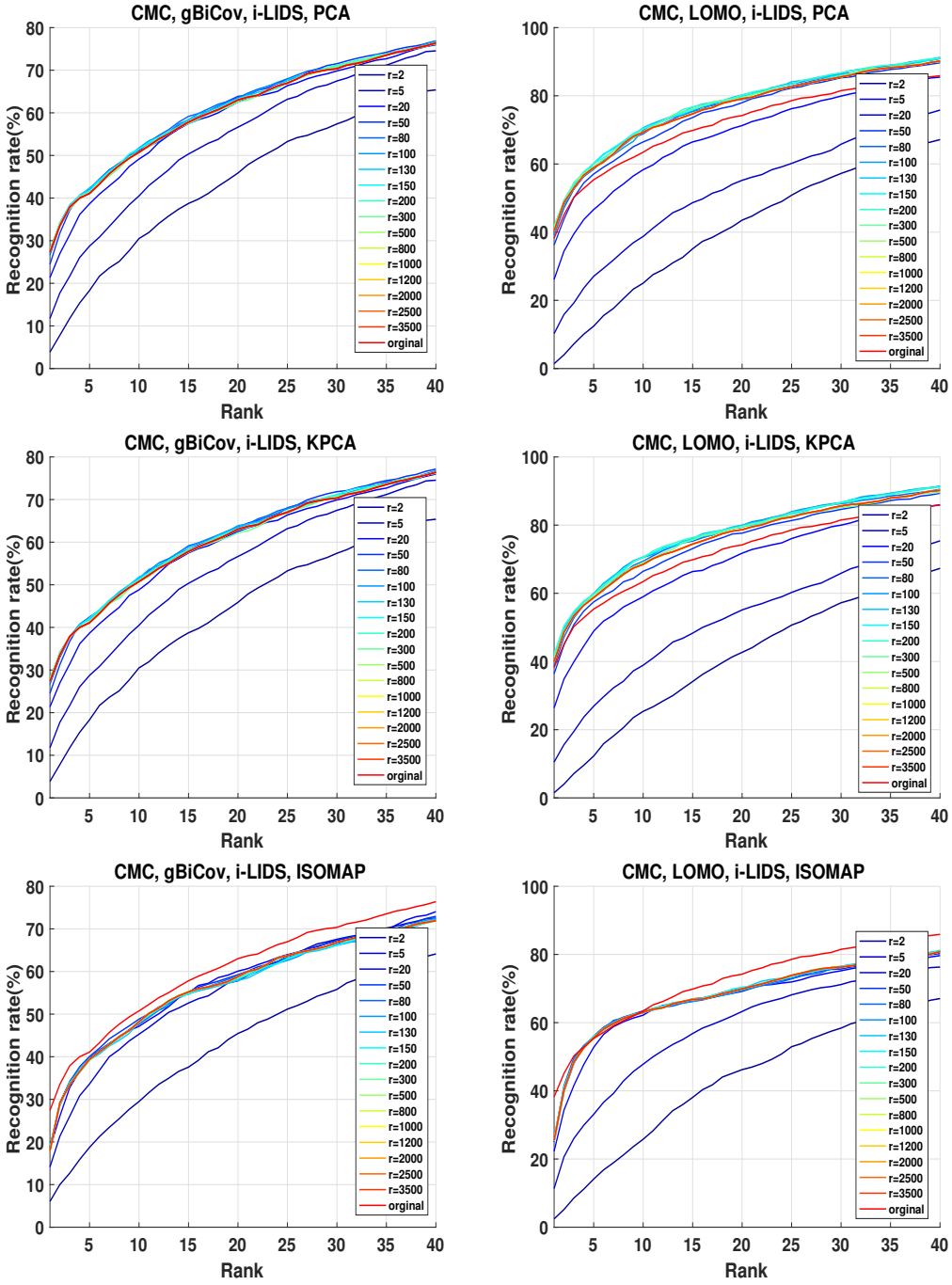

Figure 4.3: CMC curves obtained by gBiCov and LOMO descriptors on i-LIDS data set in which the feature reduction methods have been employed.

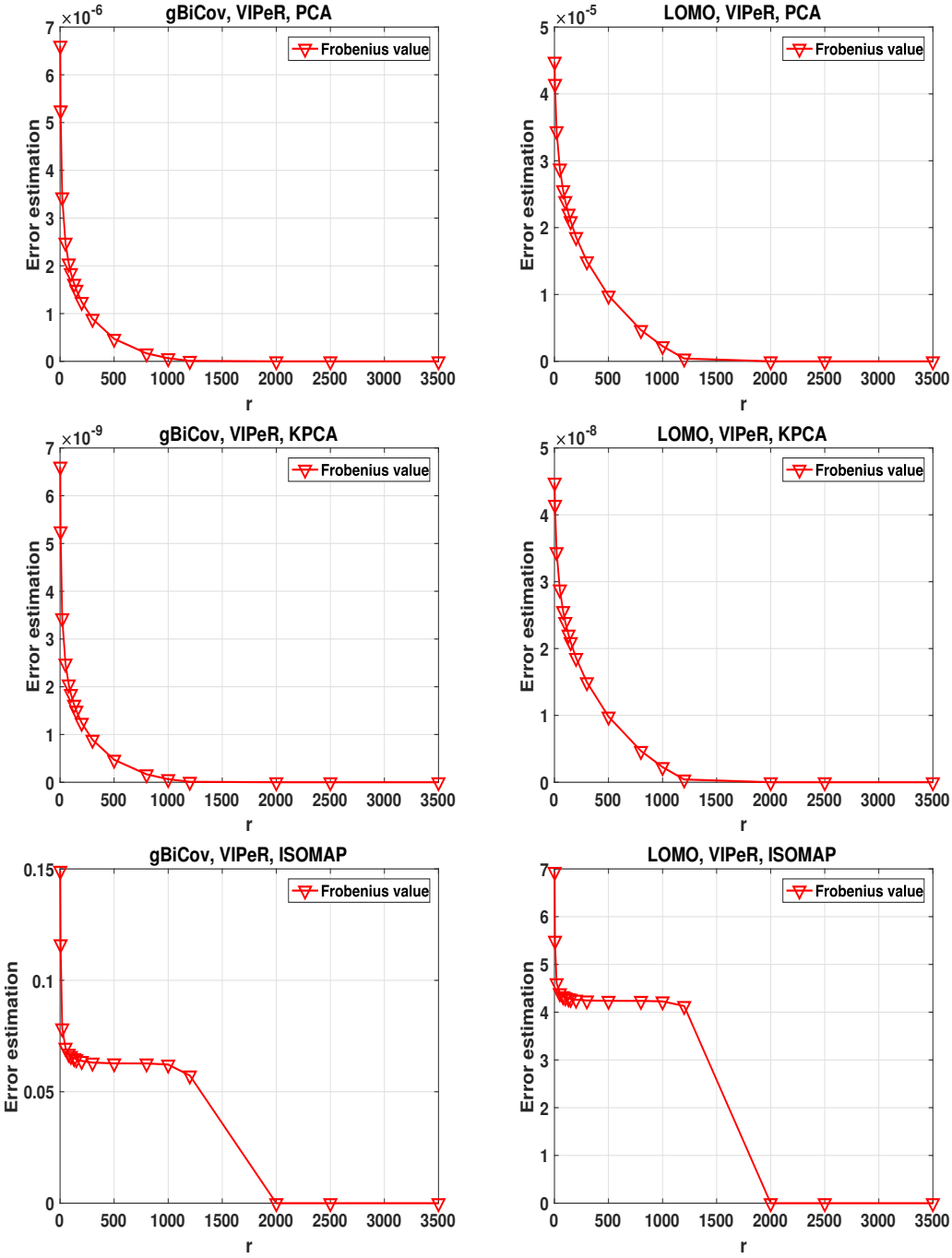

Figure 4.4: Reconstruction errors of different feature reduction methods by using gBiCov and LOMO descriptors on VIPeR data set.

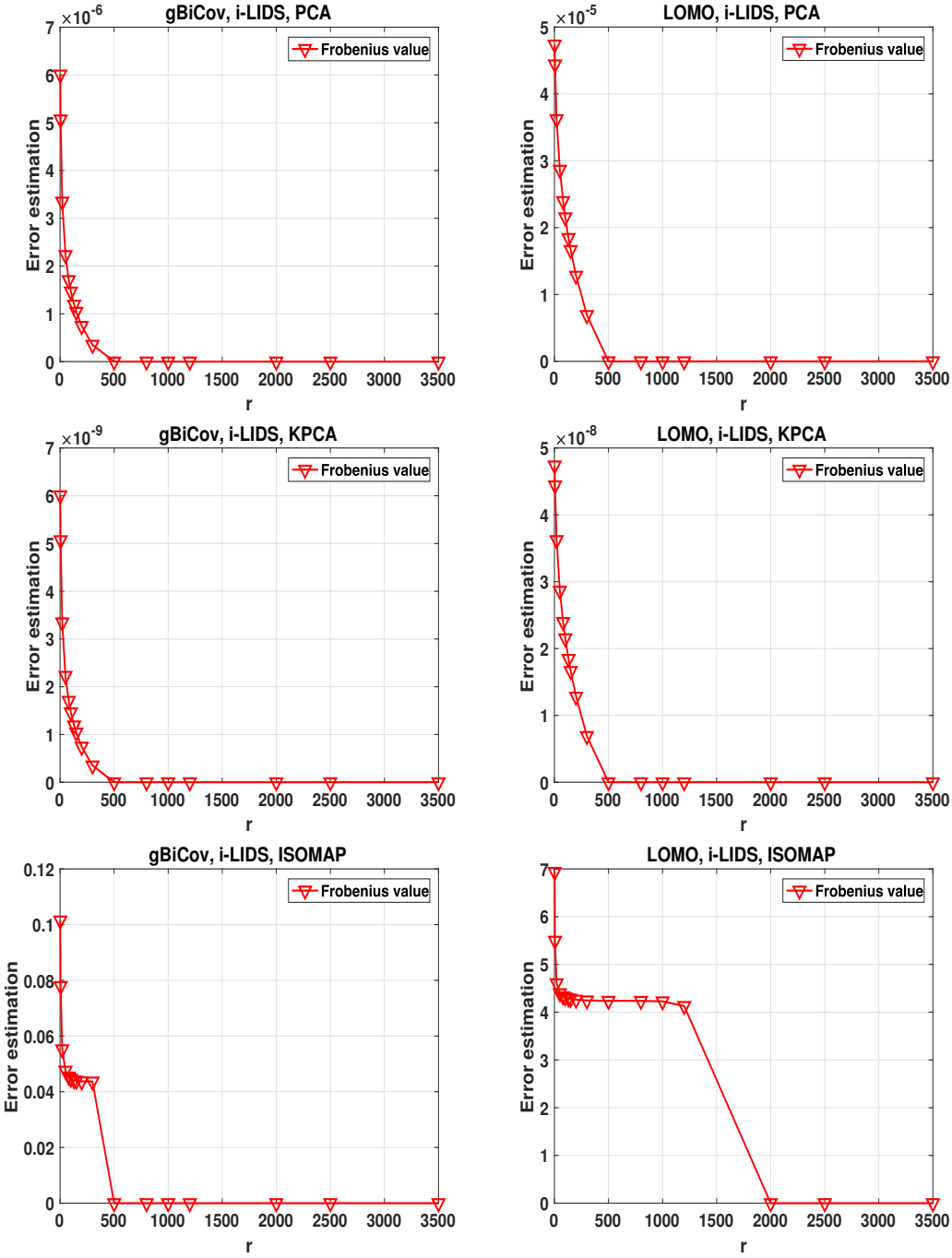

Figure 4.5: Reconstruction errors of different feature reduction methods by using gBiCov and LOMO descriptors on i-LIDS data set.

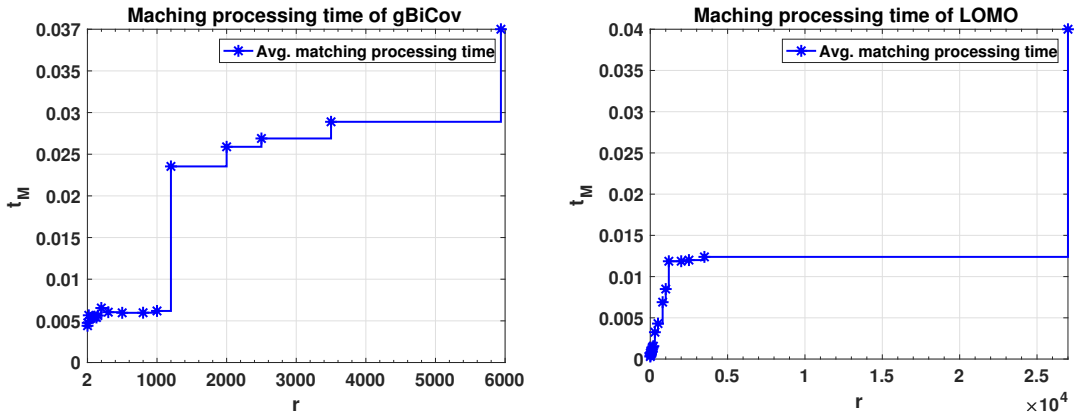

Figure 4.6: Average processing time t_M (in sec.) for computing a matching score for a single probe image and one template, for each of the two descriptors with different feature sizes r.

Chapter 5

Discussion and conclusions

This thesis work presented a contribution to the literature on *Intelligent Video Surveillance systems*, which is one of attracting and interesting task among researchers and industries due to a continuously growing demand of security and safety in crowd. In particular, the thesis addressed, specifically, the issue of the processing time for person re-identification problem, that could provide a tools of video-surveillance operators and forensic investigators to be proceeded swiftly. The ultimate goal for a generic re-identification system is the capability of it in real-time applications. To this aim, the whole re-identification procedure must be quick enough, and enables to design and implement such a task in real-world application.

This conclusive chapter closes the thesis. First, the major contributions of this work are stated in Sect. 5.1, and then, Sect. 5.2 provides future research directions to enrich and extend the presented work.

5.1 Contributions of this thesis

Person re-identification is one the most challenging tasks of intelligent video surveillance system with broad open areas of its application in numerous fields. It has received lots of attention by many researchers and the re-identification methods and recognition techniques become a long way but are still very narrow and specific to apply them on real world problems.

At this thesis work, I proposed a multi-stage ranking approach for person re-identification task; in which the goal was to achieve a trade-off between processing time and ranking accuracy for any *given* appearance descriptor of person re-identification. Th approach was inspired by a well-known multi-stage classification architecture which widely used in pattern recognition systems, and I therefore

adapted it to a ranking problem by developing an ad hoc analytical model of the trade-off between its ranking accuracy and processing time. The multi-stage ranking system could be useful in real-world applications that are supposed to be involved by interaction with human operators. It could also be useful in an application scenario when the operator cannot or does not want to scan all the ranked template images. Empirical evidence on the used data sets, using different state-of-the-art descriptors, showed that the proposed ranking approach was capable of reducing processing time, by keeping the ranking quality of the original descriptor. The proposed multi-stage ranking system has been well positioned in the application of real-time video intelligent surveillance systems. To the best of author's knowledge, multi-stage ranking approach is unique on its way that explicitly explored the issue of the processing time in person re-identification task, by its quite straightforward implementation strategy. Considering to the point: the computational complexity of many methods is too high to be used in real-time applications (e.g., SDALF); at this thesis, I addressed this issue more explicitly, however, a more thorough analysis of the requirements for real-time re-identification systems in terms of computational resources must be taken into account.

Additionally, some feature reduction methods were studied at this thesis work including: PCA, KPCA, and Isomap. The goal was to enrich a promising trade-off between processing time and ranking accuracy. On the other hand, the results evidenced that using a feature reduction method such as PCA, could be also useful to the application of such a system like the one proposed multi-stage techniques, which also relied on the issue of achieving a trade-off between processing time and ranking accuracy. The empirical experimental evaluation also proved that using a feature reduction method not only practical in the issue of the processing time, but also can be remarkable in terms of the ranking accuracy.

5.2 Future works

The research on person re-identification is a relatively young area in pattern recognition and computer vision. Also, many aspects have still to be explored, and a large amount of work has to be adopted before a re-identification system can be employed in real-world scenarios. Although some attempts in this regard have been proposed, there still exist many open issues that must be solved before a complete real-time re-identification system being successfully implemented.

The proposed multi-stage approach in chapter 3 can be improved and also the attainable trade-off can be optimized, for a given descriptor, by suitably choosing the number of stages and the parameters to be modified in order to obtain faster (and less accurate) versions of the same descriptor in

each stage but the last one. The idea of multi-stage ranking approach could further expand by investigating on some optimization techniques (e.g. the well-known Pareto optimization as a technique for multiple criteria decision making[54]) which might be jointly optimized of all critical design parameters. By this, the multi-stage approach is leaded to be constructed/designed as a novel multi-stage system which enables to autonomously determined the number of stage, as well as, constructing the different simplified versions of a given descriptor to gain the significant trade-off of between processing time and ranking accuracy.

The study which has been investigated in chapter 4, showed also the ability of feature reduction methods in terms of the reducing the processing time; which attained by reducing the feature space of the original descriptor, and improve the quality of the ranking accuracy in some descriptor and data sets (e.g. gBicov on VIPeR and i-LIDS data sets where PCA has been employed). This behaviour can be studied further in terms of the better recognition accuracy which achieved after some reduction on the original feature vector. This is typically caused due to the well-known problem in pattern recognition, so-called *the curse of dimensionality*. The curse of dimensionality is the fact when the number of features or dimensions are too large; this leads the machine learning with some difficulties in training stage (e.g. the algorithm can be easily failed caused by overfiting during the the training process). In this manner, a feature reduction method can impede the possibility of this kind of issues by removing the redundant or irrelevant patterns from the data space.

Bibliography

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015. [cited at p. 15]

[2] Slawomir Bak, Etienne Corvee, François Bremond, and Monique Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 179–184. IEEE, 2011. [cited at p. 11]

[3] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002. [cited at p. 43]

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Computer vision–ECCV 2006*, pages 404–417, 2006. [cited at p. 11]

[5] Apurva Bedagkar-Gala and Shishir K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270 – 286, 2014. [cited at p. vii, viii, 2, 20, 21, 23]

[6] Apurva Bedagkar-Gala and Shishir K Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014. [cited at p. 1, 2]

[7] Home Office Scientific Development Branch. Imagery library for intelligent detection systems (i-lids). In *Crime and Security, 2006. The Institution of Engineering and Technology Conference on*, pages 445–448. IET, 2006. [cited at p. vii, 4, 20]

[8] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016. [cited at p. 15, 17]

[9] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. [cited at p. 16, 17]

[10] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, page 6, 2011. [cited at p. 22]

[11] Tommy WS Chow and D Huang. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transactions on Neural networks*, 16(1):213–224, 2005. [cited at p. 41]

[12] Etienne Corvee, Francois Bremond, Monique Thonnat, et al. Person re-identification using spatial covariance regions of human body parts. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 435–440. IEEE, 2010. [cited at p. 11]

[13] Angela D'Angelo and Jean-Luc Dugelay. People re-identification in camera networks based on probabilistic color histograms. In *Visual Information Processing and Communication*, page 78820K, 2011. [cited at p. 10, 12]

[14] Icaro Oliveira De Oliveira and José Luiz de Souza Pio. People reidentification in a camera network. In *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*, pages 461–466. IEEE, 2009. [cited at p. 11]

[15] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959. [cited at p. 45]

[16] Mert Dikmen, Emre Akbas, Thomas S Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Computer Vision–ACCV 2010*, pages 501–512. Springer, 2010. [cited at p. 12]

[17] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. [cited at p. 17]

[18] Cristianne RS Dutra, William Robson Schwartz, Tiago Souza, Renan Alves, and Lara Oliveira. Re-identifying people based on indexing structure and manifold appearance modeling. In *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on*, pages 218–225. IEEE, 2013. [cited at p. 4, 18]

[19] Markus Eisenbach, Alexander Kolarow, Konrad Schenk, Klaus Debes, and Horst-Michael Gross. View invariant appearance-based person reidentification using fast online feature selection and score level fusion. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 184–190. IEEE, 2012. [cited at p. 11]

[20] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. [cited at p. 20]

[21] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010. [cited at p. 2, 4, 10, 11, 12, 27, 31, 33, 34, 46]

[22] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. [cited at p. 21]

[23] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188, 1936. [cited at p. 45]

[24] Itzhak Fogel and Dov Sagi. Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113, 1989. [cited at p. 11]

[25] Alexandre Franco and Luciano Oliveira. A coarse-to-fine deep learning for person re-identification. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–7. IEEE, 2016. [cited at p. 15]

[26] Alexandre Franco and Luciano Oliveira. Convolutional covariance features: Conception, integration and performance in person re-identification. *Pattern Recognition*, 61:593–609, 2017. [cited at p. 15]

[27] Tarak Gandhi and Mohan M Trivedi. Panoramic appearance map (pam) for multi-camera based person re-identification. In *Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on*, pages 78–78. IEEE, 2006. [cited at p. 12]

[28] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008. [cited at p. vii, 2, 4, 10, 12, 13, 19]

[29] Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu, and Bruno Steux. Interest points harvesting in video sequences for efficient person identification. In *The Eighth International Workshop on Visual Surveillance-VS2008*, 2008. [cited at p. 10]

[30] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011. [cited at p. 11, 21, 26, 27]

[31] Martin Hirzer, Peter M Roth, and Horst Bischof. Person re-identification by efficient impostor-based metric learning. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 203–208. IEEE, 2012. [cited at p. 2, 10, 12]

[32] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933. [cited at p. 43]

[33] Shuai Huang, Yun Gu, Jie Yang, and Pengfei Shi. Reranking of person re-identification by manifold-based approach. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4253–4257. IEEE, 2015. [cited at p. 26, 27]

[34] Ryo Kawai, Yasushi Makihara, Chunsheng Hua, Haruyuki Iwama, and Yasushi Yagi. Person re-identification using view-dependent score-level fusion of gait and color features. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2694–2697. IEEE, 2012. [cited at p. 10, 12]

[35] Cenk Kaynak and Ethem Alpaydin. Multistage cascading of multiple classifiers: One man's noise is another man's data. In *ICML*, pages 455–462. Citeseer, 2000. [cited at p. 26]

[36] Mohamed Ibn Khedher and Mounim A El Yacoubi. Two-stage filtering scheme for sparse representation based interest point matching for person re-identification. In *Advanced Concepts for Intelligent Vision Systems*, pages 345–356. Springer, 2015. [cited at p. 4, 18, 19]

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. [cited at p. 2]

[38] Bahram Lavi, Giorgio Fumera, and Fabio Roli. A multi-stage approach for fast person re-identification. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 63–73. Springer, 2016. [cited at p. 6]

[39] Bahram Lavi, Giorgio Fumera, and Fabio Roli. A multi-stage approach for fast person re-identification. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+SSPR 2016, Mérida, Mexico, November 29 - December 2, 2016, Proceedings*, pages 63–73. Springer International Publishing, 2016. [cited at p. 25]

[40] Bahram Lavi, Giorgio Fumera, and Fabio Roli. Multi-stage ranking approach for fast person re-identification. *IET Computer Vision*, January 2018. [cited at p. 6]

[41] Bahram Lavi, Mehdi Fatan Serj, and Domenec Puig Valls. Comparative study of the behavior of feature reduction methods in person re-identification task. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, pages 614–621. INSTICC, SciTePress, 2018. [cited at p. 7]

[42] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3594–3601. IEEE, 2013. [cited at p. vii, 21, 22]

[43] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31–44. Springer, 2012. [cited at p. 21]

[44] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. [cited at p. 2, 14, 21]

[45] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. [cited at p. 4, 27, 31, 34]

[46] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 441–448. IEEE, 2013. [cited at p. 26, 27]

[47] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*, 2016. [cited at p. 15, 17]

[48] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. [cited at p. 10]

[49] Jiwen Lu and Erhu Zhang. Gait recognition for human identification based on ica and fuzzy svm through multiple views fusion. *Pattern Recognition Letters*, 28(16):2401–2411, 2007. [cited at p. 10]

[50] Bingpeng Ma, Yu Su, and Frederic Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6):379–390, 2014. [cited at p. 2, 4, 11, 27, 31, 34]

[51] Niki Martinel and Christian Micheloni. Re-identify people in wide area camera network. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 31–36. IEEE, 2012. [cited at p. 10]

[52] Michael J Metternich and Marcel Worring. Track based relevance feedback for tracing persons in surveillance videos. *Computer Vision and Image Understanding*, 117(3):229–237, 2013. [cited at p. 26, 27]

[53] Jürgen Metzler. Appearance-based re-identification of humans in low-resolution videos using means of covariance descriptors. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 191–196. IEEE, 2012. [cited at p. 11]

[54] Patrick Ngatchou, Anahita Zarei, and A El-Sharkawi. Pareto multi objective optimization. In *Intelligent Systems Application to Power Systems, 2005. Proceedings of the 13th International Conference on*, pages 84–91. IEEE, 2005. [cited at p. 55]

[55] P Pudil, J Novovicova, S Blaha, and J Kittler. Multistage pattern recognition with reject option. In *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*, pages 92–95. IEEE, 1992. [cited at p. 26]

[56] Mohammad Ali Saghafi, Aini Hussain, Halimah Badioze Zaman, and Mohamad Hanif Md Saad. Review of person re-identification techniques. *IET Computer Vision*, 8(6):455–474, 2014. [cited at p. 1, 5]

[57] Riccardo Satta, Giorgio Fumera, and Fabio Roli. Fast person re-identification based on dissimilarity representations. *Pattern Recognition Letters*, 33(14):1838–1848, 2012. [cited at p. 4, 18, 19]

[58] Riccardo Satta, Giorgio Fumera, Fabio Roli, Marco Cristani, and Vittorio Murino. A multiple component matching framework for person re-identification. In *Image Analysis and Processing–ICIAP 2011*, pages 140–149. Springer, 2011. [cited at p. 4, 10, 12, 27, 34]

[59] Cordelia Schmid. Constructing models for content-based image retrieval. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2001. [cited at p. 11]

[60] Bernhard Schölkopf, Christopher JC Burges, and Alexander J Smola. *Advances in kernel methods: support vector learning*. MIT press, 1999. [cited at p. 43, 44]

[61] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, pages 732–748. Springer, 2016. [cited at p. 15]

[62] Alessandro Sperduti. Theoretical and experimental analysis of a two-stage system for classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):893–904, 2002. [cited at p. 5, 26]

[63] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491. Springer, 2016. [cited at p. 16, 17]

[64] Kirill Trapeznikov, Venkatesh Saligrama, and David Castañón. Multi-stage classifier design. *Machine learning*, 92(2-3):479–502, 2013. [cited at p. 26]

[65] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013. [cited at p. 12]

[66] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1:511, 2001. [cited at p. 5, 26]

[67] Shengke Wang, Cui Zhang, Lianghua Duan, Lina Wang, Shan Wu, and Long Chen. Person re-identification based on deep spatio-temporal features and transfer learning. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 1660–1665. IEEE, 2016. [cited at p. 15]

[68] Zheng Wang, Ruimin Hu, Chao Liang, Qingming Leng, and Kaimin Sun. Region-based interactive ranking optimization for person re-identification. In *Advances in Multimedia Information Processing–PCM 2014*, pages 1–10. Springer, 2014. [cited at p. 26, 27]

[69] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016. [cited at p. 13]

[70] Dong Yi, Zhen Lei, and Stan Z Li. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*, 2014. [cited at p. 2, 14]

[71] Guanwen Zhang, Jien Kato, Yu Wang, and Kenji Mase. People re-identification using deep convolutional neural network. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 3, pages 216–223. IEEE, 2014. [cited at p. 14]

[72] Lingsong Zhang, JS Marron, Haipeng Shen, and Zhengyuan Zhu. Singular value decomposition and its visualization. *Journal of Computational and Graphical Statistics*, 16(4):833–854, 2007. [cited at p. 43]

[73] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. [cited at p. vii, 21, 23]

[74] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. [cited at p. 13]

[75] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 649–656. IEEE, 2011. [cited at p. 10, 12]