THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Sciences Mécaniques, Acoustiques, Électronique et Robotique de Paris

Présentée par

Solène CHAN-LANG

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Closed and Open World Multi-Shot Person Re-identification

soutenue le 6 décembre 2017

devant le jury composé de:

Mme. Alice Caplier	Rapportrice
M. Andrea Cavallaro	Rapporteur
M. Fabien Moutarde	Examinateur
M. Jean-Luc Zarader	Examinateur
Mme. Catherine Achard	Directrice de thèse
M. Quoc Cuong Pham	Encadrant

Remerciements

Tout d'abord, je souhaite remercier le Laboratoire Vision et Ingénierie des Contenus du CEA LIST et l'Université Pierre et Marie Curie au sein desquels j'ai réalisé ma thèse.

Je remercie Jean-Luc Zarader pour avoir accepté de présider le jury de thèse, Alice Caplier et Andrea Cavallaro pour avoir accepté d'être rapporteurs ainsi que Fabien Moutarde pour avoir accepté d'être examinateur. Je suis reconnaissante pour le temps qu'ils ont consacré à la lecture du manuscrit et pour leurs remarques enrichissantes au sujet de mon travail. Je suis d'autant plus reconnaissante envers Alice Caplier et Andrea Cavallaro qui se sont respectivement déplacés de Grenoble et de Londres pour ma soutenance.

Je remercie Quoc Cuong Pham et Catherine Achard pour leur encadrement. J'ai tout particulièrement apprécié leur suivi régulier et leur gentillesse. Ils ont cru en moi et m'ont encouragé tout au long de la thèse. Que ce soit pour la relecture d'articles, la relecture du manuscrit ou encore pour les présentations orales, leurs conseils m'ont été précieux, ceux de Quoc Cuong montrant son recul et ceux de Catherine plus dans le détail.

J'ai beaucoup apprécié la bienveillance de l'ensemble des membres de l'équipe. En particulier, je souhaite remercier Loïc Fagot Bouquet pour les discussions nombreuses et instructives que l'on a pu échanger ainsi qu'Emma Spanjaard, Luis Tobias, Juliette Bertrand, Florian Chabot et Geoffrey Vaquette pour les moments musicaux, culturels ou encore sportifs que nous avons pu partager en dehors du laboratoire. Un grand merci à Odile Caminondo pour sa gentillesse, sa réactivité et son efficacité dans toutes les démarches administratives.

Je remercie toute ma famille, Cynthia, Sophie et Sion pour leur soutien tout au long de la thèse. Je suis très reconnaissante envers Alexandre Le Boité pour son soutien et ses conseils qui m'ont beaucoup aidé à mener à bien la thèse.

Je tiens également à remercier Teresa Colin, Arnaud Golinvaux et Pascal Grégis qui sont venus sur le plateau de Saclay pour assister à ma soutenance.

Merci à tous pour avoir été à mes côtés durant ces trois années et d'avoir contribué de près ou de loin à la réussite de ma thèse.

Abstract

More than ever, in today's context of insecurity and terrorism, person re-identification based on video surveillance images has become a hot research topic. Indeed, tracking an individual not only requires to track him within a camera, but also to re-identify him when he re-appears in other cameras.

In recent years, remarkable progress has been achieved in person re-identification, notably thanks to the availability of larger datasets composed of thousands of identities captured by several cameras where each camera captures multiple images per identity. Yet, we are still far from being able to automatically re-identify people accurately in real life.

Considering the evolution of the available research data and the real applications needs, this thesis has followed one major research axis. How can we tackle the challenging question of open world re-identification in which the person we want to re-identify might not appear in the database of known identities? A secondary research axis consisted in relevantly making use of the multiple images that are available for each identity.

The open world re-identification task we consider in this thesis consists in two subtasks: a detection task and a re-identification task. We are given a set of known identities, the gallery identities, but since we are in an open world situation, this set of known identities is supposed not to be overcomplete. Therefore, when presented a query person also referred to as probe person, the detection task aims at determining whether or not the query person is a probable known gallery person. Since the probe person might look similar to several gallery identities, the goal of the re-identification task is to the gallery identities from the most probable match to the least likely one.

Our first contribution, COPReV for Closed and Open world Person RE-identification and Verification, is mainly designed for tackling the decision aspect of the problem. We formulate the re-identification task solely as a verification task and aim at determining whether two sets of images represent the same person or two distinct people. With this information, we can find out whether the query person has been identified previously or not and if so, who he is. This is achieved by learning a linear transformation of the features so that the distance between features of the same person are below a threshold and that of distinct people are above that same threshold. The

purpose of our proposed cost function is to obtain a maximum number of well classified pairs (same or distinct people) while not favoring positive pairs (same person) or negative ones (distinct people). For a better generalization, it also encourages the distances to be as far from the threshold as possible, enabling to use the same decision threshold for the training and the testing phases.

Our second and third contributions are based on collaborative sparse representations. A usual way to use collaborative sparse representation for re-identification is to approximate the feature of a query probe image by a linear combination of gallery elements, where all the known identities collaborate but only the most similar elements are selected. Gallery identities are then ranked according to how much they contributed to the approximation. We propose to enhance the collaborative aspect so that collaborative sparse representations can be used not only as a ranking tool, but also as a detection tool which rejects wrong matches. A bidirectional variant gives even more robust results by taking into account the fact that a good match is a match where there is a reciprocal relation in which both the probe and the gallery identities consider the other one as a good match.

While our COPReV method only shows average performances on closed an open world person re-identification tasks, our bidirectional collaboration enhanced sparse representation method outperforms state-of-the-art methods for the open world scenarios.

Keywords: person re-identification, person verification, closed world, open world, sparse representation, metric learning, subspace learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

The first idea that comes in mind when evoking "person re-identification" is face recognition or recognition based on biometric data such as footprint or iris images. This kind of tasks allow long-term person re-identification, but they require high quality information extracted under strict constraints that assume the cooperation of the subject. For example, current face recognition algorithms require images captured in a frontal way with appropriate illumination and minimum image resolution. In this thesis, we work with images captured by surveillance cameras in an unconstrained way. The subject cooperation is not necessary but the drawback is that it only allows short-term re-identification. Person re-identification based on surveillance images aims at recognizing people within a day assuming that people do not change clothes. An obvious application of person re-identification is crime investigation. Since cameras are spread everywhere, person re-identification can help in tracking a suspect across multiple cameras. In the private sphere, person re-identification can be used in domotic applications for intelligent home for example. For more comfort, re-identifying each person as he enters a room enables to tune automatically the lighting and temperature according to his personal taste.

We can already sense with these two examples that depending on the application, what we expect from person re-identification differs. For crime investigation, if we are given an underground surveillance camera scene, the goal is not to identify every person present in the video, but to find out whether the suspect is in it or not and where he is heading to, so the person re-identification task would rather consists in quickly discarding dissimilar people and focus on the similar ones to determine if they really are the searched person or not. On the contrary, at home, every person should be identified so as to adapt the environment accordingly. Though many applications require to re-identify people, every application has its own specific goal and the criteria on which a person re-identification algorithm performance is evaluated should be defined consequently. This is why the person re-identification task is constantly evolving, in terms of test scenarios and in terms of evaluation metrics, so as to get closer to the actual needs of person re-identification applications.

Literally, person re-identification consists in finding someone's identity given that he has been previously identified. The set of known people is called gallery. The people whose identity we are looking for are referred to as probe or query person. A match, right match or positive pair refers to a couple of gallery and probe identities which actually represent the same person. A mismatch, wrong match or negative pair refers to a couple of gallery and probe identities which represent two distinct people. Though not equivalent, person re-identification is often described as the task of matching people across non overlapping cameras. Indeed, historically, before being first treated as an independent computer vision task in 2006, person re-identification started as a subtask of multi-camera tracking. People were tracked within a camera, and when re-appearing in another camera, person re-identification ensured that they were assigned the same identity as before, instead of creating new identities. This is the reason why in most test scenarios all gallery people are from one camera view, the gallery camera, while probe people are captured by another camera view, the probe camera. This is however changing. From pairwise camera re-identification, we are moving towards multiple cameras re-identification. In some recent person re-identification datasets, images are captured by multiple cameras (more than 2). People do not necessary appear in all the cameras. Similarly to real life, each person is captured only by a subset of cameras which differs with the person. Another evolution is the increasing number of images available per identity. Even if the person re-identification task was initially a multi-cameras tracking subtask, the test scenarios were single-shot scenarios where each identity was only represented by one image per camera. Multi-shot scenarios on the other hand allow for the exploitation of multiple images per identity per camera. These types of scenarios are developing.

Besides those changes related to an increased number of cameras and images, an even bigger evolution is occurring in the research field of person re-identification: the notion of open world is rising and brings with it many questions. Indeed, in a closed world setting, a probe person is one of the gallery people so we are certain to be able to find out his identity, even if the right match is the last one we check. On the other hand, with an open world comes the possibility that a probe person is not present in the gallery, in which case he has never been identified before and he therefore can not be re-identified (identified once more). Relaxing the closed world assumption seems more realistic, but it also brings ambiguity to the definition of the open world re-identification task. Many interpretations of "open world re-identification" are possible, so there exist different scenarios corresponding to a generalization of the closed world person re-identification task to an open world case. Some view it as an identity inference task [1] where the goal is to give a label (identity) to every bounding boxes extracted from multiple camera views during a time lapse without necessarily knowing the total number of identities. Others prefer to consider it as a group-based verification task [2, 3, 4]. The gallery is called the target set and is only composed of a few identities. The objective is to determine whether the probe person is one of the target people or not and does not require to find the exact identity of the probe person. Yet others define it as a two subtasks problem: the detection task and re-identification task. Similarly to the closed world re-id task,

there is a gallery and a probe set, and the goal is to determine the identity of the probe person but only in the case the probe identity is in the gallery. For probe people whom are not present in the gallery, they must be detected and rejected.

Despite the progress made in the field of person re-identification, the main issues are still the same as ten years ago and new challenges arise with the growing size of datasets and the apparition of new more realistic test scenarios. This is the object of the next section.

1.2 Challenges

It is a popular belief that person re-identification is an almost solved problem. This misconception is largely conveyed by investigation movies which regularly show policemen zooming on a specific area of a surveillance video image, the blurry image becomes neat and within seconds the suspect's identity is revealed thanks to key points extracted from his face. The scene seems realistic since for us, humans, re-identifying people is an easy task that we constantly perform without even consciously thinking about it. Isn't it therefore plausible for a computer to do it on a larger scale? It has been long since we acknowledged that computers perform much better than us for tasks such as calculating. Now with the booming of artificial intelligence where computers can beat even the world best GO players, people can be under the impression that computers will soon be provided with awareness and the ability of understanding, analysing and reacting to the world.

The reality is quite different. To begin with, in our daily life, our vision is not restricted to a rectangular picture and we unconsciously integrate context information. Computers are only given pictures which can moreover be of bad quality. As for images, when we see an image of a person, we immediately recognize that the images represent someone and we easily differentiate the parts of the image which represent the person and the parts of the image which belong to the background. Computers on the other hand do not understand what is represented and even distinguishing relevant information such as a pixel belonging to the person from irrelevant elements such as a background pixel is far from being an easy task. Therefore, when comparing two images, we compare information coming only from the people that are represented in the images while computers compare the two images as a whole. Computers must be told what information to extract, how to extract it, and how to compare the information extracted from different images. However, the large variability in the images captured by camera networks make it difficult to find the best description and the best comparison tool.

Moreover, it is common to find images of two distinct people that at first look more alike than two images of a given person. Indeed, depending on the conditions under which the images are taken, such as the environment, the cameras, the person detector and the people themselves, images corresponding to a given person can be quite dissimilar while distinct people's images can be visually similar.

People are asymmetric articulated body which can introduce a huge variability in the images they appear in. Depending on the phase of the walking period at which an image is captured, one person might look like an "I" or like a reversed "Y". Moreover, depending on their **pose** (front, back, side left or side right), the clothes they wear and the accessories they bring might look different and some elements might appear in some views but not in others. Some examples are shown in figure 1.1.

Figure 1.1 – Illustration of the pose variation related problems.
This figure displays 3 pairs of images, coming from the Shinpuhkan dataset [5] (left), the VIPeR dataset [6] (middle) and the ETHZ 1 dataset [7] (right). On the left, with a back view, we only see the girl's black jacket, but with a front view, her white top and skirt are visible. With the middle pair of images, two bag slings are visible in the frontal view but are not in the side view where the main part of the bag can be seen. On the right are displayed a pair of images where the person is at a different stage of the walking cycle, so his general shape is quite different.

Person re-identification is mainly based on **color** information, however those colors can undergo drastic changes due to different cameras color rendering and varying illumination conditions which mostly depend on the environment (indoor or outdoor) and the weather (sunny, cloudy, rainy, ...). Under those changing conditions, the captured clothing colors can be brighter or darker, and worst, with the presence of shadows, a uniform color garment can appear to be made of several main colors, especially from a computer point of view. Moreover, the color of people's clothing can also be greatly tampered by low image quality which can happen with video surveillance images. Some examples are shown in figure 1.2.

Figure 1.2 – Illustration of color issues.
The pair of images on the far right, taken from the GRID dataset [8], shows how low quality images can distort the color of the clothing. Since most datasets only provide images from non-overlapping camera views, the other three pairs of images, taken from iLIDS-VID dataset [9] and Shinpuhkan dataset [5], jointly illustrate the effects of having different cameras color rendering and illumination conditions. From a human point of view, despite the shadow, we can easily distinguish the red color of the coat of the man on the far left in both images. For the man in pink, it is already more difficult to tell whether his pants are actually grey of beige. For the man in black, the sun reflection on his coat might confuse a computer which could consider it being black and white while we know for sure it is a black coat.

Misalignment problems can arise due to the camera viewpoint which can be plunging or at man size, or simply due to bad detections that are not centered on the person. Moreover, due to bad detections, parts of a person can be missing and large parts of the image might correspond to irrelevant background information. Some datasets even provide bounding boxes that are wrong detections which do not depict a person. Furthermore, the size of the images can vary, which also contributes to the misalignment issue. In order to fit well in the page, the images in figures 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6 have all been either resized to the same width or to the same height, and even so, we can observe they do not all have the same proportions. Figure 1.3 illustrates the misalignment issue with some image samples.

Figure 1.3 – Illustration of misalignment and bad detections issues.
The pair of images of the far left are taken from the Shinpuhkan dataset [5] by a camera at man size and one with a plunging view. The other images are from the Market1501 dataset [10]. The bounding box around the girl in the white dress is well centered in one image, but badly cropped in the other image where her lower body is completely missing. The three images on the right are very bad detections which still capture some body parts.

Along with wrong or badly centered detections, **occlusions** also bring misleading information. There are three main types of occlusions. People can be occluded by a common fixed object (pole, sign,...) as they pass along the same area, which might make them look alike. People can be occluded by other people, so even when labelling images, people might not agree on who the image represents, leading to

a confusion in the labelling of each person. People can be occluded by a personal item, such as a luggage, which might sometimes help and sometimes hinder his re-identification.

The pair of images on the far right are taken from the Market1501 dataset [10]. They depict a woman with a white bag which partially occludes her in one of the picture, but which is not visible in the other picture. The other images are from the iLIDS-VID dataset [9]. On the left, three people with different clothing color are occluded by a fixed yellow sign. This might be interpreted by re-identification algorithms as a person's characteristic, and these people could be confused with the person wearing a yellow safety vest. In the middle right, two images of a man in white t-shirt are presented, but on one of them, he is occluded by a woman wearing a striped top who is the one at the forefront of the image.

Figure 1.4 – Illustration of the issues related to occlusion.

The **background** can play a big part in the good or the bad re-identification of people. Indeed, while we easily tell apart the person from the background, in most images, a non negligible pixel proportion corresponds to background pixels. From uniform gray background to very colourful ones, without forgetting highly textured ones, the range of background types is very large making it difficult to model and discard. Figure 1.5 shows a small panel of background types.

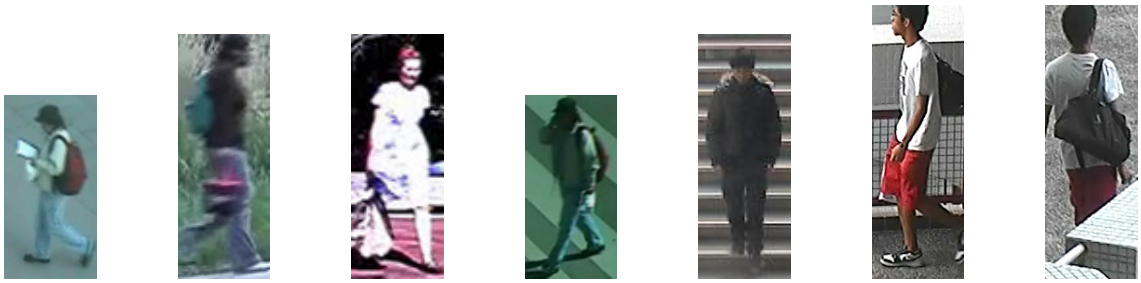


Figure 1.5 – Illustration of a small panel of background types.
Images are taken from the VIPeR [6], PRID2011 [11], CUHK03 [12] and Shinpuhkan [5] datasets. The following backgrounds are presented: a gray uniform background, a green one (grass), a pink one (racetrack), two textured backgrounds with stripes (pedestrian crossing and stairs), and two more complex background where buildings details are visible (black mesh and building wall).

All the previous elements explain and illustrate why the images of a given person can look quite different. Now this wouldn't be such a problem if different people's image looked even more dissimilar. However, since people tend to wear **similar**

clothes, two people's image can be visually very similar. A few such examples are show Figure 1.6.

Figure 1.6 – Illustration of visually similar distinct people.
Images are taken from the VIPeR [6] dataset. It shows two groups of people that look similar but who are distinct people.

The similarity between different people and the dissimilarity of images from a single person is already an issue in the closed world setting but we do know that in any case the identity of the probe person will eventually be found, should we have to look at all the gallery identities. In the **open world** setting, it becomes a more critical issue because there is no certainty about finding the identity of a probe person anymore. If the right match is not found in the first ranks, should we continue looking for it in the following ranks or declare the probe person as an imposter? The definition of open world re-identification is still evolving, its goal and evaluation as well.

Up until now, we have only seen factors that make the re-identification task very challenging. Hopefully, person re-identification datasets are growing in size and intuitively, having more images can only help. We now have access to datasets with more identities, more images per identities and more images captured by more cameras. Nonetheless, a challenge also comes with it, the **scalability** of algorithms.

Datasets might be growing in size but they are useless for most learning based methods if no annotation is available. For now, the minimum annotation required is the identity of the people represented in the images. Information about the view angle, the clothes, etc. can be a plus. Since annotations require a lot of workload, designing methods usable even with some **non labelled data** is also an interesting challenge.

1.3 Thesis objectives

At a time when new multi-cameras and multi-shot datasets are being developed and when the performance of existing methods are being questioned when applied to more realistic re-identification applications, this thesis focuses on two of the new problematics. The question of the open world re-identification task is the main research axis of this thesis. Finding a relevant exploitation of multi-shot data is also an important aspect and throughout this thesis, we only worked with multi-shot datasets. Since most existing work were conducted for the closed world person

re-identification task, for each of our proposed methods, we also present our performances for the closed world re-identification task.

For reminder, in the closed world re-identification task, the set of gallery identities are known and we have at disposal several images of each gallery person. For each probe person represented by a set of images, the goal is to return a ranked list of the gallery identities from the most likely match to the least likely one. Every presented probe person is someone who is also present in the gallery.

For the open world re-identification task, among the different tasks which aim at generalizing the closed world person re-identification task, namely the identity inference task [1], the group-based verification task [2] and detection and re-identification tasks [13], it is the last one that we tackle in this thesis and which we consider as being an open world person re-identification task. Similarly to the closed world case, the gallery set is the set of known identities and the probe set is the set of query people whose identity we are looking for. Contrary to the closed world case, a probe person might not be present in the gallery. Probe imposters are probe people who do not belong to the gallery. Non imposter probe people have their match in the gallery. The detection subtask consists in detecting and rejecting probe people that are considered as not belonging to the gallery. The re-identification subtask consists in ranking the gallery identities from the most likely match to the least likely one for a probe person who is considered as belonging to the gallery.

Closely related to this open world re-identification task we tackle is the person verification task which is never discussed in any paper and which we will evoke throughout this thesis along with the open world case. We define person verification task as the task of determining whether two sets of images corresponds to the same person or to two distinct people.

As for the multi-shot aspect, the way most existing methods exploit the availability of multiple images for each person is as follows. Given a method which assigns a similarity or dissimilarity score to a pair of images, where one image comes from a probe person and the other one from a gallery person, the method is extended to a method which assigns a score to a pair of sets of images, where one set of images corresponds to the probe person's images and the other set of images represents the gallery person, by using a simple aggregation function. The score of a pair of identity is defined as the aggregation of all the scores of the pairs of images involving these two people. The aggregation function can for example be the maximum function if the score in question is a similarity score or the minimum function if the method uses a dissimilarity score. With this approach, before the aggregation, each pair of images is considered separately. Though less emphasis has been put into this question, one of the objectives of this thesis is to propose a better exploitation of the multi-shot aspect of data which does not consider independently each image of a given person but considers them jointly.

1.4 Thesis outline

The thesis is organized as follow:

- **Chapter 1** positions this thesis with regards to the evolution of the person re-identification task. We present the main challenges of person re-identification tasks and point out the ones we focus on in this thesis. Finally, we introduce the thesis outline.
- **Chapter 2** gives an overview of the main work that have been conducted in the field of person re-identification. After a short description of the evolution of the main steps in person re-identification systems, we introduce the person re-identification datasets available for research purpose. Then we move on to the different methods proposed in the literature which we divide into two groups: methods proposed in the context of closed world re-identification and methods that specifically tackle open world re-identification. Finally we present the evaluation measures used to assess the performance of person re-identification methods.
- **Chapter 3** examines the differences between closed and open world re-id tasks. This preliminary study leads to the presentation of our COPReV method. COPReV stands for Closed and Open world Person RE-identification and Verification because this approach is meant to perform well for the usually dealt with closed world re-id task, for the open world re-id task we tackle, and for the person verification task. In COPReV, the re-id task is cast as a binary classification task. A cost function is proposed to learn a linear transformation of the features so that the Euclidean distance of the projected features of a positive pair of images (coming from the same person) is smaller than a given threshold and the distance of a negative pair (corresponding to distinct people) is bigger than the same threshold. The decision threshold is fixed at the training phase and re-used at the test phase, so that this approach can be easily employed for the open world re-identification detection subtask and for the person verification task. The formulation of the cost function aims at enforcing pairs distances to be as far from the threshold as possible during the training phase so that it generalizes well and does not favor positive nor negative pairs.
- **Chapter 4** presents a completely different approach specifically designed for the open world re-id task which is based on sparse representations and on an enhancement of the collaboration between gallery identities. After a brief introduction on sparse representations, we point out the difference between non collaborative and collaborative sparse coding which makes collaborative sparse coding much more performing than the non collaborative approach for the closed world re-identification task. Thanks to an analogy with the open world case, we propose to artificially enhance the collaboration in the usual sparse coding approach so that performances are further improved for the open world re-identification case. Collaborative sparse coding is used as ranking and a detection tool which explains its good performances for the open world task. This approach can also easily be used for the person verification task.

- **Chapter 5** offers an improvement of the approach presented in Chapter 4. Given that the collaborative sparse coding approach presented in Chapter 4 is not symmetric for probe and gallery identities and based on the intuitive idea that being similar is a reciprocal relation, Chapter 5 proposes a bidirectional collaboration enhanced sparse representation approach for the closed and open world re-identification and verification tasks. The open world results are particularly impressive.
- **Chapter 6** summarizes our contributions to the person re-identification field and recalls our main results. It also offers some ideas for future work.
- **Chapter 7** summaries our work in french.

1.5 List of publications

Part of the work presented in this thesis has already been published:

- [14] Solène Chan-Lang, Quoc Cuong Pham, and Catherine Achard. Bidirectional sparse representations for multi-shot person re-identification. In Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on, pages 263–270. IEEE, 2016.
- [15] Solène Chan-Lang, Quoc Cuong Pham, and Catherine Achard. Closed and open-world person re-identification and verification. In Digital Image Computing Techniques and Applications (DICTA), 2017 International Conference on. IEEE, 2017.

We are currently writing an article for a journal submission which covers the work that has not been published in a conference yet.

Chapter 2

Related work

2.1 Introduction

Many people around the globe work on person re-identification and a commonly adopted classification of re-identification work divide them into three groups: the feature design group, the metric learning group and the neural network group. In this short introduction of the related work, we will explain where this division comes from, and why we choose to present the related work with several more categories than the usual classification.

Two mains steps are involved in the person re-identification task. The **feature design step** extracts from images, useful information about the person it represents. The **matching step** compares probe and gallery images descriptions, and outputs a score value used for ranking gallery identities from the most likely probe match to the least likely one. This score value can be a dissimilarity score (for example a distance value) or a similarity score (for example it can be the probability that the pair of images comes from the same person).

In early methods [16, 17], these two steps formed the two main **independent** steps in the person re-identification pipeline (cf Figure 2.1). Both steps were **unsupervised**, the feature design were hand-crafted based on human intuitions, while the matching step was based on simple distances such as Euclidean or cosine distances or weighted distances.

Gradually **supervised** methods involving a training phase with training data appeared for each step (cf Figure 2.2). The feature design step becomes further divided into feature extraction and feature transformation [18, 19]. The usual distances used for the matching step such as the Euclidean distance or the Bhattacharyya for histograms are replaced by more sophisticated metrics [20, 21, 22, 23] or other classification tools learnt on the training set (such as SVM in [24, 25] for example). Some methods [26, 27, 28, 29, 30, 31, 32, 33, 34, 35] propose to add a re-ranking step at the end of the pipeline. This multiplication of intermediate steps makes the frontier between the description step and the matching step become blurrier and some methods, such as Mahalanobis metric learning methods, can be seen as dis-

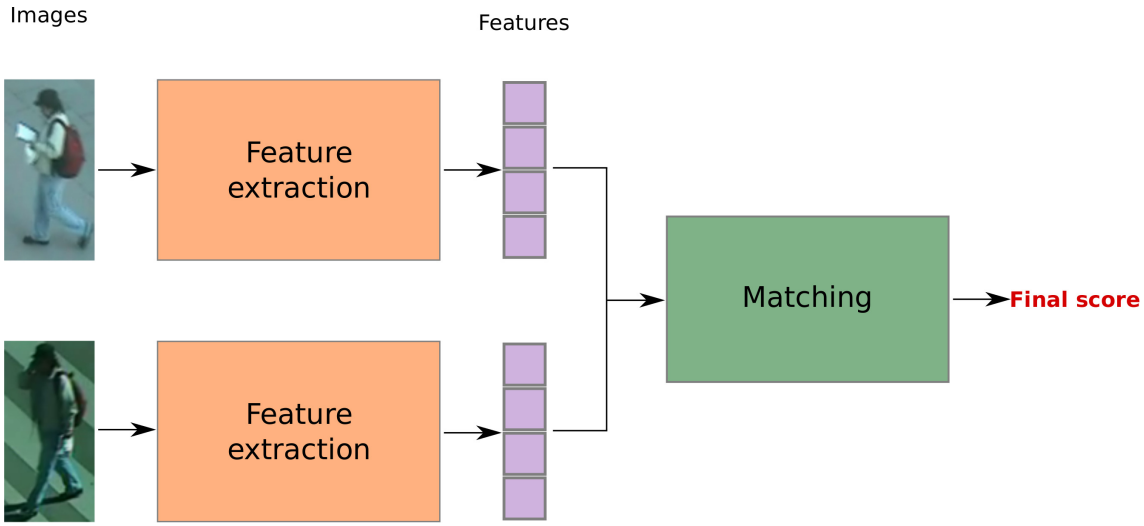


Figure 2.1 – The two main steps in person re-identification frameworks.

tance learning methods, but could also be interpreted as a feature transformation step.

Transfer learning approaches [36, 37, 38] involve even more supplementary training steps in order to transfer knowledge acquired in another domain, such as multi-class classification, clothing classification, etc., into useful information for the person re-identification task.

At the complete opposite, another trend consists in developing methods that **jointly learn** features and similarity function, so that the similarity function is adapted to the feature (cf Figure 2.3). Most of those methods are deep learning approaches [12, 39, 40, 41, 42, 43], but deep learning approaches are sometimes also used only for the feature design step [44, 45].

In the end, even if the three categories (feature design approaches, metric learning approaches and deep learning approaches) do not exactly correspond to person re-identification step, this division is often adopted because a large part of existing person re-identification approaches fall into one the three categories. The small part of methods which do not fit this classification are often only mentioned by closely related methods. In our case, the methods we present in our thesis are related to sparse representations and re-ranking approaches which are not wide spread approaches in the person re-identification field but which we must mention so as to differentiate our method from theirs.

Furthermore, the main research axis of this thesis is the open world case and most existing methods were developed for the closed world case. The few papers [1, 2, 3, 4, 13, 46, 47, 48] which evoke an open world task actually do not all tackle the same problem. This is why a whole section is dedicated to these papers so as to better highlight the difference in scenarios and methods.

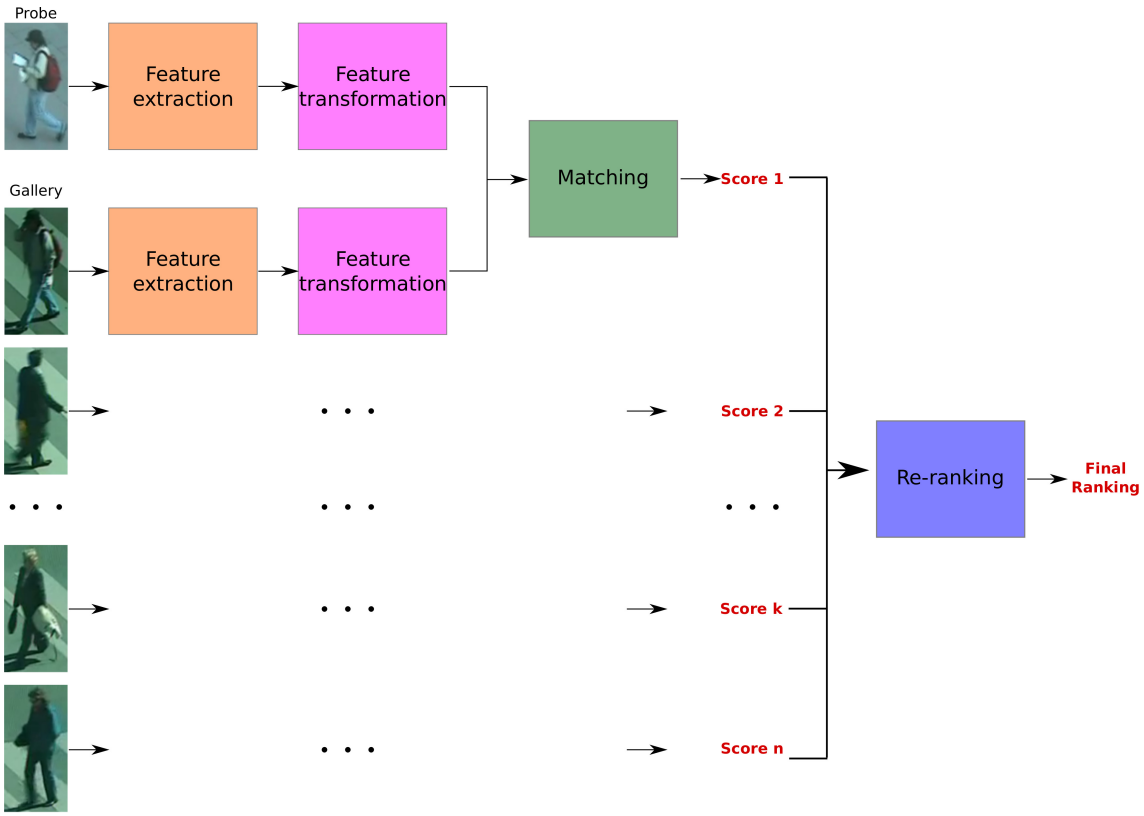


Figure 2.2 – Multiplication of intermediate steps in person re-identification frameworks.

When scenarios differ, it is natural that the evaluation also differs. For a given scenario, there can also be several evaluation measures which assess different aspects of a method. What is evaluated at the end highly influences the design of a method, therefore evaluation measures play an important role in the development of re-identification methods and deserve a separate section.

In this chapter, we first introduce the publicly available person re-identification datasets and look at their evolution over time. Then we give an overview of closed world methods, by dividing it into five parts. The first two parts correspond to the main steps of person re-identification, the representation learning methods and the metric learning methods. The third part is devoted to deep learning methods that are booming these few years. The fourth and fifth parts deal with less often mentioned approaches which however are closely related to our work, namely sparse representation methods and re-ranking methods. The presentation of closed world methods is followed by that of open world approaches in which each type of open world scenario is presented along with the methods specifically designed for tackling it. This chapter is concluded by a description of the different evaluation metrics that are employed for evaluating closed and open world re-identification approaches.

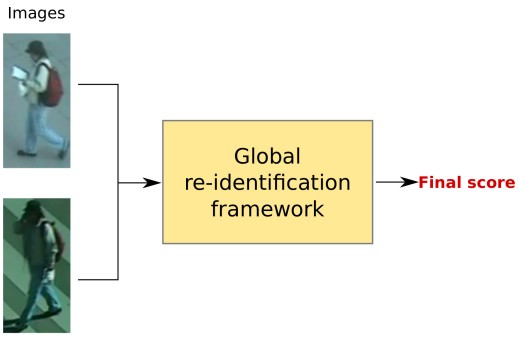


Figure 2.3 – One single global step for person re-identification

2.2 Datasets

From the first person re-identification dataset VIPeR [6] released in 2007, to the last released dataset as of today, DUKE [49] released in 2017, there exist around twenty publicly available person re-identification datasets. Half of them were released between 2014 and 2017, which clearly shows the growing interest in this field. The person re-identification datasets are becoming larger and larger, in terms of number of images, identities and cameras. Each new dataset has been created to overcome some of the shortcomings of previously existing datasets, but sometimes other aspects are overlooked, so each dataset has its own set of advantages and flaws. The main information about each dataset is summarized in table 2.1. This evolution of the datasets reflects the evolution of what is perceived as being the most useful additional data needed for improving re-identification methods and what would make the re-identification scenarios more realistic. This section presents the general trends in the evolution of person re-identification datasets. Of course these are general trends and while overcoming the shortcoming of previously existing datasets, new datasets sometimes are less comprehensive for other criteria.

More images per identity

A few datasets are single-shot datasets, they contain at most one image per identity per camera. VIPeR [6] is the most popular one. Multi-shot datasets, with several images per identity per camera, allow for more varied information. They can be divided into four groups. Some datasets contain only a few images per identity per camera, and those images are a few consecutive images from a single tracklet. It is the case of CUHK campus [53], CUHK02 [54] and CUHK03 [12] with 2 to 5 images per identity per camera. Other datasets such as PRID2011 [11] and iLIDS-VID [9], contain sequences of images, where a sequence of image can be made of up to several hundred images. However for a given identity and a given camera, these images still come from only one single tracklet so the variation is limited. Conversely, some datasets such as Caviar4Reid [17], only present a few images for each identity, but the images representing one person are visually not so similar, often with huge pose or illumination variation. Finally, most recent datasets, such as Market1501 [10], PRW [60], contain much more images per identity, often extracted

Dataset	Year	#Ids Common	#Distractors and Imposters	#Cams	#Ims	Multi shots	Long tracklets sequences
VIPeR [6]	2007	632	0	2	1264	✗	✗
ETHZ [7]	2007	146	0	1	8580	✓	✓
QMUL iLIDS [50]	2009	119	0	2	476	✓	✗
GRID [8]	2009	250	775	8	1275	✗	✗
3DPeS [51]	2011	192	0	8	1011	✓	✗
PRID2011 [11]	2011	200	185 + 549	2	94987	✓	✓
Caviar4ReID [17]	2011	50	22	2	1220	✓	✗
SAIVT-softbio [52]	2012	152	0	8	64472	✓	✓
CUHK01 [53]	2012	971	0	2	3884	✓	✗
CUHK02 [54]	2013	1816	0	5 pairs	7264	✓	✗
CUHK03 [12]	2014	1467	0	5 pairs	14097	✓	✗
OPERID [13]	2014	200	0	6	7413	✓	✗
HDA+ [55, 56]	2014	33	20	13	2976	✓	✓
RAiD [57]	2014	43	0	4	6920	✓	✗
iLIDS-VID [9]	2014	300	0	2	42459	✓	✓
Shinpuhkan [5]	2014	24	0	16	22504	✓	✓
Market1501 [10]	2015	1501	0	6	32 217	✓	✗
Airport [58]	2015	1382	8269	6	39902	✓	✗
SPRD [59]	2016	37	0	24	9619	✓	✓
PRW [60]	2016	932		6	34 304	✓	✗
MARS [61]	2016	1261	0	6	1 191 003	✓	✗
Duke [49, 62]	2017	1852	439	8	46261	✓	✗

Table 2.1 – Person re-identification datasets presentation.

from several tracklets, with eventually several images from the same tracklet which are not consecutive ones. In the Shinpuhkan dataset [5], there are even several full tracklets for each person and each camera.

More identities

After VIPeR [6] which contains 632 identities, many later released datasets emphasized more in having more images per identity rather than in capturing more identities. For example iLIDS-VID [9] is composed of 300 distinct identities, and Shinpuhkan [5] has only 24 identities. However, the trend now is to create datasets with more identities. Several datasets now contain around a thousand identities or more [53, 54, 12, 10, 61, 60, 49]. Besides datasets that simply contains more common identities (identities that appear all cameras), which is the case of CUHK datasets [53, 54, 12], more and more datasets [10, 49] include images from people that appear only in one camera view. These additional identities which only appear in one of the camera views when several camera views are available can be used as distractors (identities present in the gallery set but not in the probe set) if they are put in the gallery set or as imposter probe identities (identities present in the probe set but not in the gallery set).

More cameras

In the early datasets, apart from ETHZ [7] in which every person appears only in one camera, people were captured by 2 cameras. Then multi-cameras datasets appeared. At first, in datasets such as CUHK person re-identification datasets [53, 54, 12], the images were captured by several camera pairs, but each identity appears only in one camera pair. Only most recent datasets are really multiple cameras datasets. In Shinpuhkan [5], each identity appears in all camera. In DUKE [49] and Market1501 [10], the images are captured by several cameras and the camera subset in which a person appears is not the same of everyone.

More challenging bounding boxes

The image boxes have also evolved. In VIPeR [6], images are all resized to 48×128 pixels, and people are in the center of the image. However, with the growing size of datasets, it becomes very time consuming to manually extract people's bounding boxes and label them. Many datasets now rely on person detection and tracking algorithms to automatically provide boxes which size varies and in which people are not necessarily well centered. Some datasets even return boxes that do not contain any person and those images are included in the dataset.

2.3 Closed world approaches

In this section, we present the main approaches developed for the closed world person re-identification task. In addition to the usual description learning, metric learning and deep learning methods, we also devote a subsection to sparse representation methods and another section to re-ranking approaches which are closely related to our work.

2.3.1 Representation learning

This subsection covers approaches which mainly focus on designing good descriptors. It ranges from hand-crafted low level features to mid-level feature learning, without forgetting methods that start with simple features and encompass them into more sophisticated frameworks.

Hand-crafted features

Many methods [16, 17, 63, 64, 65, 66, 67, 68, 23], mainly early ones, propose hand-crafted features. The idea is to extract from the images what we humans consider as important clues to re-identify people. When we describe someone, we talk his about his height, the color of his skin, his hairstyle and the clothes he wears. Using images, hand-crafted features essentially aim at retaining accurate color and texture information. Information extracted from the center of the image should describe clothes information. Skin and hair information is extracted from the top of the image in the zone where the head appears. Very early methods [16, 17] presented part-based features which often required some pre-processing of the images

(segmentation, background subtraction, etc) in order to extract information only from the part of the images which were found to belong to the person represented in the image. A little bit more recent papers [63, 64, 65, 66, 67, 68, 23] often put less emphasis on accurate segmentation and alignment, because these issues are often tackled during the matching step. Instead information is extracted in a more systematic way from horizontal slides or rectangular patches.

Several early methods present features based on rough body parts. In SDALF [16], a segmentation and background subtraction pre-processing step on the images is computed to extract the shape of the represented people. Then using the symmetry and asymmetry properties of the human body, the image is divided into two parts corresponding to the torso and the legs. For each part, color and texture information are extracted via an HSV histogram, the MSCR feature (Maximally Stable Color Region) and the RHSP feature (Recurrent Highly Structured Patches). The matching between two signatures is done by a weighted linear combination of the distances of each of the features, where each feature is associated to a specific distance. In Custom Pictorial Structures [17], six body parts (chest, head, left thig, right thig, left leg, right leg) are localized by maximizing the posterior probability of having that body configuration given the image. For each part, color features (Hue and Saturation histogram and Brightness histogram) are extracted. Each histogram is weighted depending on the part it is extracted from so as to give more importance to bigger parts. They are then concatenated into a single vector and normalized. To supplement body part features, MSCR blobs are also extracted from the body mask. Similarly to [16], the matching is done based on a weighted linear combination of the distances of each feature.

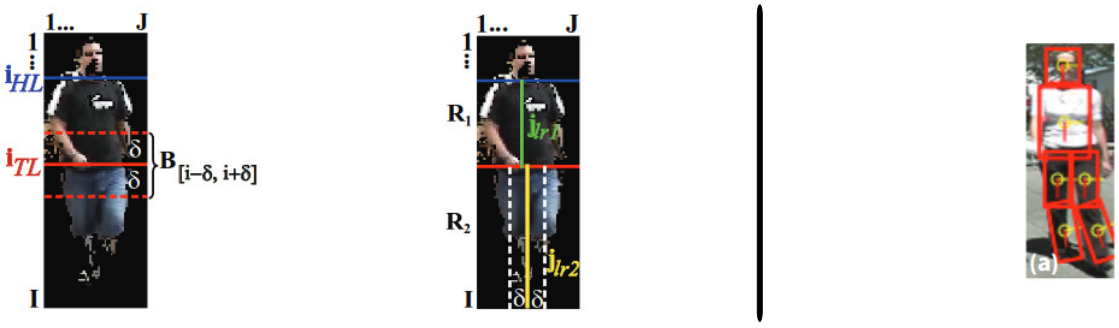


Figure 2.4 – Body parts used in SDALF [16] (left) and Pictorial Structures [17] (right) for hand-crafted feature extraction.

Extracting information from specific body parts seems natural but body parts are not often accurately delimitated, therefore some methods [63, 64, 65, 66, 67] prefer to extract features from simpler splitting of the image such as rectangular patches, where each patch has the same size [64, 65, 66, 67] or with different patch sizes [63]. Similarly to SDALF [16] and PS [17], in those methods, an image is not described by a single vector but by a set of descriptors. Each patch has a descriptor and comparing two images consists in comparing its patches. There is no direct

correspondence between the patches and the body parts, so there can be misalignment and scale issues. Different strategies are adopted. In [63], SCR use covariance matrices as features for their robustness to illumination and rotation change. These features are extracted from rectangular patches of different sizes in order to capture useful information at different scales. In [64, 65, 66, 67], the set of patch descriptors used for describing an image is a mean to deal with misalignment issues by allowing the matching of patches that are localized in the same horizontal stripe. In addition, a learning phase enables to learn the saliency of patches based on the probability of occurrence of similar patches and assuming that the less probable a patch is, the more salient it is. The similarity of two images is computed based on the matching score of the two images patches and on the saliency of those patches.

In all the hand-crafted features methods mentioned so far, the use of sets of local features extracted from body parts or from rectangular patches for describing an image enables to have local information, but these descriptors do not correspond accurately to body parts. Moreover, the way they are used for matching pairs of images not only does not completely solve the misalignment issues and but also does not take into account more global information such as the relationship between the patches of each image (because the global score is a linear combination of the matching scores of each descriptor) . Recent hand-crafted features [68, 23] do not rely on segmentation preprocessing step anymore for obtaining features corresponding exactly to body parts, and instead of using a set of local features for describing an image, local features are concatenated into one vector which acts as a global descriptor for an image. Feature are global but contain local information. In [68], the ISR feature is a concatenation of HOG, HSV and RGB histograms extracted from 15 overlapping horizontal stripes. The color histograms take into account the position of the pixels in the image: every pixel's bin's participation in the histogram is weighted by an Epanechnikov kernel. The reason for using the Epanechnikov kernel to weight the participation of a pixel's bin is the same as the reason for using a mask or a pre-segmentation of the human body in the image: extract information mostly from informative pixels coming from the person and not from the background. In [23], one of the state-of-the art features, LOMO (Local Maximal Occurrence Feature) is presented. Images are preprocessed by the Retinex algorithm [69] that considers human color perception to produce images with vivid colors and better details in shadowed regions. HSV color histogram and SILTP (an improved operator over LBP) are extracted from overlapping patches at two different scales. For each scale, each horizontal position, and each bin, the maximum value over all the patches is taken. In that way, local information at different scales are extracted. Features presented in [68, 23] have both been proposed together with a specific matching step (sparse representation for [68] cf section 2.3.4 and metric learning for [23] cf section 2.3.3) but these features can be part of other re-identification frameworks.

Features selection and weighting

Instead of designing features by hand and relying on intuitions for the choice of color and texture descriptors which besides might present redundancies, some methods

propose to use a whole set of classic color and texture features and learn which ones are important through a learning phase so as to combine them in an appropriate way.

In [18], the Ensemble of Localized Features (ELF) method extracts eight color channels and nineteen texture channels from the images. For each feature, a binary classifier is learnt to output whether two images are from the same person or from distinct people. Each binary classifier is considered as a weak classifier. Adaboost is used to learn a robust classifier defined by a weighted linear combination of the weak classifiers. In [19], there is a similar idea of weighting differently different features, but the framework is quite different. Clustering forests are used to group similar people into prototypes. Classification forests are then employed to learn a different weight vector for each prototype. The distance of a probe image to a gallery image is computed using the weights of the prototype the probe person most likely belongs to.

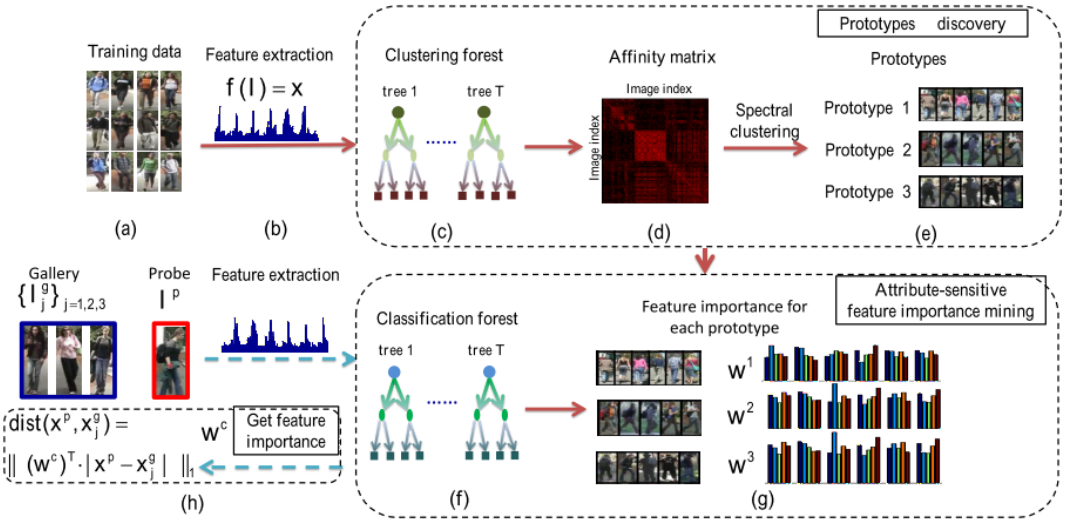


Figure 2.5 – Illustration of the framework presented in [19] for learning prototypes and weighting differently the features for different prototypes.

Adding some semantic information

In addition to usual low level features, some methods propose to consider higher level information such as semantic colors (linguistic labels we use to describe colors) or even attributes (presence or absence of an accessory, type of clothes, hair, gender, etc).

In [70], along with other descriptors (color histograms, texture features, covariance descriptors), color name descriptors are extracted from each of the 6 horizontal stripes in which an image is divided. The color name descriptor is an 11 dimensional vector, coding for a region, the probability distribution of colors on 11 basic color names. Each type of descriptor has its corresponding similarity measure (Bhat-

tacharyya distance for histograms, solution of a generalized eigenvector problem for covariance matrix, etc), and the similarity measure of each descriptor in each of the 6 regions form a weak ranker. RankBoost is used to sequentially add and weight a weak ranker to finally form a strong ranker which is a linear combination of weak rankers.

In [71], a 16 dimension salient color names descriptor (SCNCF) is proposed. This descriptor is designed to be robust to illumination variation, because not only do similar colors have similar color description, but multiple similar colors also share the exact same color description. This salient color name descriptor is extracted from 6 horizontal stripes and concatenated. It is not meant to be used alone, but in addition to usual color descriptors. In that paper, the KISSME metric learning is used for the final person matching step.

In [72, 73], mid-level attributes such as the presence of shorts, skirt, backpack, etc are used. For each of these attributes, an SVM is trained to output the probability of presence of the attribute. The vector of probability of presence of all the attributes is then fused to other usual features for the final matching step. This approach requires to have a annotated datasets with both identities and attributes labels.

To overcome this labelling issue, the paper [38] proposes a transfer learning approach to transfer knowledge learnt from datasets labelled with clothing attributes (fashion photography datasets) to the person re-identification task. Semantic attribute classifiers are learnt at patch level, using SVM once again.

Instead of semantic attributes that require labelled datasets, be it from a re-identification dataset or from other fields datasets, another type of approaches that promotes the use of mid-level attributes rely on latent attributes. In [74], ARLTM (Attribute Restricted Latent Topic Model), inspired by document topic search, uses the probabilistic model LDA (Latent Dirichlet Allocation) to re-identify people by learning latent attributes. A study of the correlation between attribute and semantic labels is integrated to better exploit related attributes such as "long hair", "skirt", "female".

Time space features

In previously mentioned approaches, features were computed per image. For datasets that present full tracklets, a few papers [9, 47] propose to extract time-space features which capture people's appearance but also motion and gait information.

Both DVR [9] and MDTS-DTWt [47] propose to perform person re-identification by video matching, ie. by matching video sequences rather than images. The videos sequences are not required to be previously aligned. In [9], based on the walking cycle that is periodic, several fragments of fixed length (21 frames) are extracted from the full sequence. This is illustrated in Figure 2.6. Each fragment is encoded by one single HOG3D descriptor. An SVM is learnt to separate positive and negative pairs, based on pairs' absolute difference. The similarity associated to a pair of

sequence is given by the maximum similarity between all pairs of selected fragments from the two sequences. In [47], instead of first selecting different fragments of a video sequence, video sequences are directly matched with a modified Dynamic Time Warping algorithm, which jointly selects and aligns fragments from both video sequences.

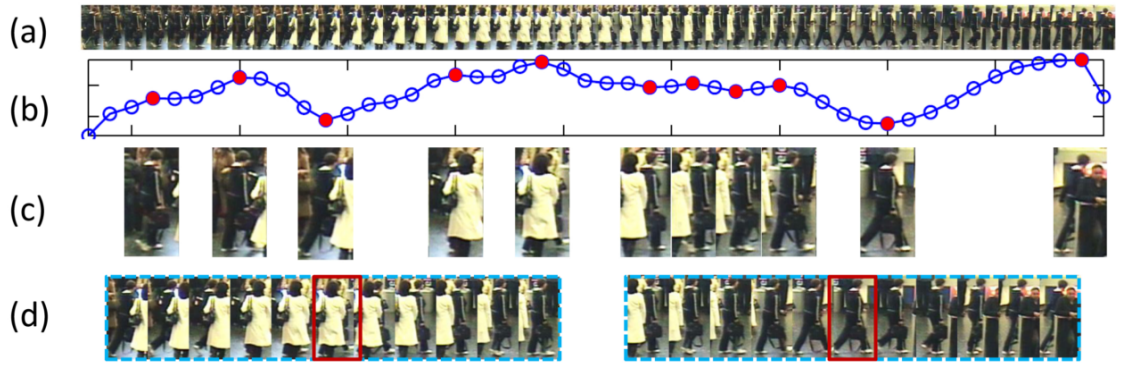


Figure 2.6 – Illustration of the DVR framework presented in [9].
The figure and the legend are directly taken from the paper [9] and they explain how the walking cycles are exploited. (a) A person sequence of 50 frames is shown, with the motion intensity of each frame shown in (b). The red dots in (b) denote automatically detected local minima and maxima temporal landmarks in the motion intensity profile, of which the corresponding frames are shown in (c). (d) Two example video fragments (shown every 2 frames) with the landmark highlighted by red bounding boxes.

2.3.2 Metric Learning

This section is dedicated to metric learning methods which aim at solving the matching step by replacing usual distances by metrics adapted to the features and the person re-identification problem.

In mathematics, a synonym of metric is distance function. It refers to a function that defines a distance between pairs of elements of a set. One of the most well-known metric is the Euclidean distance, but many other metrics exist, they must verify a small number of constraints.

A function d is a metric on the set E if it is a function

$$f : E \times E \to \mathbb{R}^+ \tag{2.1}$$

that satisfies:

$$\begin{array}{lll} 1. & d(x,y) = 0 \iff x = y & \text{identity of indiscernibles} \\ 2. & d(x,y) = d(y,x) & \text{symmetry} \\ 3. & d(x,z) \leq d(x,y) + d(y,z) & \text{subadditivity or triangle inequality} \end{array} \tag{2.2}$$

Metric learning is the task of learning a metric, ie. the task of learning a function that satisfy the conditions just described.

A popular metric is the Mahalanobis metric. A Mahalanobis distance is characterized by a matrix M that must be symmetric positive definite . The Mahalanobis distance characterized by the matrix M between two column vectors $x, y \in \mathbb{R}^d$ is given by:

$$d_M(x, y) = (x - y)^T M(x - y). \tag{2.3}$$

The Euclidean distance is a special case of Mahalanobis metric, where the matrix M is the identity matrix I:

$$d_I(x, y) = (x - y)^T I(x - y) = (x - y)^T (x - y) = ||x - y||_2^2 \tag{2.4}$$

where $||.||_2$ is the $\mathcal{L}2$-norm.

The metric learning step is often associated to the matching step of person re-identification because it replaces usual distances by metrics learnt on training data. However, a simple reformulation of Mahalanobis distance makes it evident that using Mahalanobis metric as a new distance function actually amounts to applying a linear transformation to the initial features and then use the Euclidean distance for matching. Indeed since M is symmetric positive definite, there exist a real value matrix L such that :

$$M = L^T L. \tag{2.5}$$

Therefore the equation 2.4 can be reformulated as:

$$\begin{aligned} d_M(x, y) &= (x - y)^T M(x - y) \\ &= (x - y)^T L^T L(x - y) \\ &= ||L(x - y)||_F^2 \end{aligned} \tag{2.6}$$

where $||.||_F$ is the Frobenius norm. Thus learning a Mahalanobis distance characterized by a matrix $M = L^T L$ can also be considered as a learning a linear transformation of the features characterized by the matrix L.

Most metric learning algorithms [75, 76, 21, 23] proposed for person re-identification are Mahalanobis metrics but not all. They rely on the optimization of different objective functions using training data. Some methods [77, 78, 76] impose constraints per probe person with relative ranking constraints on the pairs of images the probe person appears in. Other methods [79, 20, 21, 23, 80] adopt a more global reasoning in terms of positive pairs and negative pairs, without specifically distinguishing the constraints for each probe person. Yet other approaches [81, 82] propose to combine several metrics instead of using a single one. Metrics initially developed for other applications might also be relevant for person re-identification.

Relative ranking for each identity

Quite a few metric learning methods are based on relative ranking constraints on each training identity [75, 77, 78, 83]. Those approaches impose constraints on the relation between positive and negative pairs of each identity independently of other identities. While some of these approaches also enforce positive pairs distances to

be small or negative pairs distances to be large, no fixed threshold are imposed, or solely on negative pairs.

In LMNN [75] and TopPush [83], the constraints are directly on the distances. The main idea is to encourage positive pairs distances to be as small as possible and for each probe image, the distance of a positive pair it appears in is enforced to be smaller by a margin than the distance of a negative pair it appears in. In LMNN [75], only the most difficult negative pairs are taken into account during the optimization process. For each probe image, only mismatch gallery images that are among the probe image's k nearest neighbors and whose distance to the probe image is smaller than the positive pairs distance are considered. In TopPush [83], features are computed for video-sequences. For each probe identity, each relevant positive pair's distance should be smaller than that of all relevant negative pairs, ie. it should be smaller than the smallest distance of all related negative pairs. A probabilistic approach is adopted in [77]. Instead of direct penalization on the distances, it is probabilistic values that are considered. The PRDC method [77] aims at maximizing for each training person the probability that the distance of a positive pairs is smaller than that of a negative pair. The WARCA [78] method focuses on the ranking rather than on the distances themselves. A Mahalanobis metric is learnt by minimizing the ranks at which the right matches are found. A drop in the top ranks is more heavily penalized than a drop in further ranks. In EIML [76], there is no direct relative distance constraints. Instead, the ratio between the distance of the negative and the positive pair of a triplet is used to weight the importance of the penalization on the negative pair. The metric is learnt to minimize the distance of positive pairs and maximize that of negative pairs. The smaller the ratio distance negative pair over distance positive pair, the more penalized the negative pair is.

Minimizing positive pairs distances, maximizing negative pairs distances

Other metric learning methods formulate constraints on the positive pairs and on the negative pairs but without explicit constraints on the relations between positive and negative pairs. The intent is to minimize the intra-class variances while maximizing the extra-class variances where each class corresponds to a training identity.

In RPLM [79], the optimization of the objective function leads to globally minimizing the distance of positive pairs and maximizing that of negative pairs. In PCCA [20] and MLAPG [84], a log-logistic loss function is applied to the Mahalanobis distance of pairs of features so that positive pairs are penalized if their distance is bigger than a predefined threshold, while negative pairs are penalized when their distance is smaller than the same threshold. The LADF [22] approach on the other hand does not use a predefined threshold, but jointly learns a metric and a local threshold. In KISSME [21], using the assumption that both negative and positive pairs distribution follow gaussian distributions with zero mean, the learnt metric minimizes intra-class variance and maximize extra-class variances. XQDA [23] combines the KISSME metric with a dimension reduction method. The LFDA method proposed in [80], combined Fisher Discriminant Analysis and Locality Preserving Projection,

to minimize intra-class variance and maximize extra-class variance, while still allowing for multi-modal distributions. Following the reverse direction, in DNS [85], not only intra-class variances are reduced but the learnt metric squarely projects images of the same identity into one single point.

Combination of several metrics

Some methods consider that one metric is not good enough and propose to combine several metrics in order to obtain more robust matching.

In [81], MiMl-DML-IR uses 8 features and learns one metric for each of them. The final distance of a pair is the sum of its 8 Mahalanobis distances. This approach measures the confidence of an image extracted with a tracking algorithm being an imposter and integrates this information in the metric learning process. In [82], MtMCML deals with the person re-identification task in a camera network context, where there are more than two cameras. It is a multi-task metric learning method, where several metrics are learnt jointly. There is one metric for each camera and one for each pair of cameras.

Metric learning, kernelization, transfer metric

In addition to the metric learning methods proposed specifically for the person re-identification task, many other metric learning methods and dimension reduction methods exist and can also be used for re-identification. To cite only a few, we could use LDA (linear discriminant analysis) [86], LDML (Logistic Discriminant Metric Learning) [87] or ITML (Information theoretic metric learning) [88].

Moreover, many metric learning methods can be kernelized. The paper [89] proposes several kernel variants of existing person re-identification metric learning methods.

In [90], instead of simply learning a generic metric during the training phase and use it as it at test time, an online transfer learning phase is proposed to adapt the generic metric to candidate-set-specific metric.

2.3.3 Neural networks

In the last few years, we have witnessed an explosion of the number of person re-identification approaches based on neural networks. Long after the beginning of neural networks in the late 1950's with the perceptron, it is only from 2012 on, after a deep learning method won the ImageNet Large Scale Visual Recognition Challenge [91], that deep neural networks approaches became popular and started to be used for a large variety of domains. The better designed backpropagation strategies and the increasing size of datasets have played a large role in this renewed interest in neural networks. It is however only two years later, in 2014, that the first deep neural network was presented for the person re-identification [12]. The delay in the

apparition of deep neural network approaches for the person re-identification task is connected to the size of re-identification datasets that were previously too small, making it impossible for neural networks to generalize well to unseen data. But since then, the number of deep neural network person re-identification papers is ever increasing.

A neural network is a group of interconnected nodes where each node applies a non linear function to a linear combination of its inputs. The architecture of a network is designed depending on its application with some fixed choices of non linear function for the nodes. The non linear functions are often the $tanh$ function, sigmoid functions, or relu function (derivable approximation of the function $f(x) = max(x, 0)$). What is learnt during the training phase of a neural network are all the weights corresponding to the linear combinations of nodes input. The deeper the network, the more parameters (weights) there are to learn and the larger the amount of labelled data needed is. The learning process is based on the optimization of a cost function which is defined depending on the application.

Neural network approaches are sometimes seen as complete re-identification frameworks that take raw pairs of images as input and return similarity scores by jointly learning feature extraction and matching. Other times, they are solely used as a way of designing features. In this section, we divide the neural network approaches for person re-identification into six groups. The first group considers the re-identification task as a binary classification task and focuses on the conception of the different layers of the network. The second and third groups are interesting due to their choices of loss function which is either based on ranking constraints only or on a combination of different types of constraints. The last four groups emphasize on the information that is captured by the features learnt by the network. For that purpose, some networks rely on their recurrent architecture, others rely on the learning process to integrate information regarding the presence or absence of some attributes or the description of space-time information.

We will use the following abbreviations: NN for Neural Network, DNN for Deep Neural Network (neural networks with many layers of neurones) and CNN for Convolutional Neural Network (neural networks with convolutional layers).

NN solving a binary classification problem for feature extraction and matching

The methods adopted in [12, 39, 40, 41, 42, 43] use training data to learn a network for feature extraction and matching. The re-identification task is cast as a binary classification problem. During training and testing phases, the network is used in the same way. It is given a pair of images as input, and it outputs a score value that represents the probability that the images comes from one single person or from two distinct people.

FPNN [12] is one of the first neural network framework for person re-identification.

The architecture of the network is designed so that patch matching between the patches from a pair of images is performed by taking into account possible horizontal displacements. The Figure 2.7 shows the architecture of this filter pairing neural network. In [39, 40] a siamese architecture (DML) is presented. Instead of patches, the images are partitioned into three horizontal parts. Each part is treated with a different CNN, but the weights of the CNNs are shared for the two input images. The cost function is based on binomial deviance which enables the network to train more on hard samples than on easy ones. The cosine similarity function is used for matching. In [41], the MBDML network is a combination of the siamese DML architecture [39] used for the person re-identification task and the bilinear CNN architecture [92] used for fine-grained classification. By incorporating bilinearity with outer product of features in DML [39], more spatial information are captured. In [42], in between the first two layers of convolution with max pooling layer, which extract features from each image, and the last fully connected layer, it is the middle layers that capture the cross-input local mid-level and high-level features through a neighborhood difference layer, a patch features summary layer, and an across patch features layer. In [43], the deep embedded metric network makes a parallel between the weights of the fully-connected layer (FC) of the neural network and the coefficients of the matrix characterizing a Mahalanobis metric. The loss function aims at minimizing the distance of positive pairs and encourages distances of negative pairs to be bigger than a predefined threshold, under the constraints that the weights of the FC layer are such that the final Mahalanobis metric does not deviate too much from the Euclidean distance. Furthermore, for learning a metric that does not distort the manifold, an emphasize is put on the choice of positives pairs that should be moderate, ie. not too hard nor too easy.

The approach presented in [93] also takes pairs of images as input and outputs a similarity score. It differs from the previously presented neural network methods in the fact that the network is jointly considered as a network that outputs a similarity score given a pair of images and as a network that provides features for single images. The proposed network is composed of two subnetworks, SIR and CIR, that share the same first part. SIR stands for single-image representation, CIR stands for cross-image representation. The SIR subnetwork computes features from single images such that the Euclidean distance of positive pairs are below the threshold $b_{SIR} - 1$ and the Euclidean distance of negative pairs are above the threshold $b_{SIR} + 1$. Slack variables are added to enlarge the field of possible solutions. The CIR subnetwork computes a feature for a pair of images. The associated loss function ensures that an SVM can perform binary classification on the CIR feature. The final output is the similarity score computed by a weighted sum of the Euclidean distance of the SIR features of each image and of the SVM score on the CIR feature.

NN learnt with ranking constraints

Several neural network approaches rely on ranking constraints on pairs of images. In order to impose such constraints, at least triplet of images need to be considered during the training phase. For testing phase however, a neural network is either

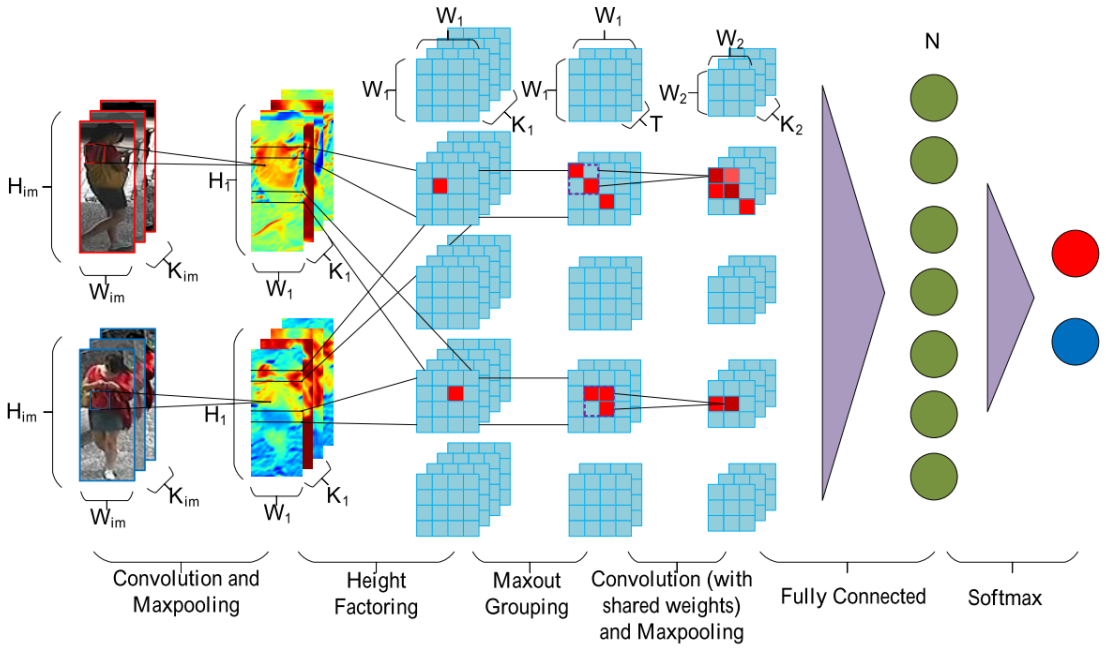


Figure 2.7 – Illustration of the Filter Pairing Neural Network (FPNN) presented in [12].

used for providing features for each image, or for returning a score for the matching of two images. Therefore, the way the network is used at testing phase is a little bit different than at training.

The neural networks architectures presented in [94, 95, 45] are learnt through ranking constraints. In [94], a deep ranking framework made of two consecutive components is proposed. The first component, the deep joint representation learning component takes a pair of images as input, and outputs a similarity score. The second component, the learning to component, takes as input pairs' similarity scores and minimizes for each probe identity, the number of gallery disorders. During test phase, only the first component is exploited, it returns for a pair of images, the associated similarity score. The neural network presented in [95] is learnt to provide image features by keeping the last fully connected layer. During the training phase, relative distance constraints are imposed on triplets. All three images of a triplet are passed through the same network, and the loss function aims at minimizing the Euclidean distance of the features of the positive pair of the triplet while maximizing that of the negative pair of the triplet. The neural network TCP presented in [96], is also used for providing image features that are compared using the Euclidean distance. During the training phase, TCP takes triplets of images as input. Each image of the triplet passes through the same CNN. That CNN is a multi-channel part-based CNN that extracts features from the whole body, and from four body parts through 5 independent channels, and the features extracted from each part are then concatenated into a single feature. The improved triplet loss function is not a simple relative ranking constraint. For a triplet, in addition to constraining the Euclidean distance of the positive pair to be smaller than that of the negative pair

by a margin τ_1, it also constraints the Euclidean distance of the positive pair to be smaller than a predefined threshold $\tau_2 < \tau_1$. The idea is that in addition to relative ranking constraints, the intra-class variance should be smaller than the extra-class variance.

NN with several loss functions

A few neural network approaches aim at producing features adapted for re-identification thanks to their ability to solve several tasks.

In [97], the Multi-Task Deep neural network MTD is trained on triplets of images. A ranking loss function enforces that the distance between the wrong pair is bigger than that of the matching pair by a margin. Furthermore, to ensure that the network is able to solve the binary classification task, a binary logistic regression loss is applied to the pairs (positive and negatives) involved in the triplets. Once learnt, the network is used for feature extraction. An additional transfer learning phase enables to adapt the network weights for use on a different dataset. This is interesting for datasets that are too small for obtaining good generalization using the usual learning process. In [36], the Deep Transfer Learning (DTL) network is trained on two different tasks: an ID classification task and a pairwise verification task. The network itself is composed of three blocks. Starting with an existing network (GoogLeNet is used but other networks could replace it), two blocks are added: an identity classification block and a verification block. The network undergoes two fine-tuning processes, one for each of these two blocks. The identity classification block is associated to the task of being able to relate the training images with their right identity index. The verification block is associated to the task of finding whether two images represent one single person or two distinct people. For testing, the network is truncated to extract features that are compared with the Euclidean distance. Coincidentally, both MTD and DTL approaches rely on two types of loss functions and on transfer learning. However, these two points are completely independent ideas that can be implemented separately. Moreover transfer learning simply refers to the idea of re-using the knowledge acquired in a domain for another domain. The domains can be more or less related. In MTD [97], the network weights are first learnt using a big person re-identification dataset, and this knowledge is transferred to smaller person re-identification datasets through the transfer learning phase. In DTL [36], the knowledge is acquired in an object classification task and it is transferred to the person classification and verification tasks.

NN for attribute feature extraction

In this paragraph, we introduce some approaches [44, 45] where a special importance is given to capturing understandable attribute features, such as the type of clothes worn, the presence of some accessories, the hair length, etc.

The deep attribute learning network SSDAL presented in [44] is trained so that its last layer outputs a -dimensional feature where each value corresponds to a clothing attribute. A semi-supervised attribute learning approach is proposed for

training the network. This approach not only requires a person re-identification dataset with identity labels but also requires an independent dataset with annotated attributes. The learning phase incorporates several steps: a fully supervised training on the attribute dataset, a fine-tuning step on a person re-identification dataset using predicted attributes, and finally a fine-tuning step with both attribute and person re-identification datasets.

In [45], the ResNet-50 pretrained on ImageNet object classification task is fine-tuned on person re-identification datasets annotated with both identities and attribute labels. During the training phase, the network takes one image as input, and predicts people's identity and attributes. There is one output per attribute and one output per training identity, so the final size of the output vector is the sum of the number of attributes and the number of training identities. For the test phase, the network output vector is used as an image descriptor.

Recurrent NN

Recurrent neural networks (RNN) are a type of neural network that can process sequences of information thanks to their internal memory. Most RNN approaches [98, 99] use it for computing space-time features from sequences of consecutive images. Others [100] use it to sequentially treat sequences of information coming from single images.

Both RFAnet [98] and RCPVPR [99] approaches deal with the video-based person re-identification task. The recurrent networks aggregate temporal information of sequences of images and are used at test time as a mean to provide a single feature descriptor given a sequence of images. In the RFAnet approach [98], the network is trained on a multi-class classification task, where each class is a training identity. The weights are optimized for maximizing the log of the probability of finding the right match. At test time, the descriptor computed using the trained recurrent network can either be directly compared with a cosine distance or through a RankSVM model learnt during the training phase. In [99], the RCPVPR network is trained using pairs of sequences of images by optimizing jointly an identification task and a verification task.

In [100], the recurrent neural network is used to sequentially process image regions rather than image sequences. This allows for capturing interactions between features extracted from different regions of an image. It is a siamese network that takes pairs of images as input. It is optimized with the contrastive loss that minimizes the distance of positive pairs, and penalizes the negative pairs that have a distance larger than a margin. After the learning phase, the network is used to provide image features and pairs of images are compared with the Euclidean distance.

NN and cross-dataset re-identification

The person re-identification performances are always evaluated on several datasets, but most of the time each dataset is evaluated separately. For each dataset, a model is learnt on the training set, and then tested on the testing set. Using the model

learnt on a dataset but tested on another dataset often leads to poor results. For deep neural network approaches an additional problem comes on top of it due to the fact that the training phase requires a huge amount of data that is not available for each dataset, especially small ones.

Most methods which tackle this issue propose an additional training steps for adapting the weights already learnt on a larger dataset for the smaller dataset which is the case in [97]. A different approach is adopted in [101] where it is the dropout process that enables to tackle the cross-dataset issue. The network is trained on several datasets and to obtain generic features that are effective on all datasets, some neurons are shared for all datasets while some are used only for a subset of datasets. During training, depending on the dataset, the dropout rate of a neuron varies. The dropout rate of a given neuron depends on its impact on the outputs of the samples of the chosen dataset. At test time, the features are computed using the learnt network in which each neuron is applied a dataset dependent mask.

2.3.4 Sparse representations

Following the success of sparse coding for face verification and recognition, some person re-identification methods [68, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113] also adopted sparse coding in their framework.

Sparse coding consists in representing an element with a sparse vector. Given a column vector $x \in \mathbb{R}^{d\times 1}$ and a dictionary $D \in \mathbb{R}^{d\times n}$, a sparse code of x is a weight vector $\alpha \in \mathbb{R}^{n\times 1}$ such that $x \approx D\alpha$ where α is sparse, ie. α contains only a few non-zero elements. In the end, x is approximated by a linear combination of a few dictionary elements (columns of D) where the weights are the non zero values of α. In order for α to be sparse, the dictionary elements that participate in the reconstruction of x are most often those that are the most similar to the reconstructed x. The sparse code is not unique, there can be many different linear combinations of dictionary elements that are sparse and approximate well x. α is determined by solving an equation of the following form:

$$\alpha* = \arg\min_{\alpha} ||x - D\alpha||_F^2 + \psi(\alpha) \tag{2.7}$$

where ψ is often chosen as the $\mathcal{L}_1$ norm, but other choices of ψ exist.

For person re-identification, sparse representation can be exploited in different ways. In [68, 102, 103, 104], the gallery features form the dictionary, and probe elements are reconstructed by linear combinations of gallery elements. The reconstruction errors are used for ranking gallery identities. Sparse code can also be considered as new features, as in [105, 106, 107, 108, 109] where a dictionary is learnt during the training phase. Using that dictionary, the sparse code of gallery and probe elements are computed and then compared for matching. Apart from these two main categories, some papers propose still other ways of exploiting probe sparse representation [110, 111, 112, 113].

Matching with residual errors

Methods [68, 102, 103, 104] use residual errors for matching gallery and probe identities. A residual error is a reconstruction error where the reconstruction does not take into account all the dictionary elements corresponding to non zero values in the sparse representation, but where only part of the dictionary elements is considered.

What differentiates [68] and [102, 103], is the way the sparse code is computed. In [68], in order to deal with a ranking issue inherent to sparse coding methods, sparse codes are computed by an iterative method. Since a probe element is approximated by a sparse linear combination of gallery elements, most gallery elements do not participate in the approximation, their associated sparse code α_i is a null vector, so their residual error equals $||x||$, the Euclidean norm of the probe element. Therefore it is not possible to those gallery identities because they have the same residual error. In [68], after computing an initial sparse representation of a probe element using the whole dictionary D, another sparse representation is computed using only the dictionary elements that participated in the initial sparse representation, where each dictionary element is penalized differently depending on the value of its weight in the initial sparse representation. The gallery identities that were involved in the initial sparse representation are ranked according to their new residual error. This process is iterated using the overall dictionary D deprived of the dictionaries associated to gallery identities that have already been ranked. None of the other existing sparse representation methods for person re-identification tackle this issue of lack of clear ranking from a certain rank onward. The approaches presented in [102, 103] aim at better exploiting the multi-shot aspect of gallery identities. A group penalization with $\mathcal{L}2$ norm is introduced to obtain sparse codes that are not only sparse in terms of non zero elements, but also in terms of identities involved. Among two sparse codes that would involve the same number of non zero elements, the one spread over $l_1 < l_2$ identities would be preferred over the one which involves l_2 identities. In [104], instead of whole images, SURF points are extracted from each images. For each image, each SURF point is assigned to an identity by considering the identity that has the smallest residual error, then majority decision rule is applied, and the image is labelled with the identity that has the most SURF points vote. In the approaches just presented [68, 102, 103, 104], the dictionaries are not learnt but simply composed of features from the gallery.

Matching with sparse representations as new features

Approaches presented in [105, 106, 107, 109, 108] rely on dictionary learning, and sparse representations are simply new feature descriptions. Training data are used to learn either a single dictionary [105, 106, 107] or a coupled dictionary [109] so that the sparse representation of gallery and probe elements are similar and can be compared for the final matching.

In [105], a dictionary is learnt by minimizing the overall reconstruction error of every probe and gallery training elements and the associated $\mathcal{L}1$ norm of the sparse codes, with the constraint that for each identity, the Euclidean norm of the dif-

ference in sparse code of a positive pair of (probe,gallery) images is smaller than that of a negative pair. During test, the sparse code of probe and gallery elements are computed using the learnt dictionary, and for each probe query, gallery identities are ranked by increasing Euclidean distance of their sparse code with that of the probe query. In [106], in addition to the reconstruction error and the $\mathcal{L}1$ norm of the sparse code, a graph Laplacian regularisation term is introduced. Just as the constraint in [105], the aim of the additional regularization term is to force pairs of probe and gallery images of the same identity to have similar sparse code. The regularization term is based on a cross-view correspondence matrix of training elements, which encodes information about the K-nearest neighbours in terms on cosine distance. This method can exploit labelled as well as unlabelled data. At test time, the sparse code of probe and gallery elements are compared using the cosine distance. In [107], instead of using the $\mathcal{L}1$ norm operator $\psi = ||.||_1$, a robust graph regularization is used. This regularization is designed to further alleviate the effect of outlying samples. During test, probe and gallery sparse codes are computed with $\psi = ||.||_2$ and their cosine distance is used for ranking.

While only one common dictionary was learnt in [105, 105, 106], the approach [108] proposes to learn several dictionaries for tackling the cross-dataset issue. When using several datasets, in addition to a common dictionary used for all datasets, two dictionaries are learnt for each dataset. For tests on a chosen dataset, the sparse codes of gallery and probe elements are computed jointly on the three dictionaries and the cosine distance of the concatenation of the sparse codes on the three dictionaries is used for ranking.

In [109], probe and gallery sparse codes are not computed using the same dictionary. A coupled dictionary is learnt during the training phase. Through a subspace learning phase (Canonical Correlation Analysis), projection matrices for each camera is learnt. Then a gallery dictionary and a probe dictionary are jointly learnt by minimizing for all training examples, their reconstruction error and the weighted $\mathcal{L}1$ norm and $\mathcal{L}2$ norm of their sparse code. At test time, the sparse codes of probe elements are computed using the probe dictionary while the sparse codes of gallery elements are computed using the gallery dictionary. A modified cosine similarity is used as a measure of the similarity of gallery and probe elements sparse codes.

Sparse representation, an intermediate step for other methods

This paragraph introduces a few other person re-identification approaches which use sparse representations but neither for ranking with the residual errors, nor as new features.

In [110], like in [109], coupled probe and gallery dictionaries are learnt for three body parts (head,body, legs). The method is semi-supervised and can make use of unlabelled data. At test time, the sparse codes of probe elements are computed using the probe dictionary. The obtained sparse codes are used to recover new features using the gallery dictionary. The matching is based on the sum for each body part

of the Euclidean distance of the new features with the existing gallery features.

In [111], instead of a coupled probe and gallery dictionary, it is a coupled low and high resolution patch dictionary that is learnt, as well as a mapping matrix to convert features corresponding to low-resolution images into features that could be extracted from high-resolution images. In the person re-identification scenario considered, it is assumed that gallery images, which correspond to known identities, are high resolutions images. The query images on the other hands are in low resolution. Similarly to [110], the sparse codes of a probe image patches obtained through optimization with the low-resolution dictionary are used to compute new high resolution features thanks to the high resolution dictionary and the mapping matrix. Gallery image patches are approximated by a sparse linear combination of elements of the high resolution dictionary, and a gallery image new feature is the concatenation of all its patches reconstructions. For a given probe image, the gallery images are ranked based on the distance of their new high resolution features.

In [112], the proposed semi-supervised coupled dictionary method is used to obtain sample specific SVM. For each training probe identity, an SVM is learnt to separate positive pairs and negative pairs. Then a coupled dictionary composed of a feature dictionary and an SVM weights dictionary are learnt jointly with a mapping between the two dictionaries. At test time, from the coding of the probe image using the feature dictionary, the sample specific SVM weights is recover and used for determining the matching score of that probe image with each of the gallery images.

In [113], sparse codes of probe elements are computed using the gallery as the dictionary. But instead of using the residual errors for ranking the gallery identities, they are simply used as a mean to assess the reliability of a matching.

2.3.5 Re-ranking methods

Most of the methods presented so far are what some call train-once-and-deploy scheme which involve a training phase in which a model is learnt and considered final, ie. it is not modified anymore afterwards and is directly applied in the testing phase. However training datasets are yet still not big and varied enough to be representative of images extracted from any surveillance camera. Therefore, methods do not generalize so well on new data. Some researchers propose to update the model online thanks to user feedback [114, 115] to improve the ranking performances. Other papers [26, 27, 28, 29, 30, 31, 32, 33, 34, 35] tackle it in an automatic way and propose re-ranking approaches that do not require human interaction.

Human feedback for ranking and re-ranking

Person re-identification algorithms are far from being perfect, and in the end there is always a human who checks the returned list of identities that are supposed to

be ranked from the most similar individuals to the least similar ones. However, in practice, the top ranks gallery identities often contain identities that we would not consider as similar. The approaches described in [114, 35] make use of this kind of information that can easily be given by the final user so as to improve the ranking algorithm in an online manner.

In [114], a Post-rank OPtimization (POP) method is proposed to refine the ranking using user feedback. For each probe instance, the user selects an image which is visually dissimilar to the probe instance but which is in the top of the ranked list of gallery identities (strong negative image). Optionally he can also select images that are visually similar to the probe instance but which do not represent the right person (weak negative images). An affinity graph describing the similarity between all gallery instances and synthesized positive pairs is computed. The negative pairs information obtained by user feedback is propagated from the labelled elements to the unlabelled ones thanks to the post rank function. The post rank function is learnt through a Lapalacian SVM optimization problem which besides encouraging positive pairs to have a similarity score bigger than 1 and negative pairs to have a similarity score smaller than -1 also takes into account the affinity graph by enforcing the post-rank function to take close values when the inputs are close gallery instances.

Contrary to POP [114] where the final matching score is a weighted linear combination of the initial ranking score and the refined ranking, in [115], the online incremental learning method proposed for human in the loop person re-identification does not require pre-labelled data and initial ranking, it only requires feedback about whether probe and gallery images are similar or dissimilar. For each new probe, a new Mahalanobis metric is learnt through the optimization of a loss function similar to that of the WARCA method [78] which aims at quickly pushing up the true matches to the top ranks. A Bregman divergence regularization ensures that the new metric is close to the previous metric thus incrementally improving the metric learnt with previous instances with the new human feedback. Once human feedback is not available anymore, instead of directly using the last updated Mahalanobis metric for matching, the paper [115] proposes to learn a strong ensemble model from the set of metrics obtained throughout the feedback process which are considered as weak models.

Automatic re-ranking

Instead of relying on human feedback as in POP [114] and Human In the Loop [115] for obtaining better ranking performances, some methods propose to use existing person re-identification methods and add an additional re-ranking step to improve the performances of the initial ranking. Those methods assume that true matches appear in the first ranks without necessarily be the first ones. To find out the most likely true match among the top ranks gallery identities, they develop different strategies mostly based on top ranks neighbours.

Some approaches only make use of ranking information [35, 26, 27, 28]. In [35], the Common Near-Neighbor-Analysis proposes a refined ranking based on the weighted linear combination of two dissimilarities, the direct and the relative dissimilarities. Given two identities i and j, the direct dissimilarity returns the minimum rank between the rank of person i in the ranked list of person j and the rank of person j in the ranked list of person i. The relative dissimilarity is the sum of the ranks of the nearest neighbors of person j in the ranked list of person i and of the ranks of the nearest neighbors of person i in the ranked list of person j. The direct dissimilarity gives an information about whether at least one of the identity is among the nearest neighbour of the other one. The relative dissimilarity evaluates the similarity between the sets of nearest neighbors of the pair of identities.

The approaches [26, 27, 28] also consider the top k-ranks gallery identities and their nearest neighbours to refine the ranking list. A good match should not only appear in the top ranks, but its nearest neighbor set should significantly overlap the probe's nearest neighbor set. This is captured by the Jaccard similarity, which is the ratio of the cardinality of the intersection of two sets over the cardinality of the union of those two same sets. The similarity score of a gallery identity is redefined as the Jaccard similarity weighted by a coefficient related to the original rank of the gallery person. In [27], considering that different descriptors might capture different types of similarities, the probe nearest neighbors among the gallery identities and the gallery nearest neighbors also among the gallery identities are obtained using different feature descriptors. If probe nearest neighbors are obtained using global descriptors, then gallery's nearest neighbors are found using local descriptors, and vice-versa. In [26], in addition to a first re-ranking step that considers the sets of top ranking gallery identities and their nearest neighbors sets, a second refinement is conducted using dissimilarities. Elements that are strongly dissimilar to the probe element should also be strongly dissimilar to its true match. The more often an identity appears among the nearest neighbors of the bottom ranked gallery identities, the more their ranking is penalized. The paper [28] combines the ideas of the two previous papers [27, 26] by making use of similar and dissimilar people, and using different ranking methods for computing the nearest neighbors of the probe identity and of that of its top ranked gallery identities.

Other approaches make use of both ranking and similarity value information [29, 34]. In [29], the main idea is once again that the probe person and the true gallery match should share many common nearest neighbors. The Jaccard similarity captures the similarity between sets but all nearest neighbors are considered the same way, regarding of whether they are the first nearest neighbor of the k^{th} one. A revised Jaccard similarity that takes into account the distance of gallery images to the probe image is proposed. Along with it, a new feature encoding, the k-reciprocal feature, enables to compute set comparisons by vector calculation. The method proposed in [34] also requires ranking information and similarity value, but it applies only to the case the people we wish to re-identify appear at the same time in one camera observation view or several non overlapping camera views. Based on the fact that people can not appear twice at the same time at different places, the proposed method penalizes for a probe person, the ranks of gallery identities that

are already very likely to be the true match of another probe person that appears concurrently.

Different variants of the Discriminant Context Information Analysis (DCIA) are proposed in [30, 31, 32]. DCIA aims at learning what distinguishes people that look similar, it focuses on the visual ambiguities shared between the first ranks. A first training step learns a model to distinguish true and false matches. Then the second training step requires the introduction of two notions, the content set and the context set. The content set contains gallery images that have low dissimilarity with the probe person. The context set contains gallery images that have low dissimilarity with both the probe person and a gallery person that belongs to the content set. Based on these content and context sets, a feature transformation step is conducted to remove the visual ambiguities of the first ranks. Finally, a new model is learnt on the new features for checking whether two images are from the same person or not. At test time, after obtaining a first ranking thanks to the first model, only the content sets are re-ranked using the transformed features and the second model.

2.4 Generalizing person re-identification

Person Re-identification methods performances have tremendously risen in the past few years for the closed world case. However, the situation where every query person has been previously identitified and is thus included in the gallery set, does not correspond to the context of most real applications. Therefore the open world re-identification topic has emerged and is gaining strong interest.

However, many ways of relaxing the closed world assumption exist and the person re-identification research commonity has not yet plebiscite one situation over another. Indeed, "open" simply refers to the fact that the gallery set of known people is not comprehensive. It could correspond, for example, to a case where the gallery is empty and grows as new identities come in, or to a case where not all identities are included in the gallery set but the gallery remains unchanged even if a new identity is captured. Many more situations could match the expression "open world", each being specific and corresponding to different applications with distinctive goals. This is why even though most methods first developed for the closed world task could apply to open world tasks, and vice-versa, we chose to present the methods specifically tackling an open world task in a separate section.

We grouped the existing papers generalizing person re-identification task into three groups corresponding to the type of scenario tackled: identity inference, group-based verification, detection and re-identification.

2.4.1 Identity Inference

Literally, **identity inference** consists in inferring people's identity. This task is closely related to multi-person tracking. Given a video sequence or a collection of

sequences where multiple people appear, the goal of identity inference is to assign an identity to every person detection in the sequence(s). This task can be considered as a generalization of closed world person re-identification because images are not assumed to be already grouped by identity. This task actually send us back to where person re-identification started as an independent vision task. Instead of assuming that images are already grouped by identities thanks to tracking algorithms, here all detections must be labelled.

Both papers [116, 1] which discuss this topic use a Conditional Random Field model (CRF) for minimizing an energy function composed of a unary energy function and a pairwise energy function. In [116], the model assumes that the gallery is known and a few images of each identity are available. The unknown query detections however are not grouped by identity. The unary potential penalizes the cost of associating a detection to an identity according to its L1-distance to the closest gallery representation of that identity. The binary potential favors attributing the same label to similar detections. The method presented in [1] can be used in a more general context where many of the closed world re-identification assumptions are relaxed. Not only does it not require to have a set of known labels (gallery), but it also does not assume that every identity appears in several cameras. Therefore, by assigning a label to every person detection, the method [1] also returns the number of people present in the scene. Spatio-temporal information are taken into account. The labels are first initialized using information within each camera, before being further refined for matching identities across cameras. The identities inference are based on a global inference using all the video sequences across all the cameras.

The scenario in [1] illustrates well an open world case: not only the gallery set is not comprehensive, but there is actually no gallery. However, this situation is actually more related to tracking than to re-identification. Camera relations are supposed to be known. The whole scene, with all people's detection, is needed to determine each person's identity.

2.4.2 Group-based verification

Literally, **group-based verification** consists in determining whether a person belongs to a given group or not. The associated scenario is the following. A small set of identities is known, it is referred to as the target set. When given a query identity, the goal is to verify whether he is one of the target identities or not. This task is strongly related to the forensic application where a group of wanted people are the target set, and the query images corresponds to all the pedestrian detections captured by different surveillance cameras. This group-based verification task can be considered as an open world re-identification task since the target set of known identities is only a small subset of all the existing people.

Three papers [2, 3, 4] tackle this group-based verification task. Contrary to the usual closed world case, where the training and the testing sets contain non overlapping sets of identities, for the group-based verification, the target set of identity

stays the same in the training and the testing phases. The non target set is different in training and testing phases. In [2, 3, 4], metric learning approaches are proposed to learn from the target set and the non target training set. It is assumed that only one image of each target identity is available, while the identities in the non target training set can be represented by multiple images. Therefore, to compensate for the lack of information about target identities intra-class variance, the non target set images are exploited. The main ideas of the work in one-shot group-based verification are the following. The intra-class variance of non target identities should be smaller than the extra-class variance, in both target and non target set. The extra-class variance of the target set should be smaller than the extra-class variance between identities in the target set and in the non target set.

Intuitively, it might seem that in order to verify the membership of a person to a group, one must find a target identity with whom the probe person matches, but the previous methods showed that it is possible to determine that a person belongs to a group without finding out his identity. Therefore, the group-based verification scenario is not really a re-identification task since even if the query person is found to be among the target set, the group-based verification task does not require him to be re-identified.

2.4.3 Detection and Re-Identification

In this subsection, the open world task is defined by two subtasks: a Detection task and Re-Identification task. As it was the case for closed world re-identification, there is a list of known people, the gallery. When presented images of a query person, two questions need to be answered: has that person been previously identified? If so, who is he?

The paper [13] introduces the OPERID dataset for person re-identification with images captured by 6 cameras such that each identity appears only in a subset of those cameras. Benchmark open world test partitions are provided with the dataset. In those partitions, the probe set contains identities that are not present in the gallery. In addition, the multiple cameras aspect also make the re-identification scenario difficult to tackle. In the test partitions, gallery images come from one camera view, while probe images come from all five other cameras, but each probe image is not a priori associated with other probe images of the same person. This paper [13] does not propose any new re-identification method, but highlight the fact that existing methods are not yet performing enough for the open world detection and re-identification task. It is also this paper that proposes the DIR versus FAR evaluation metric presented in the next subsection.

In [46], it is also the detection and re-identification task that is considered, but in a one-shot case. A semi-supervised method is presented, but it relies on the additional assumption that for each identity in each camera, there is at most one image. Though presented as a method for open set re-identification, the open world aspect, meaning that the gallery is not overcomplete and the probe person might

not be in the gallery, is not directly taken into account. The proposed method is a subspace learning method that aims at minimizing the variance of similar pairs of images that are captured by different camera view and at maximizing the variance of similar pairs of images captured by the same camera. In doing so, the variations in appearance due to different camera view is minimized, while the variation associated to distinct people is reinforced.

2.4.4 Drone based

Paper [48] generalizes the person-re-identification task based on fixed cameras into several tasks where cameras are not fixed anymore but in movement on flying drones. This leads to very challenging open world re-identification tasks. Three types of scenarios are considered: watchlist verification, within-flight re-identification and across-flight re-identification.

2.5 Evaluation measures

2.5.1 Closed world measures

Closed world person re-identification measures mainly evaluate the relevance of the relative ranking of gallery similarity to each probe person. We present these measures from the most widely used to the least used.

- **CMC (Cumulative Matching Characteristic)** represents the proportion of right matches found in the r top ranks. The ranks range from one to the number of gallery identities. To use this measure, it is assumed that for each probe person, the person re-identification method returns a ranked list of the gallery identities, where each gallery identity appears only once. In the multi-shot case, the similarity between a probe person and a gallery person is often chosen as the maximum value of the similarity between their image pairs.
- **mAP (mean Average Precision)** is used for multi-shot cases. Gallery images of the same person are not grouped together so as to return a single similarity value for each gallery person. Instead, a ranked list of gallery images is returned for every query person. For a set of queries $\mathcal{Q}$, the mean average precision is the mean of the average precision of each query:

$$mAP = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} AP(q) \tag{2.8}$$

  where AP is the average precision.

  The average precision is the mean of the precision scores after each relevant gallery image is retrieved. It is defined by:

$$AP(q) = \frac{1}{R} \sum_{r} Precision_r(q) \tag{2.9}$$

  where r is the rank of each relevant gallery image, R is the total number of relevant gallery images and $Precision_r(q)$ is the proportion of relevant gallery

images in the top r gallery images. What we call here a relevant gallery image is a gallery image which represents the query person.

- **MRR (Mean Reciprocal Rank)** is the average of the reciprocal ranks for a set of queries $\mathcal{Q}$. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct match. The other matches are not taken into account. MMR is given by:

$$MRR = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{r(q)} \tag{2.10}$$

  where $r(q)$ refers to the rank position of the first relevant gallery image for the query q.

- **PUR (Proportion of Uncertainty Removed)** is the normalized entropy reduction between a randomized rank and any given method's output rank. It accommodates information from across the entire CMC, rather than at arbitrary values of rank r. It is defined in [80] by:

$$PUR = \frac{log(N) - \sum_{r=1}^{N} CMC(r) log(CMC(r))}{log(N)} \tag{2.11}$$

  where N is the number of gallery images.

2.5.2 Open world measures

For each of the open world scenarios, a different measure is employed.

- When open world is considered as a **binary classification problem**, or as an **identity inference problem**, as in [116, 1], the performances are assessed by information retrieval measures.

  A Positive pair is a pair of probe-gallery images or identities which corresponds to one single person.
  A Negative pair is a pair of probe-gallery images or identities which corresponds to two distinct people.
  A **True Positive** pair (**TP**) is a correctly accepted pair, ie. a pair of probe-gallery images or identities which represents the same person and which is detected as so.
  A **True Negative** pair (**TN**) is a correctly rejected pair. ie. a pair of probe-gallery images or identities which corresponds to two distinct people and is detected as so.
  A **False Positive** pair (**FP**) is a wrongly accepted pair, ie. a pair of probe-gallery images or identities which corresponds to two distinct people but which is classified as the same person.
  A **False Negative** pair (**FN**) is a wrongly rejected pair, ie. a pair of probe-gallery images or identities which corresponds to one person but which is classified as corresponding to two distinct people.

TP, TN, FP, FN also refer respectively to the number of true positive pairs, the number of true negative pairs, fthe number of alse positive pairs and the number of false negative pairs.

Based on these notions of TP, TN , FP and FN, the following evaluation measures can be used.
Recall is also called TP rate (true positive rate). It is the ratio of true positive pairs (TP) over the number of positive pairs (TP+FN):

$$recall = \frac{TP}{TP + FN} \tag{2.12}$$

Specificity is also called TN rate (true negative rate). Is is the ratio of true negative pairs (TN) over the number of negative pairs (TN+FP):

$$specificity = \frac{TN}{TN + FP} \tag{2.13}$$

Precision is the ratio of the number of true positive pairs (TP) over the number of pairs that are found to be positive (TP+FP):

$$precision = \frac{TP}{TP + FP} \tag{2.14}$$

Accuracy is the proportion of well classified pairs:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.15}$$

F-score is the harmonic mean of precision and recall

$$F = \frac{2}{\frac{TP+FP}{TP} + \frac{TP+FN}{TP}} = \frac{2TP}{2TP + FP + FN} \tag{2.16}$$

- In the case open world re-id is cast as a **group-based verification task**, as in [2, 3, 4], True Target recognition Rate (**TTR**) and False Target recognition Rate (**FTR**) are used as evaluation measures. They are defined by:

$$TTR = \frac{|\mathcal{TTQ}|}{|\mathcal{TQ}|} \tag{2.17}$$

$$FTR = \frac{|\mathcal{FNTQ}|}{|\mathcal{NTQ}|} \tag{2.18}$$

where
$\mathcal{TTQ}$ = { query target images from target people }
$\mathcal{NTQ}$ = { query non-target images from non-target people }

$\mathcal{TQ}$ = { query target images that are verified as on of the target people }
$\mathcal{FNTQ}$ = { query non-target images that are verified as one of the target people }.

- In the case open world re-id is cast as a **Detection and Identification task**, as in [13], performance is evaluated with Detection and Identification Rate (DIR) versus False Acceptance Rate (FAR) which are defined by:

$$DIR(\tau, r) = \frac{|\{i | i \in P \cap G, rank(i) \leq r, dist(i^g, i^p) \leq \tau\}|}{|P \cap G|} \tag{2.19}$$

$$FAR(\tau) = \frac{|\{i | i \in P \backslash G, min_{j \in G} dist(j^g, i^p) \leq \tau\}|}{|P \backslash G|} \tag{2.20}$$

where P is the set of probe identities, G is the set of gallery identities, $P \cap G$ are common identities, $P \backslash G$ are probe imposter identities and $dist(j^g, i^p)$ is the distance between the set of images of person j in the gallery and the set of images of person i in the probe set. $DIR(\tau, r)$ represents the proportion of common identities that are re-identified before rank r with a distance smaller than τ and $FAR(\tau)$ is the proportion of imposter identities whose distance to their closest gallery identity is smaller than τ.

2.6 Conclusion

In this chapter, we have given an overview of the main work that has been conducted on the person re-identification task since its beginning a decade ago. From a subtask of the multi-camera tracking task, person re-identification in the closed world setting has become an integral task. With the growing interest in the open world problematics, person re-identification tasks are even starting to be splitted into several groups, the identity inference group, the group-based verification group and the detection and re-identification group. More groups might appear since this problematic is still a novel one in the process of emerging.

This evolution of the person re-identification task is actually a consequence of the increase in performances of re-identification approaches which has allowed researchers to relax some assumptions and deal with tasks that are closer to the actual need of practical applications. Mainly based on intuitions, early methods [16, 17] proposed hand-crafted features and matching schemes. With the apparition of supervised methods, especially metric learning approaches [75, 20, 21, 23], performances have improved a lot. While these metric learning approaches were applied on previously developed features, the apparition of deep learning methods [12] allowed for supervised methods to learn from raw images. Today, deep learning methods are very popular and often outperform other methods. A large majority of the recent papers are deep learning based approaches. Nonetheless, aside these huge trends, some papers still propose some alternative methods based on sparse

representation or re-ranking for example.

The successful development of supervised methods has been made possible by the creation of larger datasets. Indeed, the learning phase requires some training data. Deep learning methods in particular need huge amounts of data during the training process so that it doesn't overfit and generalizes well at test time. Besides the need for more training data, datasets are also evolving in order to become more representative of the data that could be collected in a large scale by real applications. It contains images from more identities captured by multiple cameras and the bounding boxes are automatically generated.

Overall, the evolution of the datasets, the evolution of the re-identification approaches, the improvement of the performances and the change in the definition of re-identification tasks go hand in hand. The evolution of one component influences the evolution of another one leading to better performances and more and more realistic scenarios and datasets. However, these improvements are achieved thanks to a learning phase on training data and the learnt model often overfit the dataset used for training. Better results are obtained mostly because features and metrics are adapted to each dataset and to the closed world scenario but all the challenges mentioned in the introduction are still unsolved.

2.7 Position of our work

Given that approaches based only on designing robust hand-crafted features perform poorly, they must be combined with a learning step. Therefore in this thesis we chose to focus on the matching step for which we proposed two approaches, a metric learning approach and a sparse representation based approach. Nonetheless, considering that features do play a major role in the re-identification process, all the experiments have been conducted using several types of features so as to show the robustness and the relevance of the proposed matching methods regardless of the features.

Moreover, even though many datasets now contain multiple images for each person, the information brought by these multiple images is often not exploited to its fullest. A few approaches do extract temporal information when full tracklets are available but they are not usable if the multiple images come from different tracklets. In the methods we propose, we do not exploit temporal information. Instead, in our metric learning approach, we exploit the multiple images but also attempt to lessen the impact of images that could potentially be outliers. In our sparse representation approach, the variability of the images of a person is kept. Be it a gallery or a probe person, a person is represented by all its images with his different poses, under varying illumination conditions, etc. The similarity score however is computed per pair of gallery-probe identities in such a way that the image of a probe person does not need to be similar to all the images of a gallery person to be deemed corresponding to the same person, but it is only required to be similar to at least one or a few of them.

As for the re-identification scenarios, we consider both the usually tackled closed world person re-identification scenario and the novel open world detection and re-identification scenario. The metric learning approach we proposed focuses on the open world aspect assuming that when the open world task is solved, the closed world task is solved as well. The sparse representation approach on the other hand considers the specificities of both the closed and the open world task: being able to rank gallery identities and being able to take a decision on whether two images are from the same person or not are both important.

Chapter 3

COPReV

In this chapter, we first re-expose what closed and open world re-identification tasks are in order to better point our their differences in terms of scenario and evaluation. Then, we present our COPReV method (Closed and Open Person RE-identification and Verification method) which addresses the open world re-identification problem by adopting a binary verification perspective because we focus on tackling what we consider is the main difference between closed and open world tasks, namely the necessity of returning a decision about whether a test person belongs to the gallery or not. Finally we evaluate the proposed COPReV approach on closed and open world re-identification tasks and on the person verification task.

3.1 Motivation

3.1.1 Closed world re-identification

What is commonly referred to as the person Re-ID task, is what we refer to by closed world Re-ID task. It literally consists in re-identifying people, ie. finding their identity given that they had been previously identified. A query probe person is also one of the gallery people and the aim is to find which one is the right match. Since we are never a hundred percent certain that our best conjecture is the right one, what is actually returned is not the identity of the gallery person who seems to be the best match for the presented probe person, but a ranked list of all the gallery identities from the most probable match to the least likely one. Therefore the rank at which the right match is found is given a huge importance in all the existing evaluation measures for assessing the performance of closed world re-id approaches. The most often used closed world evaluation is the CMC (Cumulative Matching Curve) which represents the proportion of right matches found in the top r ranks. Papers often report the CMC value for a few ranks (rank 1, 5, 10 and 20). What is assessed is the ability of a method to rank well the gallery identities. A method which has a higher recognition rate at first rank is considered better than a method which has a lower recognition rate at first rank.

However, the ranking ability in question is only a relative ranking ability com-

puted independently for each probe person. We talk about relative ranking because the only requirement is for the right gallery match to be better ranked than the false gallery matches for a given probe person. The similarity score of the right gallery match for a given probe person does not need to be bigger than the similarity score of a wrong gallery match involving another probe person. Let's take a look at the toy examples in Figure 3.1 for a better understanding of relative ranking.

In Figure 3.1, three cases are presented for the same test data, corresponding to methods A B and C. These are not real examples, but help illustrate the relative ranking issue. There are five identities in total. Every row corresponds to a probe person. The y-axis corresponds to dissimilarity scores. Each dot represents the dissimilarity score of a pair of probe-gallery identities. Green dots represent positive pairs (same identity) and red dots represent negative pairs (distinct people). In all three cases, for each probe person (ie. on each row), the green dot is on the left of the red dots, ie. the dissimilarity score of the positive pair is smaller than the dissimilarity of the negative pairs. The right match is always the first match, so all probe identities are re-identified at first rank, the recognition rank is of 100% from the first rank on and all three methods have perfect CMC scores. Nonetheless, we can notice that for methods B and C, there are wrong matches with smaller dissimilarity scores than some right matches. In the CMC evaluation, there is no comparison between the dissimilarity of a right match for a given probe person and the dissimilarity of wrong matches corresponding to other probe people.

3.1.2 From closed world re-id to open world detection and re-identification

In the open world re-identification task we tackle in this thesis, we relax the assumption that the person to be re-identified has been identified before. Thefore strictly speaking, open world re-identification is not a re-identification task. The open world re-identification task we adopt here is actually a generalization of the closed world re-identification task which can be decomposed into two subtasks, the detection and the re-identification subtasks. The detection task consists in determining whether the presented probe person should be matched with one of the gallery people or if he should be rejected as an imposter (someone who is not present in the gallery). The objective of the re-identification task is to rank the gallery identities whom are considered to be possible right matches for the presented probe person.

Contrary to the closed world task, for the open world task, besides the relative ranking aspect, the detection aspect also needs to be evaluated. The evaluation commonly adopted in the few papers which dealt with the same problem ([13, 46, 47]) is the one presented in the Operid paper [13]. The proposed DIR vs FAR evaluation (defined in section 2.5.2) is a unified measure for detection and ranking. For reminder, the Detection and Identification Rate $DIR(\tau, r)$ represents the proportion of common identities that are found in the first r ranks with a dissimilarity score smaller than τ and the False Acceptance Rate $FAR(\tau)$ is the proportion of imposter identities whose dissimilarity score to their closest gallery identity is smaller than τ. Papers mostly report the DIR at first rank for given values of FAR. In that case,

only the first rank matches are considered. Indeed, FAR is only a function of τ and it only takes into account the worst wrong match, ie. the one with the smallest dissimilarity score, the other wrong matches for bigger dissimilarity score are not taken into account in the FAR value. DIR is a function of the rank r and of the threshold τ, so DIR at first rank is solely a function of the threshold τ, and it only takes into account the first match. DIR at first rank versus FAR are functions of the threshold variable τ and it reflects the proportion of well re-identified non imposter probe identities for given proportions of wrongly matched imposter probe identities, where the dissimilarity score of matched elements are below the threshold τ.

Let's compare once again the methods A, B and C presented in Figure 3.1 but in the open world case this time. For the sake of the example, we examine two open world partitions of the identities. In both partitions, the probe set of identities is composed of all 5 identities. In the first partition, the gallery set contains people 1, 2 and 3. In the second partition, the gallery set contains people 3, 4 and 5. Since there are only two probe imposters, the non zero False Acceptance Rate (FAR) only takes 2 values, 50% when only one of them is wrongly accepted and 100% when both are wrongly accepted. For each FAR value, the value of the dissimilarity score of the wrongly accepted probe imposter is used as the decision threshold τ to determine the corresponding DIR rate. In our example, since there are two possible values of FAR, there are two decision thresholds τ_1 and τ_2. Figure 3.2 presents the open world situations for the two described partitions. For each probe person, only the first match is taken into account, ie. we only consider the pair with the smallest dissimilarity score. They are circled in black in the figure. For each FAR value and its corresponding decision threshold value, the DIR value is computed by counting the proportion of non imposter probe identites re-identified with a dissimilarity value smaller than the decision threshold. On the first partition, for all three methods, the open world DIR vs FAR results are perfect (100% recognition at first rank when the false acceptance rate equals 50%). Indeed, the circled green dots corresponds to dissimilarity values that are smaller than the decision threshold τ_1 associated to the first wrong accepted probe imposter identity. On the second partition however, the performances are really different depending on the method. While method A still achieves perfect DIR vs FAR results, method B and C perform poorly. The open world results are provided in Table 3.1.

Partition 1		
FAR (%)	50	100
Method A	100	100
Method B	100	100
Method C	100	100

Partition		
FAR (%)	50	100
Method A	100	100
Method B	0	33.3
Method C	0	0

Table 3.1 – DIR vs FAR performances for the two partitions considered for the toy examples presented in Figure 3.2.

This toy example shows how different the closed and open world re-identification tasks and evaluations are. While the three methods A, B and C had perfect CMC

results in the closed world setting, in the open world case, the DIR vs FAR performances greatly differ between methods but also depending on the partitions for methods B and C. This is because CMC only evaluates ranking whereas DIR at first rank versus FAR evaluates the adequacy of first rank matches to a decision rule (threshold).

In an open world task, a decision has to be taken about accepting or rejecting a probe person as one of the gallery identities. Therefore, a method with good ranking abilities alone does not guarantee good open world performances. The methods B and C are good examples. Even when for each probe identity the right matches are ranked first before the wrong matches, the dissimilarity score of right matches of a given probe person are not necessarily smaller than those of negative matches involving a different probe person, which can lead to small DIR values when some gallery identities are removed from the gallery set for the open world scenario.

On the contrary, a verification method or a binary classification method that perfectly determines whether two sets of images come from the same person or from distinct people solves both the closed and open world Re-ID problems. The method A is a good illustration of such a method. The decision rule of the verification method is not necessarily a fixed threshold decision rule where the distance of positive pairs is smaller than the specified threshold and the distance of negative pairs is bigger than the threshold. It could be some sophisticated decision rule, but the important point is that once you are able to distinguish positive pairs from negative pairs, you can always rank positive pairs before negative pairs, and therefore such a method can also perform well in closed world settings. Notice that even in the closed world case, there is actually no ground truth ranking of wrong matches, the right matches are the only ones that should be ranked before the wrong ones.

To sum up, due to the relaxation of the assumption that every presented probe person is present in the gallery, the open world task we consider can not simply be assessed by the CMC evaluation and it requires to employ a better adapted evaluation measure. The toy examples we assessed using CMC and DIR vs FAR evaluation brought to light the shortcoming of the CMC evaluation which only measures a relative ranking ability and which can therefore not be representative of the open world performances.

3.1.3 Existing closed world re-id approaches used in open world re-identification

Since the person re-identification task has been tackled only in the closed world setting for a long time and were evaluated using CMC, a few methods are only based on ranking constraints per probe identity. It is for example the case of metric learning methods EIML [76] and WARCA [78] or neural network approaches [94, 95, 45]. Although it is possible that some of these approaches perform well for the open world re-id task, nothing in their formulation ensures it.

There exist some methods which instead of being based on ranking constraints per identity are based on minimization problems which involve constraints on positive and negative pairs and which do not specifically distinguish these constraints per probe person. It is for example the case of metric learning methods KISSME [21] and XQDA [23] where the objective is to minimize intra-class variance and maximize extra-class variance, or RPLM [79] whose objective is to minimize the distance of positive pairs and maximize the distance of negative pairs. However, these kind of methods do not specify any threshold under which the intra-class variance and the positive pairs distances should be nor a threshold above which the extra-class variance and negative pairs distances should be. Therefore, depending on the dataset, or even on the partitions, the decision threshold corresponding to a given FAR value could vary a lot.

Some methods developed for closed world person re-identification, such as the PCCA [20], cast the re-identification task as a binary classification task and introduce a threshold during the training phase. However, even when a threshold is used during the training phase to separate positive pairs from negative pairs, the best decision threshold on the test set might not be the one used for the training phase because the distribution of the distances of test data are often shifted to the right (bigger values) because negative pairs are better modeled.

3.1.4 Existing open world re-id approaches

Some papers specifically tackle an open world re-identification task. Among them, some [1, 2, 3, 4] tackle different open world tasks from the one we deal with while a few others [13, 46, 47] also tackle the two subtasks detection and re-identification open world task we consider.

The papers [1] which have a different definition of the open world re-identification task from ours present methods that focus on the specific aspect of their definition of the open world re-identification task. In the multi label inference paper [1], the optimization of the cost function aims at grouping images by identity and assigning a label to each detection. Rather than having to determine whether a person is a known gallery person or not, the total number of probe identities is unknown and to be determined by the algorithm there is actually no gallery. In [2, 3, 94] the gallery is assumed to be composed of a small group of target identities and the goal is to determine whether a probe person is one of the gallery identities or not, without necessarily giving the exact identity of a probe person who is found to be someone from the target set. In all three methods, the optimization is based on relative ranking constraints.

As for papers which evoke the same open world person re-identification task as we do, they do not focus on what we argue makes the open world re-id task so different from the closed world re-id task. The paper [13] does not propose any new method to tackle the open world re-id task but experimentally shows that

several existing metric learning methods do not perform well for the open world re-id task. The method proposed in [47], focuses on feature design (it exploits space-time information from tracklets) and their matching without specifically tackling closed or open world re-identification. In [46], the method is supposed to be designed for an unsupervised single-shot open world re-id task, however, its formulation which enforces visually similar pairs from the same camera to be pushed apart and visually similar pairs from different camera to be pulled together does not seem to specifically take into account the difference between closed and open world re-identification.

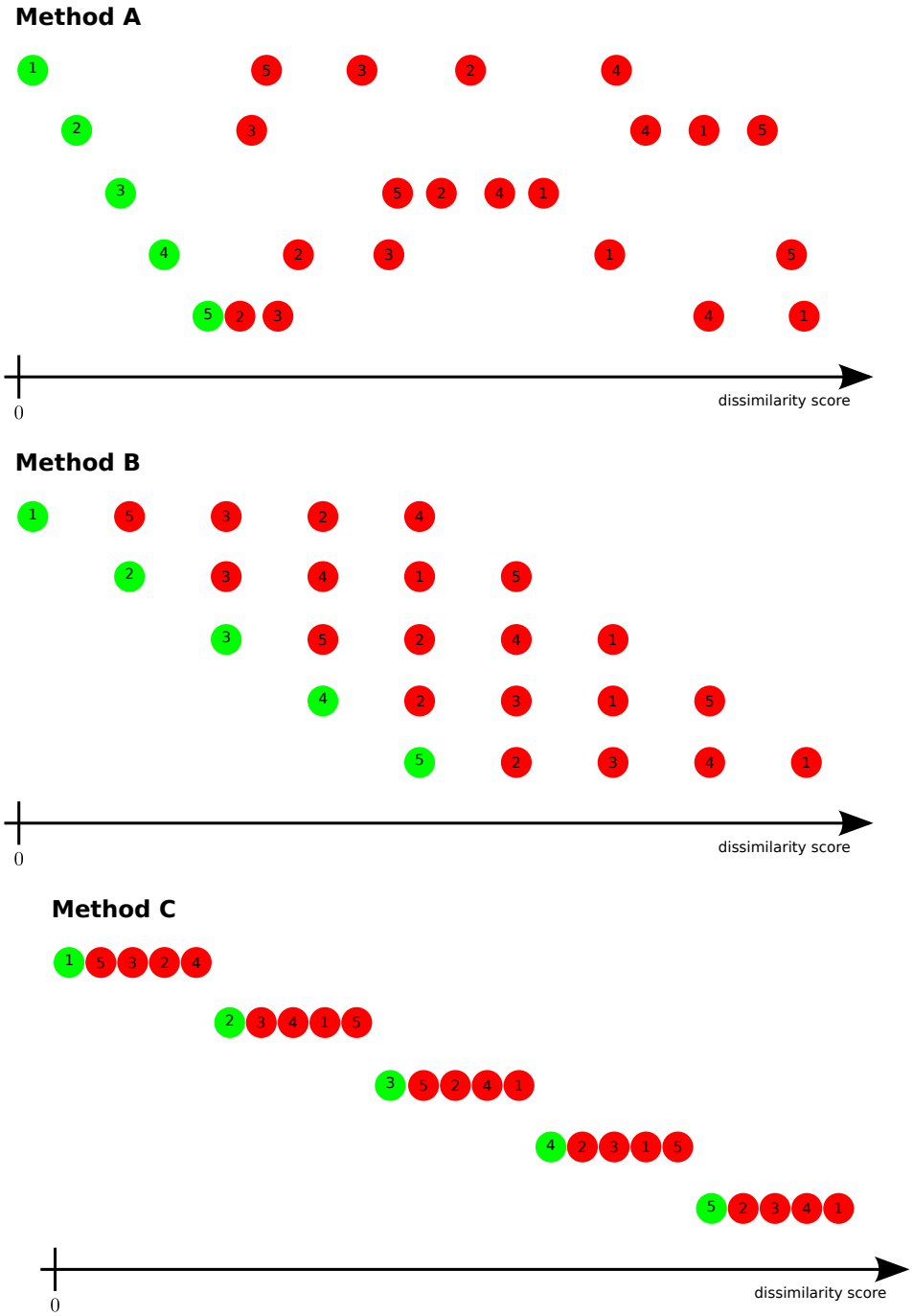


Figure 3.1 – Three toy examples of closed world dissimilarity scores leading to perfect CMC evaluation.
There are 5 people. On each row are presented the dissimilarity scores (green and red dots) of one probe person with every gallery person. Each dot corresponds to the dissimilarity of a pair of probe-gallery identities, where the probe identity is given by the row and the gallery identity is given by the number in the dot. Green dots corresponds to right matches and red dots to wrong matches. For all three methods A B and C, for each probe person, the right matches have smaller dissimilarity scores than wrong matches.

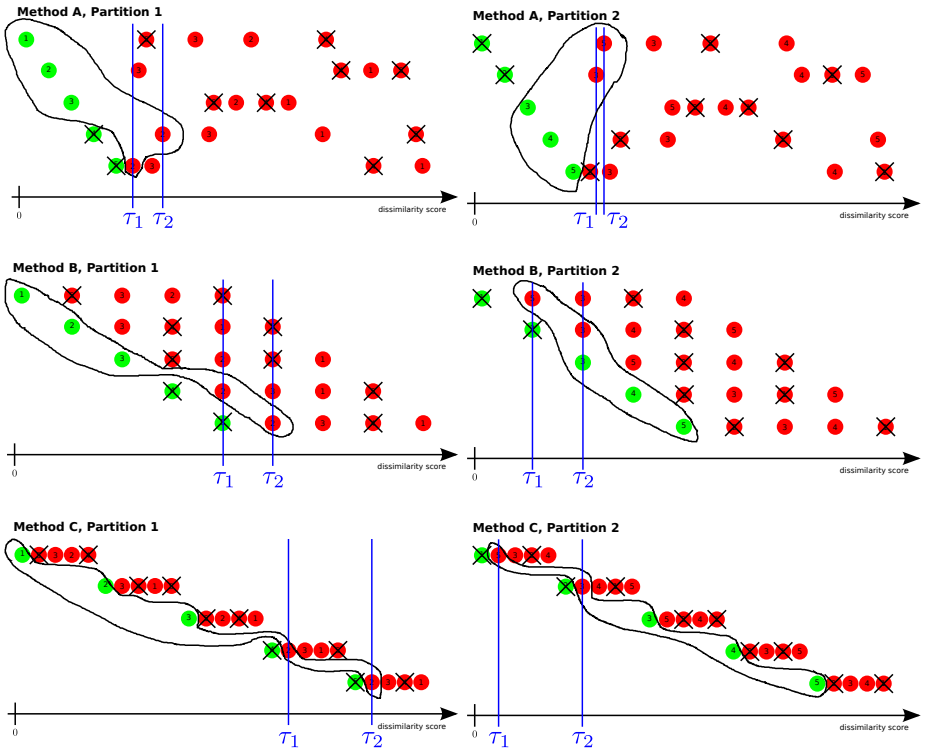


Figure 3.2 – Three toy examples for 2 open world partition sets leading to very different DIR at first rank vs FAR performances.
The dissimilarity scores for the three methods A, B and C are displayed for all probe-gallery pair of identities. We consider 2 open world partition sets. In the first partition the gallery only contains identities 1, 2 and 3. In the second partition, the gallery only contains identities 3, 4 and 5. Since some gallery identities are not present, the associated pairs are crossed. The DIR at first rank versus FAR evaluation only takes into account the best match for each probe person. These best matches are circled in black. FAR can take 2 non zero values, 50% when one of the two probe imposters is not rejected and 100% when both probe imposters are accepted. The threshold decision values τ_1 and τ_2 are shown in blue for each FAR value (50% and 100%). DIR can take 4 values, 0, 33.3, 66.7 and 100 depending on the number of non probe imposters re-identified at first rank with a dissimilarity score smaller or equal to the decision thresholds τ_1 or τ_2.

3.2 COPReV

3.2.1 Overview

Considering that for open world scenarios a decision needs to be taken about whether a probe person belongs to the gallery or not and given that a perfect binary classification method would solve the closed and open world re-identification tasks and the verification task, we formulate the re-identification task as a verification task. We name our method COPReV for Closed and Open world Person RE-identification and Verification. In order to have an easily usable method where the decision threshold is clear, we introduce an arbitrary threshold during the training phase to learn a linear transformation of the features such that distances of positive pairs are below the threshold and that of negative pairs are above that threshold. For the person verification task, the same threshold can be used during the test phase for distinguishing positive and negative matches. We propose a cost function that limits the effect of unbalanced data and outliers during training by promoting a high proportion of well classified positive and negative pairs for every identity. Our method therefore handles the unbalanced number of positive and negative pairs, as well as the unbalanced number of images per identity, and most importantly, though cast as a verification task, it works for closed and open world scenarios.

3.2.2 Problem notations

Let $\tau \in \mathbb{R}$ be an arbitrarily fixed threshold and $L \in \mathbb{R}^{d' \times d}$ the linear transformation matrix we look for, where d is the dimension of the initial features, and $d' \leq d$ is the dimension of the final transformed features. Let $\mathcal{I}$ be the set of identities in the training set and K its cardinality. $x_{il} \in \mathbb{R}^d$ represents the feature in the initial space of the l^{th} image of person $i \in \mathcal{I}$, with $l \in [1, \ldots, n_i]$, where $n_i \in \mathbb{N}$ denotes the number of images of person i. Let $D_{ii} = \{x_{il} - x_{il'}\}_{l,l' \in [1,n_i], l<l'}$ be the set of difference between positive pairs of features of person i and m_{ii} its cardinality. Let $D_{ij} = \{x_{il} - x_{jl'}\}_{i \neq j, l \in [1,n_i], l' \in [1,n_j]}$ be the set of difference between negative pairs of features involving person i and j, and m_{ij} its cardinality.

3.2.3 Mathematical formulation

We cast the open world re-identification scenario as a binary classification task. Our goal is to find a linear transformation L of the features such that the distances of positive respectively negative pairs are smaller respectively bigger than the decision threshold. For an easy use, the decision threshold which enables to accept or reject a probe person is fixed during the training phase and the same threshold can be used for test.

We reckon that penalizing misclassified pairs in a linear way with respect to their distance to the decision threshold as done in PCCA [20] is sensible to outliers. Reasoning in terms of number of misclassified pairs seems more robust to outliers, but this favors elements that are widely represented such as identities that are captured

by many images and negative pairs that are much more numerous than positive pairs. Therefore, we will minimize proportions of misclassified pairs rather than the number of misclassified pairs.

To ensure that every couple of identities is given the same importance, regardless of the number of images each identity appears in, the loss associated to a pair of images of a given couple of identities is weighted by the inverse of the total number of pairs involving that couple of identities. Similarly, to avoid favoring negative pairs over positive ones, the maximum cost associated to positive and negative pairs should be the same. Consequently, we minimize for each identity the proportion of misclassified positive pairs and for each couple of distinct identities, $\frac{1}{K-1}$ times the proportion of misclassified negative pairs. This is expressed by the following cost function:

$$E(L) = \sum_{i \in \mathcal{I}} \left[\frac{1}{m_{ii}} \sum_{y \in D_{ii}} \mathcal{L}_+ \left(\|Ly\|_2^2 - \tau \right) + \frac{1}{K-1} \sum_{j \in \mathcal{I} \setminus i} \left(\frac{1}{m_{ij}} \sum_{y \in D_{ij}} \mathcal{L}_- \left(\tau - \|Ly\|_2^2 \right) \right) \right] \tag{3.1}$$

where $\mathcal{L}_-$ and $\mathcal{L}_+$ are the loss functions respectively applied to negative and positive pairs.

To strictly stick to the previous interpretation of our cost function, the loss functions should be Heaviside functions. However Heaviside functions are not derivable, so instead of replacing them by a random derivable S-shape function that would simply act as a counting function, we take this opportunity to choose a flexible S-shape function that can tackle some more issues. Generalized logistic functions

$$S(z) = A + \frac{B - A}{(C + De^{-\lambda z})^{\frac{1}{\nu}}} \quad \text{with} \quad C > 0, D > 0 \tag{3.2}$$

that are parameterized by six parameters, A, B, C, D, λ, ν, are an extension of sigmoid and logistic functions.

The first aim of the loss function is to bound the loss associated to each pair to a value between 0 and 1 so that it is easier to balance the importance given to each pair, identity or couple of identities. Imposing that the generalized logistic function takes the value 0 when elements are very well classified and the value 1 when they are completely missclassified implies that :

$$\begin{cases} \lambda > 0 \\ A = 0 \\ B = C^{\frac{1}{\nu}} \end{cases} \tag{3.3}$$

or

$$\begin{cases} \lambda < 0 \\ A = 1 \\ B = 1 - C^{\frac{1}{\nu}} \end{cases} \tag{3.4}$$

Thanks to their S shape, generalized logistic functions penalize differently elements that are close to their inflexion point or far away from it. We use this

property to limit the influence of outliers image pairs. When a pair's distance is far from the threshold, it is already very well classified or completely misclassified so a little variation of the distance will not change the classification outcome. Conversely, when a pair's distance is close to the threshold, either it is misclassified but almost in the right class in which case it should be highly encouraged to vary in the right direction, either it is already well classified and it should be strongly prevented from varying in the wrong direction. Therefore, the closer a pair's distance is to the threshold, the more its variation should hold importance, and the further away from the threshold, the less its variation should matter. This implies that the inflexion point of the generalized logistic function corresponds to the case when a pair's distance is equal to the threshold. The second derivate of the generalized logistic function in zero equals zero imposes

$$D = C\nu. \tag{3.5}$$

The two previous constraints reduce to two the number of parameters that need to be chosen. We can only use generalized logistic functions where $\nu > 0$ that are defined by:

$$\begin{cases} \lambda > 0 \\ S(z) = \frac{1}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}}} \end{cases} \tag{3.6}$$

or

$$\begin{cases} \lambda < 0 \\ S(z) = 1 - \frac{1}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}}} \end{cases} \tag{3.7}$$

In Figure 3.3, on the first row, we visualize on the left some generalized logistic functions verifying Eq. 3.6 and on the right some generalized logistic functions verifying Eq. 3.7. On the second row, the corresponding gradient functions are plotted. The generalized logistic functions values are between 0 and 1 and the inflexion point is in 0.

For $\nu = 1$, the loss function is a sigmoid which is symmetrical with respect to its inflexion point. In that case, a variation in opposite direction of the distance of two pairs that are at the same distance to the threshold, one being well classified and the other misclassified, will have no effect on the overall value of the cost function. Well classified elements should be encouraged to be further away from the threshold only when they are very close to it while misclassified elements should still be penalized even when they are far from the threshold. To fulfill this requirement, ν should be smaller than 1 in the Eq. 3.6, and bigger than 1 in the Eq. 3.7.

$$\begin{cases} \lambda > 0, \\ \nu < 1, \\ S(z) = \frac{1}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}}} \end{cases} \tag{3.8}$$

or

$$\begin{cases} \lambda < 0, \\ \nu > 1, \\ S(z) = 1 - \frac{1}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}}} \end{cases} \tag{3.9}$$

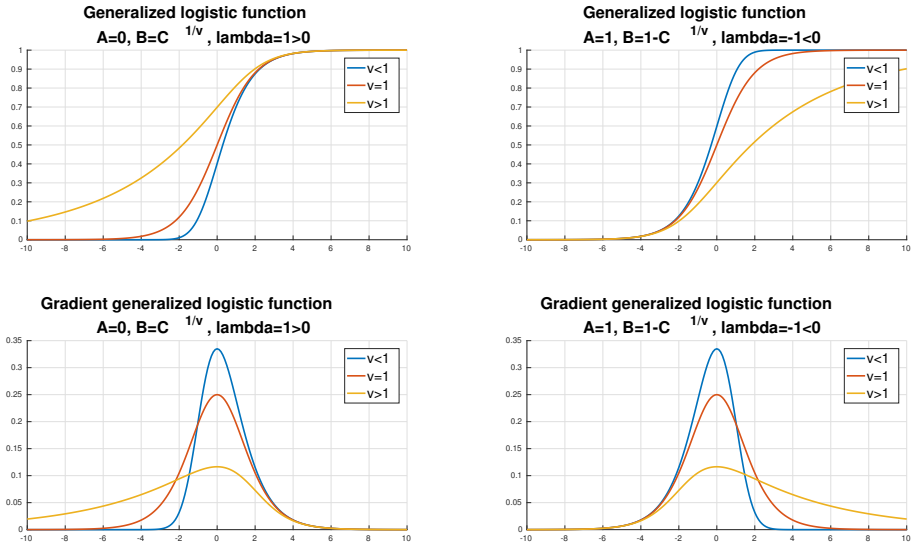


Figure 3.3 – Examples of generalized logistic functions and their gradient.
On the first row are plotted generalized logistic functions verifying Eq. 3.6 (left) and Eq. 3.7 (right). The corresponding gradient functions are plotted on the second row. Red curves are sigmoides functions. For S shaped functions verying Eq. 3.6 (left), the gradient is bigger for positive values than for negative values when $\nu < 1$ (blue). For S shape functions verying Eq. 3.7 (right), the gradient is bigger for positive values than for negative values when $\nu > 1$ (yellow).

Distances have 0 as a lower bound but do not have an upper bound. Therefore, well classified negative pairs can be far from the threshold but well classified positive pairs can not. To compensate for this asymmetry, the loss functions $\mathcal{L}_+$ and $\mathcal{L}_-$ associated to positive and negative pairs should be different.

In the end, for a given choice of threshold τ, there are four parameters to choose: ν_+ and λ_+ for $\mathcal{L}_+$, and ν_- and λ_- for $\mathcal{L}_-$. The value of the threshold itself is not important. To obtain the same results with another choice of threshold τ_2, the parameters ν_+ and ν_- are unchanged while the λ_+ and λ_-parameters should be multiplied by the ratio $\frac{\tau}{\tau_2}$. What is important about the threshold is that it is the same one that is used during the training phase and for the test phase of the verification task.

3.2.4 Optimization

Our cost function is bounded and not monotonous, so there are minima. The optimization of the cost is done using gradient descent algorithms. The gradient is given by:

$$\begin{aligned}\frac{\partial E}{\partial L}(L) = \; & 2\sum_{i\in\mathcal{I}}\left[\frac{1}{m_{ii}}\sum_{y\in D_{ii}}\mathcal{L}'_{+}\left(\|Ly\|_2^2-\tau\right)Lyy^T\right. \\ & \left.+\frac{1}{K-1}\sum_{j\in\mathcal{I}\setminus i}\left(\frac{1}{m_{ij}}\sum_{y\in D_{ij}}\mathcal{L}'_{-}\left(\tau-\|Ly\|_2^2\right)Lyy^T\right)\right]\end{aligned} \tag{3.10}$$

where

$$\mathcal{L}'(z) = \frac{|\lambda|e^{-\lambda z}}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}+1}} \tag{3.11}$$

3.3 Experimental results

3.3.1 Feature extraction

Our method is not attached to a specific feature so we tested it with two types of features. We use LOMO [23], a hand-crafted feature designed for person re-identification. We also tested our approach with the more generic feature extracted with the Inception-Resnet-v2 neural network [117] that has been trained for classification tasks. We will refer to the Inception-Resnet-v2 features by the abbreviation IR.

The dimension of the LOMO and IR features are rather large, respectively 16960 and 1536. Therefore, before learning our transformation matrix, we first perform a supervised dimension reduction step using multiclass LDA (Linear Discriminant Analysis). Using all the images of the training set, but considering only cross-view image pairs (pairs of images coming from two different cameras), we compute the between class scatter matrix Σ_b and the within class scatter matrix Σ. We solve the generalized eigenvalue problem

$$\arg\max_w \frac{w^T\Sigma_b w}{w^T\Sigma w}, \tag{3.12}$$

and keep the eigenvectors w corresponding to eigenvalues that are superior to 1, ie. we keep the directions in which the between class variance is bigger than the within class variance. If there are more than 128 eigenvectors with eigenvalues bigger than 1, we only keep the 128 eigenvectors associated to the biggest eigenvalues. By projecting our initial features on the base formed by the selected eigenvectors, we obtain new features that have a dimension smaller or equal to 128.

This multiclass LDA dimension reduction step is also used in the XQDA [23] approach in which it is follow by a KISSME metric learning step [21]. We experimentally found that using XQDA [23] or multiclass LDA as a first dimension reduction step lead to similar results. This is understandable since the composition of two linear transformations is still a linear transformation, and COPREV $\circ$ XQDA which is COPREV $\circ$ KISSME $\circ$ LDA can be reduced to COPREV $\circ$ LDA. This is why throughout this chapter we will compare our method to XQDA [23].

3.3.2 Implementation details

During the training phase, we use cross-view image pairs (ie. one image comes from the probe camera, and the other from the gallery camera). Unless specified otherwise, training is performed using 200 randomly selected positive pairs per identity and 200 randomly selected negative pairs per couple of distinct identities.

We learn a square transformation matrix L by optimizing our cost function with the Matlab fminunc solver with the quasi-newton algorithm.

The value of threshold τ is fixed to 1. The presented results correspond to tests conducted with the following parameters: $\nu_+ = 4$, $\nu_- = 4$, $\lambda_+ = -4$ and $\lambda_- = -16$. Figure 3.4 shows a visualization of the loss functions.

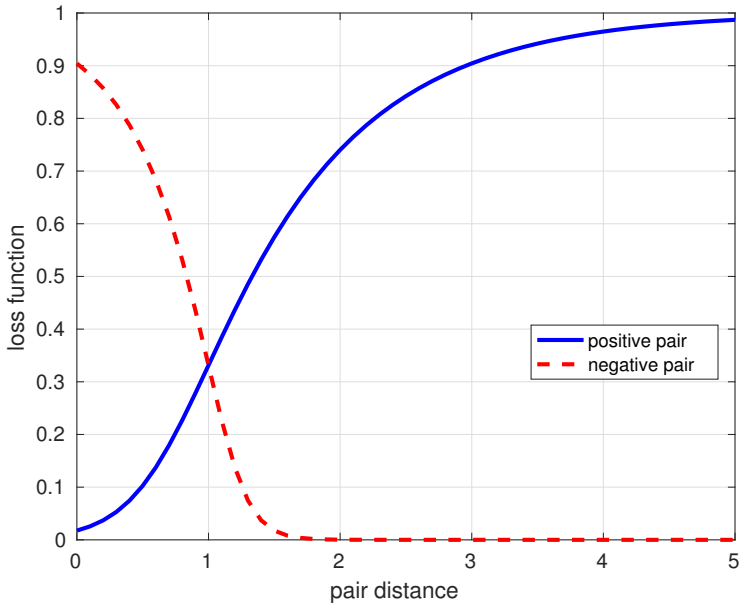


Figure 3.4 – Visualization of the loss function for positive (solid) and negative (dash) pairs.

3.3.3 Datasets and Re-ID scenarios

As far as we know, [47] is the only paper which evaluated its method on both the closed world re-id task and the type of open world re-id scenario we tackle, and does so on a multi-shot dataset. For fair comparison, we therefore chose to use the same datasets and test protocols as in [47].

iLIDS-VID [9] contains sequences of variable length, with 22 to 192 images, from 300 people. The images are captured in a busy airport by 2 cameras. The main difficulty comes from the occlusions.

PRID2011 [11] contains sequences of variable length, with 5 to 675 images, from 934 people captured by 2 cameras. Only 200 people appear in both cameras. The remaining people are captured either only by camera A or only by camera B. For this dataset, even for the closed world person task, several test protocols exist. Some papers include in the gallery some distractor identities that appear only in the gallery camera, but in most papers only the 200 common identities are used. Yet other papers, among which [47], use an even more reduced part of the dataset. Those methods are mostly based on spatio-temporal features that capture information about people's gait and which require a minimum number of images in the sequence. We follow the test protocols of [9] and [47] where only images from 178 identities that appear in both cameras with more than 21 images per sequence are used.

We adopt a **multi-shot** protocol using all the images of the sequences. During test, the distance between a probe identity and a gallery identity is chosen to be the mean of the distances of all the pairs of images from those two people.

For **closed set scenario**, we follow the test protocol from [9] using the provided 10 splits of the data into training and testing sets, each containing half of the identities, ie. 150 people for iLIDS-VID and 89 for PRID2011. Probe images are captured by camera A and gallery images comes from camera B.

For **open set scenario**, we follow the test protocol from [47] using the same 10 splits as in the closed world scenario, but the gallery set of the test set is reduced to 100 people for iLIDS-VID and to 60 people for PRID2011. One third of the probe identities are not present in the gallery. Since we use the same splits, the training is actually done once for each split, but there are separate evaluations for closed and open world scenarios.

3.3.4 Precision about the evaluations

The closed world tests are evaluated using CMC, Cumulative Matching rank Curve. For a given rank r, it represents the proportion of probe identities for whom the right match is at a rank smaller or equal to the rank r. We report CMC values at ranks $1, 5, 10, 20$.

For the open world scenario, paper [47] reports DIR (Detection and Identification Rate) values at first rank for several FAR (False Acceptance Rate) values ($1\%, 10\%, 50\%$ and 100%). In the case of iLIDS-VID, there are 50 probe imposters, so when only one of them is wrongly accepted, the FAR value already equals 2%. Therefore, what we actually report for iLIDS are the DIR values when FAR equals ($2\%, 10\%, 50\%$ and 100%). In the case of PRID2011, there are 29 probe imposters, so when only one of them is wrongly accepted, the FAR value already equals 3.4%. Therefore, what we actually report for PRID are the DIR values when FAR equals ($3.4\%, 10.3\%, 48.3\%$ and 100%).

3.3.5 Evaluation on closed world re-id scenario

In Tables 3.2 and 3.3, we report CMC values at rank 1, 5, 10 and 20, for our COPReV method tested with two features (LOMO [23] and Inception-Resnet-v2 [117]). We also report the performance of other state-of-the-art methods that were evaluated on the same closed world re-id scenario.

Dataset	PRID2011			
Rank	1	5	10	20
MDTS-DTW [47]	69.6	89.4	94.3	97.9
DVR [9]	77.4	93.9	97.0	99.4
XQDA+IR	41.2	68.9	79.9	90.2
COPReV+IR	53.0	80.8	91.5	98.1
XQDA+LOMO[23]	**86.4**	**98.3**	**99.6**	**100.0**
COPReV+LOMO	82.8	97.8	**99.6**	**100.0**

Table 3.2 – Evaluation on closed world person re-identification task. CMC value at rank 1, 5, 10, 20 for PRID2011 dataset.
Best results are in bold red.

Dataset	iLIDS-VID			
Rank	1	5	10	20
MDTS-DTW [47]	49.5	75.7	84.5	91.9
DVR [9]	51.1	75.7	83.9	90.5
XQDA+IR	11.1	29.0	39.6	51.5
COPReV+IR	21.9	51.2	66.9	81.3
XQDA+LOMO[23]	**55.9**	**83.4**	90.5	96.1
COPReV+LOMO	53.9	**83.4**	**91.6**	**97.9**

Table 3.3 – Evaluation on closed world person re-identification task. CMC value at rank 1, 5, 10, 20 for iLIDS-VID dataset.
Best results are in bold red.

Let's first compare the performance of our COPReV method with the XQDA approach [23]. The results obtained with the hand-crafted LOMO feature designed for person re-identification performs much better than the more generic IR feature, for both COPReV and XQDA approaches, which highlights the well-known fact that the choice of the features plays an important part in the re-identification performances. But of course only good features are not enough, a good matching distance is also necessary. For the IRfeature, COPReV performs clearly better than XQDA, with around 10% difference in the recognition rate at all ranks on the PRID dataset, and up to 30% on the iLIDS-VID dataset. For the LOMO feature, XQDA is slightly better at first rank, but the two methods are equivalent for the other ranks.

As for the MDTS-DTW [47] and DVR [9] approaches, they are based on spatio-temporal features and the matching involves matching video fragments rather than matching pairs of images and fusing their results. Therefore, those methods differ from ours in both the representation part and the matching part. Thus, we can not

point out what makes one method better than the other, we simply observe that for both datasets, we obtain better results than in [47, 9] by applying our COPReV transformation to the LOMO features.

3.3.6 Evaluation on open world re-id scenario

We report in Tables 3.4 and 3.5 the DIR values at first rank ($r = 1$) when FAR equals $1\%, 10\%, 50\%$ and 100%.

Dataset	PRID2011			
FAR(%)	1	10	50	100
MDTS-DTW [47]	42.7	55.2	70.5	72.8
DVR [9]	**46.8**	**58.3**	78.3	79.7
XQDA+IR	3.0	8.7	24.7	47.7
COPReV+IR	8.3	15.8	40.0	60.5
XQDA+LOMO [23]	21.0	40.5	80.3	**90.3**
COPReV+LOMO	26.5	43.5	**81.0**	87.5

Table 3.4 – Evaluation on open-world re-identification task. DIR values at rank 1 for several values of FAR (1%, 10%,50% and 100%) for PRID2011 dataset.
Best results are in bold red.

Dataset	iLIDS-VID			
FAR(%)	1	10	50	100
MDTS-DTW [47]	12.7	**32.6**	**51.8**	57.3
DVR [9]	**17.3**	29.1	49.9	57.8
XQDA+IR	0.6	2.0	8.6	13.7
COPReV+IR	1.2	5.7	17.4	25.8
XQDA+LOMO [23]	5.6	15.4	45.8	**59.9**
COPReV+LOMO	3.9	21.0	47.9	59.1

Table 3.5 – Evaluation on open-world re-identification task. DIR values at rank 1 for several values of FAR (1%, 10%,50% and 100%) for iLIDS-VID dataset.
Best results are in bold red.

Similarly to the closed world case, the DIR vs. FAR results with LOMO are better than with the IRfeatures. Though not negligible for the LOMO features, the contribution of our COPReV approach compared to XQDA is more noticeable for the IRfeatures, especially for medium and big values of FAR.

However, we must acknowledge that the DIR values at first rank when FAR equals 1% are still far from the open world results of MDTS-DTW [47] and DVR [9] approaches and we only manage to obtain similar results to those methods when FAR takes big values (50% or 100%).

3.3.7 Discussion on the evaluation measures and practical uses

With the closed world results in Tables 3.2 and 3.3 and the open world results in Tables 3.4 and 3.5, we can observe in a real case, that having good performances in terms of CMC does not guarantee to have good results in terms of DIR vs. FAR.

While CMC evaluation only assesses a relative ranking ability, the DIR vs. FAR evaluation takes into account both ranking and false acceptance rate. However, reporting DIR results only for rank 1 is not sufficient to assess the performance of an approach, the other top ranks are also important. Let's consider an extreme case where the right match is at the second rank for every probe identity. Then the DIR value at first rank is zero for any value of FAR, yet an approach that would be able to achieve such a result would be quite a good one.

Moreover, the False Acceptance Rate FAR only takes into account the most similar false match for imposter probe identities. It does not capture the difference between a situation where a probe imposter could be wrongly matched with one gallery identity or with many of them. In real applications however, we do not know in advance that the person has never been identified previously. Therefore, we would check the whole list of gallery identities who have a distance smaller than a threshold value τ before stating that the person is an imposter. The fact that the list contains only one gallery identity, or many of them makes a huge difference. The reader can take a look at the toy example in Figure 3.2 and compare the method B and C for the second open world partition.

People report DIR values for only first rank for FAR varying from 1 to 100, but we are actually more interested in DIR values for several top ranks and only for low FAR values.

Besides, since we expect to find the right match among the first r gallery identities whose distances are smaller than a threshold τ, that threshold must be chosen at some point. Based only on the DIR vs FAR evaluation, we can not determine which threshold to use. Moreover, the best threshold to base the decision on might depend on the dataset, the application, etc., and nothing in the DIR vs. FAR evaluates the difficulty to find a good decision rule.

In summary, CMC only evaluates ranking, while DIR vs. FAR is a unified measure for Detection and Ranking, but reporting DIR values at rank 1 only for FAR varying from 1 to 100 does not reflect well what we actually care about in practical use: up until which rank and up until which distance value should we look for the right match or reject the probe person?

The open world task we tackle is composed of two subtasks, the detection and the re-identification task. The re-identification part is already assessed by the CMC metric. For the detection part, it seems biased to considerer differently a negative pair where the probe identity is an imposter person to the case where the probe

identity is a known gallery person, as it is the case with the DIR vs. FAR metric. We reckon we should evaluate an open world re-id method on its ability to make the decision on whether an image pair corresponds to one single person or to two different people in an non biased way. Therefore, we propose to simply evaluate the detection part using true positive rate (TPR or recall) and true negative rate (TNR or specificity) so as to know how good we are in retrieving positive pairs and how good we are in rejecting negative pairs.

3.3.8 Evaluation on the verification task

We evaluate the verification task with TP rate and TN rate where TP rate $= \frac{TP}{TP+FN}$ and TN rate $= \frac{TN}{TN+FP}$.

We base our decision on whether a pair of images corresponds to one person or to distinct people on a simple threshold rule. That decision threshold is not a parameter to be determined after the learning phase. We use as decision threshold the threshold used at the training phase. Indeed, the whole purpose of our COPReV approach is to learn a projection so that distances of negative and positive pairs are well distributed on either side of the threshold, so it is only right that for the test phase we use the same threshold, and even if the generalization is not perfect, we still expect that positive pairs distances will be under the threshold and that negative pairs distances will be above the treshold.

In Table 3.6, we report the recall and specificity rates obtained on PRID2011 and iLIDS-VID for the two tested features. We did not report these rates for XQDA since XQDA metric learning method does not make use of a threshold at train phase and finding the best threshold to be used for verification would require a whole process of cross-validation which is precisely what we want to avoid with our method.

	PRID2011		iLIDS-VID	
	TP rate	TN rate	TP rate	TN rate
COPReV+IR	89.2	87.7	57.9	93.5
COPReV+LOMO	99.2	88.1	94.2	90.8

Table 3.6 – Evaluation for the verification task on PRID2011 and iLIDS-VID. Recall (or TP rate) and specificity (or TN rate) values are reported.

By using at test time the same threshold as in training, we achieve high true positive rate and high true negative rate (around 90%) with the LOMO features. For the IRfeatures, the results are lower. The two rates (TP rate and TN rate) are quite balanced in most of the cases, which is exactly what we aimed to obtain.

In complement to the recall and specificity values, the Table 3.7 shows the mean value of the distances of positive pairs and negative pairs. The threshold has been set to 1 and we can observe for both datasets and both features that the average

distance of positive pairs is closer to the threshold than the average distance of negative pairs. This is the case even when the TP rate is larger than the TN rate. This is because distances have 0 as lower bound, and it is difficult to have very small distances. Distances do not have upper bound, and even if from a certain value of distance on we do not enforce negative pairs to be bigger, it does not mean they can not take bigger values.

	PRID2011		iLIDS-VID	
	mean distance		mean distance	
	pos. pair	neg. pair	pos. pair	neg. pair
COPReV+IR	0.88	1.64	0.99	1.62
COPReV+LOMO	0.51	1.81	0.70	1.74

Table 3.7 – Evaluation for the verification task on PRID2011 and iLIDS-VID. Average positive pairs distances and average negative pairs distances are reported.
The decision threshold is set to 1.

3.3.9 About the initialization

Even though the objective function is not convex, experimentally, we found that under some easily verified conditions, the results were not much dependent on the initialization of the projection matrix. We tested initializing the projection matrix with several uniformly distributed random values and normally distributed random values and most of the time our cost function converge to the same value resulting in the same performances. The only point that must be taken into account is that the distance of positive and negative pairs obtained by using the initial projection matrix should be in the range of the fixed threshold. Indeed, if it is not the case, the gradient of the positive and negative pairs loss functions will probably be near zero and the optimization will not even start. If this is the case, simply multiplying the initial matrix by a relevant constant coefficient will do the trick.

However, this is true when the features have already been preprojected on the lower dimension using LDA. Without this preprojection step, the proposed COPReV method is more sensible to local minima.

3.3.10 Robustness to unbalanced data

The unbalanced data in the training set is handled thanks to the normalization coefficients which weight the importance of the loss function associated to each pair of images. To assess the relevance of that normalization, we conduct some experiments on the first partition of the iLIDS-VID dataset, keeping the same test set as previously, but using only a part of the training set to simulate unbalanced data. The new training sets are formed by randomly selecting 10 people out of the 150 training identities for whom we keep up to 20 images per camera while the remaining people are only represented by 3 images per camera. The training is computed 10 times, with 10 random selections of the identities that are overrepresented compared

to other identities. All pairs of images of the reduced training set are used during the training process. We compare the performance of our COPReV method with a variant of it where the normalization coefficients are replaced by 1. We use the LOMO features [23] as initial features.

Tables 3.8 and 3.9 are about the relevance of the positive and negative pairs distributions with respect to the fixed decision threshold $\tau = 1$. In Table 3.8 we report the mean TP rate and TN rate. In Table 3.9, we report the mean over the 10 rounds of the mean positive pairs distances and mean negative pairs distance.

	TP rate	TN rate
without normalization	50.8	100
with normalization	88.9	94.6

Table 3.8 – Recall and specificity on iLIDS-VID open world test set, first partition, for variants of our COPReV formulation, based on an unbalanced training set.

	mean positive pair distance	mean negative pair distance
without normalization	1.47	2.80
with normalization	0.83	1.88

Table 3.9 – Mean distance of positive and negative pairs on iLIDS-VID open world test set, first partition, for variants of our COPReV formulation, based on an unbalanced training set.

Without the normalization, only half of the positive pairs have a distance that is smaller than the decision threshold. Moreover, the mean value of positive pairs distance is much bigger than the threshold. The much bigger number of negative pairs is very likely to be the cause of it. With the normalization coefficients, we have balanced TP and TN rate even if the training set was unbalanced. The mean of positive pairs distance is under the threshold, and the distance of negative pairs distance is above the threshold, just as it is supposed to be.

To assess how much the unbalanced data affected the results, we compute for every pair of identities, the standard deviation of its distance over the 10 rounds. We report in Table 3.10 the mean of the standard deviation for positive pairs distances and for negative pairs distances separately. To avoid any misunderstanding, we would like to point out that this is not at all equivalent to the standard deviation of the mean of the distance of positive pairs and of negative pairs.

Given that the decision threshold is $\tau = 1$, a standard deviation of 0.3 for a pair's distance is quite a lot, but this is the case in average for the negative pairs when we do not apply any normalization coefficients. With our normalization formulation, the standard deviation for negative pairs distances is also large (0.2), but already much smaller. For the positive pairs distances, the mean standard deviation in also smaller in the case we weight the loss functions than in the case we do not weight

	mean std positive pairs	mean std negative pairs
without normalization	0.14	0.32
with normalization	0.08	0.21

Table 3.10 – Mean of the standard deviation of positive and negative pairs distances over the 10 rounds.

them. This shows that weighting the loss functions helps lessen the sensibility to unbalanced data.

3.4 Conclusion

In this chapter, we discussed about the person re-identification task, and more specifically the difference between the closed and open world re-identification tasks. We highlighted the fact that ranking methods alone are not sufficient for tackling open world scenarios so we proposed to adopt a verification perspective. Through the optimization of our cost function, COPReV learns a linear transformation of the space such that the number of misclassified pairs is minimized while not favoring negative pairs nor largely represented identities. Rather than reasoning on the distances of misclassified pairs, or directly on the number of misclassified pairs, we reasoned on the proportions of misclassified positive and negative pairs. This enabled our method to be less sensitive to unbalanced training data. For both tested datasets, and for both closed and open world evaluations, COPReV improved the performances for the different tested features. However, while our closed world performances are at the state-of-the-art, for the open world case, DVR [9] and MDTS-DTWt [47] methods still outperform our approach. This could be due to the features because both DVR [9] and MDTS-DTWt [47] use time space features. However, we reckon that casting the re-identification task only as a verification problem without any ranking constraints might be one of the main reasons why the performances of COPReV are mitigated.

Chapter 4

Sparse representations with enhanced collaboration

In the previous chapter, the presented COPReV method focuses solely on the decision that needs to be taken about the presence or absence of the probe person in the gallery set. If a perfect decision method existed, it would make it unnecessary to also integrate ranking constraints. However this is not the case. For the open world re-identification task, focusing only on the decision aspect without taking ranking aspects into account is not enough. A good method should consider both decision and ranking aspects. In that regard, collaborative sparse coding seems to be a perfect tool. On one hand, by its collaborative aspect, collaborative sparse coding integrates ranking considerations. On the other hand, the sparsity aspect could be exploited to manage the decision part of the open world re-identification task. However, would this really be sufficient to tackle the open world re-identification task? Does collaboration manage absolute ranking or only relative ranking? Will the sparsity of representations enable to reject most of the wrong matches or solely a small part of obvious wrong matches?

In addition to its relevance for the open world task, another advantage of sparse representations is its relevance for tackling multi-shot scenarios. Indeed, sparse representation allows to compare in one go a probe image with all the images of a gallery person so that two people can be deemed similar even if not all their images are alike but only a few of them match, for example those where they appear with the same pose.

After introducing in the first section the notations and important notions needed for the rest of the chapter, the second section of this chapter studies the difference between collaborative and non collaborative sparse coding when applied to the person re-identification task in order to make the most of both the collaborative and the sparsity aspects of collaborative sparse coding. Based on the observations made in the second section, the third section presents a collaborative sparse coding approach designed for the open world re-identificaton task where the collaboration is enhanced. The last section is dedicated to the analysis of the experimental results on closed and open world re-identification tasks and on person verification tasks.

4.1 Preliminaries

4.1.1 Notations: training and testing data

Let T_g and T_p be the matrices corresponding to the concatenation of column features extracted from training images respectively from the gallery camera and from the probe camera. For our sparse coding approach, the identity associated to each training image does not matter, therefore unlabelled training images can be exploited.

As for the testing set, let $\mathcal{K}$ be the set of gallery identities and $\mathcal{L}$ be the set of probe identities. In a closed world setting, $card(\mathcal{K}) = card(\mathcal{L}) = C$ and we could use $c \in [1, \dots, C]$ to refer to each person's identity. In an open world setting, there are common identities ($\mathcal{K} \cap \mathcal{L} \neq \emptyset$) but the two sets of identities are not equal ($\mathcal{K} \neq \mathcal{L}$). We call probe imposter an identity who is in the probe set but not in the gallery set ($\mathcal{L} \setminus (\mathcal{K} \cap \mathcal{L})$) . A distractor refers to a person who is present in the gallery set but not in the probe set, ie. a person who is in $\mathcal{K} \setminus (\mathcal{K} \cap \mathcal{L})$.

Let $K = card(\mathcal{K})$ and $L = card(\mathcal{L})$ be the number of identities respectively in the gallery set and in the probe set. The set of gallery identities is given by $\mathcal{K} = \{k_1, \dots, k_K\} = \{k_i\}_{i \in [1,K]}$ and the set of probe identities is described by $\mathcal{L} = \{l_1, \dots, l_L\} = \{l_j\}_{j \in [1,L]}$.

In a multi-shot dataset, every probe and gallery person are represented by several images and the number of images available for each person varies. Let n_{k_i} be the number of images of gallery person k_i and m_{l_j} be the number of images of probe person l_j. Each image is described by a single feature column vector of length d. We refer by the letter g a feature from a gallery image and by the letter p a feature from a probe image. The matrix $G_{k_i} = [g_{k_i,1}, \dots, g_{k_i,n_{k_i}}]$ is the gallery person k_i's dictionary, it is the concatenation of all its features. The concatenation of all the gallery features forms the gallery dictionary $G = [G_{k_1}, \dots, G_{k_K}]$. Similarly the matrix $P_{l_j} = [p_{l_j,1}, \dots, p_{l_j,m_{l_j}}]$ is the concatenation of the features of all the images of probe person l_j.

To lighten the notations, when there is no ambiguity, we drop the i and j subscripts, and refer to a probe identity by the letter l and refer to a gallery identity by the letter k.

4.1.2 Notations: sparse coding

In section 2.3.4 we have seen that a sparse representation depends of the dictionary with which the sparse representation has been computed and on the choice of the sparsity penalization function. In the rest of the thesis, we will always specify the dictionary with which a sparse code has been computed and we adopt the $\mathcal{L}1$ norm

as sparsity penalization function.

For a column vector x of dimension d and a dictionary D of dimension $d \times u$, a sparse representation of x using dictionary D is a column vector $a_{x,D}$ of dimension u solution of the following Lasso problem:

$$a_{x,D} = \arg\min_{a} ||x - Da||_2^2 + \lambda||a||_1. \tag{4.1}$$

where λ is a parameter with a value between 0 and 1.

Let's consider that the dictionary D is the concatenation of Q subdictionaries, ie.

$$D = [D_1, D_2, \ldots, D_Q] \tag{4.2}$$

where D_q for $q \in [1, Q]$ is of dimension $d \times v_q$, and $\sum_{q=1}^{Q} v_q = u$.

The sparse representation $a_{x,D}$ can be decomposed accordingly to the subdictionaries into:

$$a_{x,D} = \begin{bmatrix} a_{x,D,D_1} \\ a_{x,D,D_2} \\ \vdots \\ a_{x,D,D_Q} \end{bmatrix} \tag{4.3}$$

where a_{x,D,D_q} is a column vector of dimension v_q.

Since $a_{x,D}$ is a sparse representation of x computed with the dictionary D, we have an approximation of x by a sparse linear combination of dictionary elements (columns of D) :

$$x \approx Da_{x,D} = [D_1, D_2, \ldots, D_Q] \begin{bmatrix} a_{x,D,D_1} \\ a_{x,D,D_2} \\ \vdots \\ a_{x,D,D_Q} \end{bmatrix} \tag{4.4}$$

The reconstruction error of x using the dictionary D is defined by:

$$e_{x,D} = ||x - Da_{x,D}||_2^2. \tag{4.5}$$

The residual reconstruction error of x using only the elements from dictionary D_q for $q \in [1, Q]$ while the sparse representation $a_{x,D}$ of x has been computed with dictionary D is given by:

$$r_{x,D,D_q} = ||x - D_q a_{x,D,D_q}||_2^2 \tag{4.6}$$

In our problem, we compute the sparse representation of many elements, so we prefer to adopt a matrix formulation. Let's suppose we are given N column vectors $x_1, x_2, \ldots, x_N$ and we want to compute their sparse representations $a_{x_1,D}, a_{x_2,D}, \ldots, a_{x_N,D}$ using dictionary D. It can be done by solving the Lasso problem of Eq 4.1 for each of the column vectors. It is equivalent to consider the matrix X of dimension $d \times N$,

concatenation of the N column vectors $X = [x_1, x_2, \ldots, x_N]$ and solve the following Lasso problem:

$$A_{X,D} = \arg\min_{A} ||X - DA||_F^2 + \lambda||A||_1 \tag{4.7}$$

where $||.||_F$ is the Frobenius norm. The dimension of $A_{X,D}$ is $u \times N$. The n^{th} column of $A_{X,D}$ corresponds to the sparse code of the n^{th} column of X. Each row of $A_{X,D}$ corresponds to the weight of the participation of a column of D for the reconstruction of X. $A_{X,D}$ can be decomposed into:

$$A_{X,D} = [a_{x_1,D}, a_{x_2,D}, \ldots, a_{x_N,D}] = \begin{bmatrix} a_{x_1,D_1}, a_{x_2,D_1}, \ldots, a_{x_N,D_1} \\ a_{x_1,D_2}, a_{x_2,D_2}, \ldots, a_{x_N,D_2} \\ \vdots \\ a_{x_1,D_Q}, a_{x_2,D_Q}, \ldots, a_{x_N,D_Q} \end{bmatrix} \tag{4.8}$$

a_{x_n,D_q} is the submatrix of $A_{X,D}$ containing the weight coefficients corresponding to the participation of subdictionary D_q to the reconstruction of elements x_n.

The reconstruction error of X using dictionary D is the sum of the reconstruction error of each column of X:

$$e_{X,D} = ||X - DA_{X,D}||_F^2 \tag{4.9}$$

The mean reconstruction error of X using dictionary D is the mean of the reconstruction error of each column of X, it is given by:

$$E_{X,D} = \frac{||X - DA_{X,D}||_F^2}{N} \tag{4.10}$$

The residual reconstruction error of X using only the elements from dictionary D_q for $q \in [1, Q]$ while the sparse representation $A_{X,D}$ of X has been computed with dictionary D is given by:

$$r_{X,D,D_q} = ||X - D_q A_{X,D,D_q}||_F^2 \tag{4.11}$$

The mean residual reconstruction error of X using only the elements from dictionary D_q for $q \in [1, Q]$ while the sparse representation $A_{X,D}$ of X has been computed with dictionary D is given by:

$$R_{X,D,D_q} = \frac{||X - D_q A_{X,D,D_q}||_F^2}{N} \tag{4.12}$$

The reader should simply remember that when there are 2 or 3 subscripts, the first subscript refers to the element whose sparse code has been computed, the second subscript refers to the dictionary used for computing the sparse representation and the third subscript refers to a subdictionary of the dictionary used for computing the sparse code.

4.1.3 Features prerequisites

The method we propose is not specific to a given type of features, but we require to have one descriptor per image and every descriptor must have a unit $\mathcal{L}2$ norm.

If the chosen descriptor is not designed as such, one simply needs to normalize every descriptor before applying our sparse representation method. This feature normalization step is necessary and is an important point of this method. Firstly, the features that compose a dictionary should have the same norm so that there is a fair competition between each element. Indeed, suppose the norm of one feature is equals to $C > 1$ times another one. Then the one with a bigger norm will always be preferred to the smaller one because for the same decrease in the reconstruction error term, the sparsity penalization term will be C times smaller. Secondly, the features for which we compute the sparse representation should also have the same norm otherwise we should modify the λ parameter that balances the importance given to the reconstruction error and the sparsity accordingly to the norm of the feature to be reconstructed.

4.2 Collaborative versus non collaborative sparse coding

4.2.1 Non collaborative sparse coding of probe elements

A non collaborative sparse coding method for re-identification requires the computation of a sparse code for each of the probe identities, using independently each of the gallery identities's specific dictionary.

When given a probe person l's features P_l, we compute the sparse representation $A_{P_l,Gk}$ of P_l with respect to the gallery identity k using gallery dictionary G_k by solving the following Lasso optimization problem:

$$A_{P_l,G_k} = \arg\min_A ||P_l - G_k A||_F^2 + \lambda ||A||_1 \tag{4.13}$$

where $||.||_F$ is the Frobenius norm, $||.||_1$ is the $\mathcal{L}_1$ norm and λ is a trade-off parameter between the reconstruction error and the sparsity penalization term, and its value is chosen between 0 and 1.

The dissimilarity score $s(l,k)$ between probe person l and gallery person k is defined as the mean reconstruction error of probe elements using gallery dictionary G_k:

$$s(l,k) = E_{P_l,G_k} = \frac{||P_l - G_k A_{P_l,G_k}||_F^2}{m_l} \tag{4.14}$$

The gallery identities are ranked from the most similar ones to the least similar ones by increasing reconstruction errors.

In this thesis, we name this non collaborative sparse coding method Lasso DNC. Lasso refers to the fact that the sparse representations are computed by solving a Lasso problem. D stands for Direct. There is a need to specify this because in the next chapter there will be a reverse approach. NC stands for Non Collaborative.

4.2.2 Collaborative sparse coding of probe elements

In a collaborative sparse coding approach for person re-identification, only one sparse representation is needed and it involves all gallery elements at the same time.

Given a probe person's features P_l we compute its sparse representation $A_{P_l,G}$ using the collaborative gallery dictionary G, concatenation of all gallery dictionaries, by solving the following Lasso optimization problem:

$$A_{P_l,G} = \arg\min_A ||P_l - GA||_F^2 + \lambda||A||_1 \tag{4.15}$$

Since $G = [G_1, \ldots, G_K]$, $A_{P_l,G}$ can be written as

$$A_{P_l,G} = \begin{bmatrix} A_{P_l,G,G_1} \\ A_{P_l,G,G_2} \\ \vdots \\ A_{P_l,G,G_K} \end{bmatrix} \tag{4.16}$$

where A_{P_l,G,G_k} is the sparse submatrix of $A_{P_l,G}$ which describes the participation of dictionary elements G_k. For $k \in [1, K]$, the dimension of $A_{Pl,G,Gk}$ is of $n_k \times m_l$.

The dissimilarity score $s(l, k)$ between probe person l and gallery person k is defined as the mean residual reconstruction error of the reconstruction of probe features P_l using only the elements from gallery dictionary G_k given the sparse representation computed with the dictionary D:

$$s(l, k) = R_{P_l,G,G_k} = \frac{||P_l - G_k A_{P_l,G,G_k}||_F^2}{m_l} \tag{4.17}$$

Gallery identities are ranked by increasing dissimilarity score, ie. the smaller the residual error, the most likely the right match.

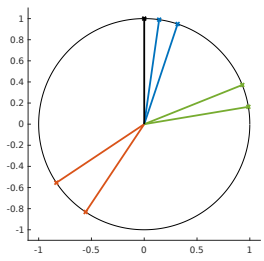
We call Lasso DC this approach where DC stands for Direct Collaborative.

4.2.3 Comparison of non collaborative and collaborative sparse coding

We have presented how non collaborative and collaborative sparse coding could be used for the person re-identification task. In reality, no paper has proposed the non collaborative sparse coding approach for tackling person re-identification. The aim of this section is to point out what exactly makes the collaborative sparse coding approach much more suitable for closed world person re-identification than the non collaborative approach and to stress what is still lacking to tackle the open world case.

In order to compute the sparse representation of probe features, in both the collaborative and the non collaborative cases, we optimize Lasso problems, respectively Eq. 4.15 and Eq. 4.13. In the Lasso problem, the cost function is the sum of two terms, the reconstruction error term and the sparsity term. Even if the sparsity

term penalizes reconstructions that involve many dictionary elements, it is very unlikely that the sparse matrix will be equal to the null matrix, some elements will be selected for reducing the reconstruction error until a balance is found between the two terms. In Figure 4.1 a two dimensions toy example is given.



$$P_l = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$G_1 = \begin{bmatrix} 1/\sqrt{50} & 1/\sqrt{10} \\ 7/\sqrt{50} & 3/\sqrt{10} \end{bmatrix}$$

$$G_2 = \begin{bmatrix} -2/\sqrt{13} & -3/\sqrt{13} \\ -3/\sqrt{13} & -2/\sqrt{13} \end{bmatrix}$$

$$G_3 = \begin{bmatrix} 6/\sqrt{37} & 5/\sqrt{29} \\ 1/\sqrt{37} & 2/\sqrt{29} \end{bmatrix}$$

Lasso optimization equations 4.13 and 4.15 with $\lambda = 0.2$ give the following sparse representations, reconstruction errors and residual errors:

	Lasso DNC			Lasso DC		
k	A_{P_l,G_k}		E_{P_l,G_k}	A_{P_l,G,G_k}		R_{P_l,G,G_k}
1	[0.79	0.00]	0.25	[0.79	0.00]	0.25
2	[-0.63	0.00]	0.59	[0.00	0.00]	1.00
3	[0.00	0.17]	0.95	[0.00	0.00]	1.00

Figure 4.1 – Comparison of non collaborative (Lasso DCN) and collaborative sparse coding (Lasso DC) approaches on a toy example.
The dimension of the features is $d = 2$. The probe person is represented by only one feature (black). There are 3 gallery identities (blue, red, green), each of them are representated by 2 features. The sparse representations obtained by the Lasso DCN and the Lasso DC approaches are presented in the table, as well as the reconstruction errors (for Lasso DCN) and the residual errors (for Lasso DC). In the Lasso DC case gallery people 2 and 3 do not particpate in the reconstruction of probe element and only gallery person 1 who is the most similar to the probe person participates. Therefore the residual errors for people 2 and 3 are both equal to 1. In the Lasso DNC case, however, even if the green vectors are almost perpendicular to the black vector (gallery person 3 is very dissimilar to the probe person), the non collaborative sparse code for gallery person 3 is not null and even though the reconstruction error is not small, it is not null either.

For the sake of the example, let's consider that the probe person is represented by only one image, so we only compute the sparse representation of one probe feature with unit $\mathcal{L}2$ norm. In the rare case the sparse representation is a null vector, the cost function equals the reconstruction error which is in this case also the norm of the probe feature, ie. it equals 1. When the sparse representation is not null, the cost function must therefore be equal or smaller to 1, which also means that the reconstruction error is strictly smaller than 1 because the sparsity term is no longer equal to zero.

In the non collaborative approach, the dissimilarity score is the reconstruction error value. Therefore the dissimilarity score will hardly ever be equal to 1, but most of the time it will be strictly smaller than 1. Even if some participate more than others, every gallery person does participate in the reconstruction of the probe feature, and there is no clear distinction between the reconstruction error value of a gallery person who is not at all similar to the probe person and a gallery person who is somehow similar to the probe person. Nonetheless, the reconstruction error still enables to rank gallery identities because a gallery person who is more similar to the probe person should still have a smaller reconstruction error than a less similar gallery person.

In the non collaborative approach, the sparsity does not play an important role

because every gallery person does participate in the reconstruction of the probe feature anyway. In the collaborative approach however, the sparsity plays an important role. Because the matrix $A_{P_l,D}$ is sparse, some submatrices A_{P_l,D,D_k} are null matrices. This means that some gallery identities do not participate in the reconstruction of the probe feature. All these gallery identities have the same dissimilarity score which equals 1. Therefore they are either not ranked at all or they are all given the same rank. This issue is tackled in the paper [68] which proposes to use iterative sparse coding to rank all gallery identities. We argue that since these gallery identities do not participate in the reconstruction of probe features, it means they are not similar enough to the probe person and are therefore obvious wrong matches. The sparsity aspect of collaborative sparse coding enables to eliminate from the gallery list the gallery identities that are the less likely to be the probe person.

As for the collaborative aspect, the collaborative reconstruction puts into competition elements from different identities. Elements that are selected for the reconstructions are the ones that are more similar to the probe person than the other elements. The more a gallery person's dictionary participates in the reconstruction, the more likely he is to be the right match. The gallery identities who did participate in the reconstruction of the probe feature are ranked by increasing residual error.

However, we must emphasize on the fact that such collaborative representations only allow for a relative ranking and nothing ensures that the gallery person who has the smallest residual error is actually similar to the probe person. At this point, we must distinguish the closed world and the open world re-identification cases. In the closed world case, the probe person is present in the gallery set. Therefore, the right match will certainly participate a lot in the reconstruction of probe elements, and as a consequence the participation of other gallery identities will be greatly reduced. This will be the case for every probe person, so everything goes well. In the open world case, the probe person might not be in the gallery set, which makes a huge difference. If the presented probe person is an imposter probe person, even if he is not present in the gallery, there will still be some gallery elements that will be selected so that the reconstruction error decreases. We are back to a case similar to the non collaborative sparse coding method: a probe person's features must be reconstructed with gallery elements none of which are actually related to the probe person.

Moreover, even if the sparsity enables to find gallery identities who are clearly dissimilar to the probe person, we can not reject a probe person and consider him as an imposter probe person who has no right match among the gallery identities. Indeed, there will always be some gallery elements who will participate in the reconstruction of the probe person's features, so we won't obtain a residual error equal to 1 for every gallery identity.

4.3 Collaboration enhanced sparse coding for open world person re-identification

4.3.1 Enhanced collaboration for open world re-identification

For tackling the open world re-identification, relative ranking is not sufficient, we must also reject imposter probe people. In order to reject a probe person, the dissimilarity score (or mean residual error) of every gallery person should be big, equal to 1, or close to it. It means that no gallery person should participate in the reconstruction of the probe features, or at least not much. However, we have seen that a sparse representation matrix is hardly ever null, it is simply sparse. We therefore propose to supplement the collaborative gallery dictionary with an additional dictionary D which aim is to relieve the participation of the gallery dictionaries for the reconstruction of imposter probe people's features.

The optimization problem is similar to Eq 4.15, but there is now an additional dictionary D and the sparse representation $A_{P_l,[G,D]}$ of probe features P_l using the collaborative dictionary $[G, D]$, concatenation of gallery dictionary G and the additional dictionary D, is computed by optimizing the following Lasso problem:

$$A_{P_l,[G,D]} = \arg\min_A ||P_l - [G, D]A||_F^2 + \lambda||A||_1 \tag{4.18}$$

The dissimilarity score $s(l, k)$ between probe person l and gallery person k is defined as previously by the mean residual reconstruction error of the reconstruction of probe features P_l using only the elements from gallery dictionary G_k, but the sparse representation is the one computed with dictionary $[G, D]$:

$$s(l, k) = R_{P_l,[G,D],G_k} = \frac{||P_l - G_k A_{P_l,[G,D],G_k}||_F^2}{m_l} \tag{4.19}$$

The gallery identities are ranked by increasing dissimilarity score, and for a probe person l the best match is given by:

$$k^* = \arg\min_k R_{P_l,[G,D],G_k} \tag{4.20}$$

Notice that the additional dictionary D is only used for the computation of the sparse representations. The residual error corresponding to the additional dictionary D is not exploited. Therefore the nature of D is not important. For example, D could be the result of a dictionary learning process, a clustering process, etc. The columns of D could also correspond to features extracted from real images and the identity of the people represented in those images would not matter, so unlabelled data can be used.

We refer to this collaboration enhanced sparse representation approach by Lasso DCE where D stands for Direct, and CE for Collaboration Enhanced. The Figure 4.2 shows an overview of this Lasso DCE approach.

4.3.2 Additional dictionary D

The only data available for us to construct the additional dictionary is the training data, and we propose here a simple selection process to form the additional dictionary.

The first point we would like to raise attention to is the camera from which the images are captured. Maybe in a few years this point won't be so relevant anymore, but at this stage of development of person re-identification features, even the best image descriptors and metric learning methods, are not yet capable of providing features or transformed features which really are invariant to camera views. Images taken from the same camera have the same color rendering and it generally shows people with similar pose and viewpoint which is still reflected in the final features. Therefore, intraclass variability between images from a given person captured by gallery and probe cameras can be higher than the interclass variability between images from different people all captured with the same camera. We wish to have a fair competition between identities, so since the elements from the additional dictionary will be competing against the collaborative gallery dictionaries (with features extracted from images captured by the gallery camera), they should also come from the gallery camera. Therefore, in our method, we do not use training data from the probe camera, we solely use training data from the gallery camera.

The additional dictionary is essentially meant to help avoid false matches by reducing the participation of gallery identities in the reconstruction of imposter probe elements. However, we know nothing about the imposter probe people. Probe imposters might be similar to some gallery identities. They might also be dissimilar to every gallery people. In order to reject a maximum number of probe imposters, we include in the additional dictionary all the training data available coming from the gallery camera. While this additional dictionary should enable to reject more probe imposters, the re-identification of non imposter probe people should not be affected. Indeed, collaborative sparse coding selects the elements that are the most similar to the elements to be reconstructed. If there is a match among the gallery people for the probe person, the elements from the dictionary of this gallery person should be the most similar to the features of the probe person so they should be among the few selected elements, even if that gallery dictionary competes with elements from an extended collaborative dictionary.

In the end, the additional dictionary we propose for our collaboration enhanced sparse coding approach is simply composed of all the features from the gallery camera images of the training set: $D = T_g$.

4.3.3 A method also relevant for person verification

In the person verification task, the goal is to determine whether a pair of identities are the same people or distinct people. In the COPReV method, we had to specify a decision rule that was used during the training phase and re-used at test time. With our collaborative sparse coding approach, 1 is a natural decision threshold. Indeed,

features are normalized to unit $\mathcal{L}2$ norm. We have seen that when gallery identities do not participate in the reconstruction of the probe person, their dissimilarity scores equal 1. When gallery identities do participate in the reconstruction of the probe features, most of the time the dissimilarity score is strictly smaller than 1 although it might happen that it is bigger than 1.

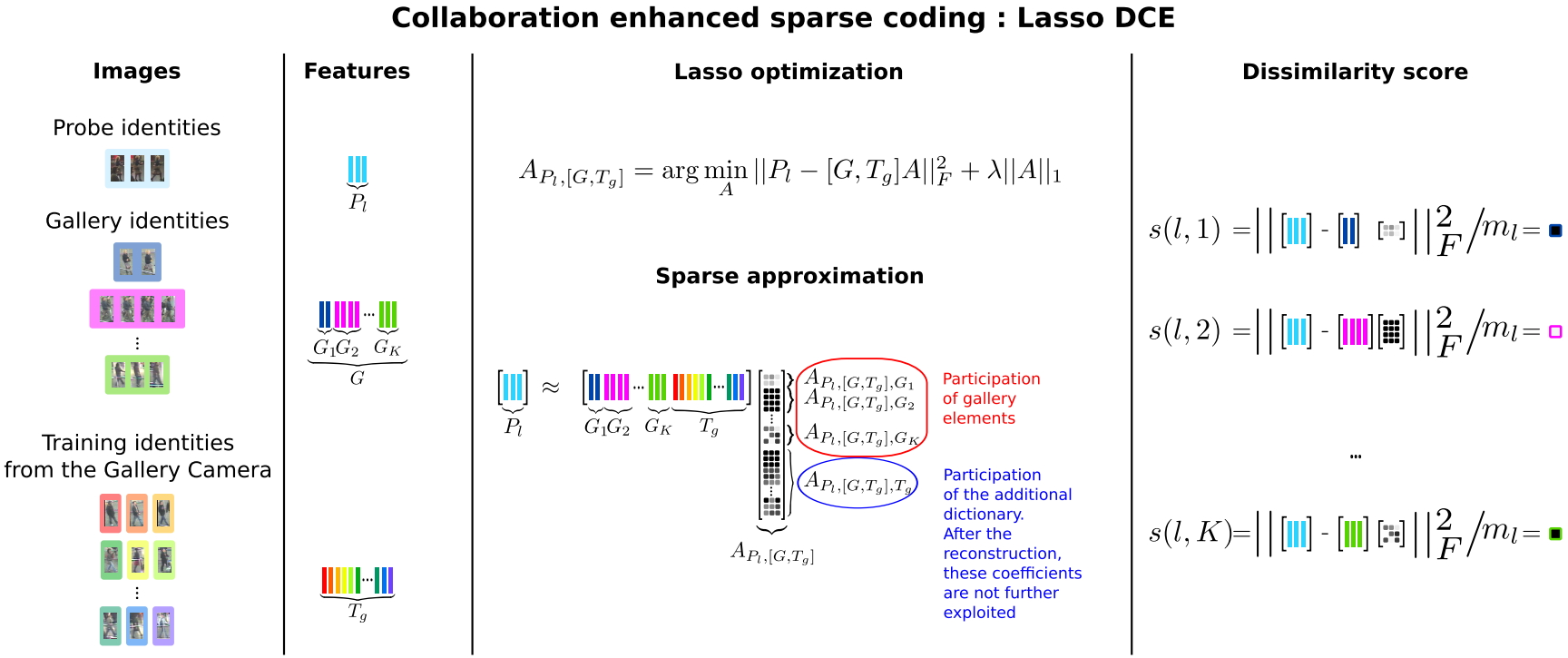


Figure 4.2 – Overview of Lasso DCE (Direct Collaboration Enhanced) approach.
In multi-shot datasets, probe and gallery identities are represented by several images and the number of images can vary with the person. m_l is the number of images of probe person l. In this example it is equal to 3. A column feature is computed for each image. Probe person's features are approximated by a sparse linear combination of gallery features and additional features corresponding to gallery camera images taken from the training set. The sparse representation is found through the optimization of a Lasso problem. Sparse matrices and dissimilarity scores (mean residual errors) are in gray scale, the darker the square, the smaller the value.

4.3.4 About the exploitation of multi-shot data

In this collaborative sparse coding Lasso DCE method, we exploit the multi-shot aspect of probe data in the usual way by computing for each probe person, the mean of the dissimilarity score of each of its images with regards to each gallery person.

The multi-shot aspect of gallery data is better exploited. When comparing a probe person's image to a gallery person, instead of comparing the probe image to each of the gallery images separately, all of that gallery person's images are considered jointly. The residual error of a probe feature is computed by gallery identity rather than by gallery image.

4.4 Experimental results

4.4.1 Implementation details and feature extraction

Optimization. All the optimization problems we consider are $\mathcal{L}_1$-norm minimization problems which can be be solved using proximal algorithms. We used the mexLasso function of the SPAMS library [1].

Parameters. The parameter λ has been set to 0.2 in all our experiments.

Features. We use as our base feature, LOMO features [23] that are already transformed by a metric learning algorithm and a normalization step. Images are resized to 128×64 pixels and we extract LOMO features of dimension 26960. All the training images are used to learn the XQDA [23] matrix M which is symmetric definite positive and which can thus be decomposed into $M = L^T L$. To obtain our final features LOMO_{pn} where the subcript pn stands for projected and normalized, we project LOMO features using the projection matrix L into a lower dimensional space and then normalize the projected features to unit $\mathcal{L}_2$-norm.

Since our approach is not specific to a given feature, the main tests have also been conducted with other features. We use generic features that have not specifically been designed for the person re-identification task, they are extracted from a well-known neural network, the Inception-Resnet-v2 network [117] which has been trained on the ImageNet dataset for classification tasks. In the tables we will refer to this feature by the abbreviation IR for Inception-Resnet feature. We tested our approach with $\mathcal{L}_2$-norm normalized XQDA projected IR features (IR_{pn}) and with $\mathcal{L}_2$-norm normalized IR features (IR_n) where the subscript n stands for normalized.

4.4.2 Datasets, training and testing sets, testing protocols, evaluation

Please refer to section 3.3.3.

[1] spams-devel.gforge.inria.fr/downloads.html

4.4.3 Evaluation on closed and open world re-identification tasks

In order to assess the relevance of collaborative sparse coding for the re-identification task and the relevance of enhanced collaboration for the open world case, we conduct the tests with differents features to show the improvement they bring regardless of the features. The closed world CMC performances are reported in Tables 4.1 and 4.2. The open world DIR vs FAR evaluations are reported in Tables 4.3 and 4.4.

	Rank 1	Rank 5	Rank 10	Rank 20
MDTS-DTW [47]	49.5	75.7	84.5	91.9
DVR [9]	51.1	75.7	83.9	90.5
IR + Eucl. dist.	2.8	8.7	14.3	22.6
IR_n + Lasso DNC	2.1	8.8	15.0	26.3
IR_n + Lasso DC	2.9	10.9	18.3	31.5
IR_n + Lasso DCE	2.7	11.4	18.6	32.1
IR + XQDA	11.7	30.9	41.5	53.6
IR_{pn} + Lasso DNC	31.7	52.9	64.9	76.5
IR_{pn} + Lasso DC	40.7	66.6	77.1	85.9
IR_{pn} + Lasso DCE	40.1	65.1	76.2	85.9
LOMO + XQDA	55.3	83.1	90.3	**96.3**
$LOMO_{pn}$ + Lasso DNC	56.1	81.9	88.5	94.5
$LOMO_{pn}$ + Lasso DC	64.9	**87.1**	**92.5**	96.1
$LOMO_{pn}$ + Lasso DCE	**65.1**	86.6	92.4	96.1

Table 4.1 – Evaluation on closed world re-identification task. CMC value at rank $1, 5, 10$ and 20 for iLIDS-VID dataset.
Best results are in bold red.

	Rank 1	Rank 5	Rank 10	Rank 20
MDTS-DTW [47]	69.6	89.4	94.3	97.9
DVR [9]	77.4	93.9	97.0	99.4
IR + Eucl. dist.	14.4	39.2	52.7	70.0
IR_n + Lasso DNC	26.5	53.3	67.2	78.2
IR_n + Lasso DC	30.6	55.4	69.7	81.3
IR_n + Lasso DCE	29.0	55.5	70.2	80.9
IR + XQDA	43.4	71.9	82.4	91.6
IR_{pn} + Lasso DNC	68.5	89.7	94.7	97.4
IR_{pn} + Lasso DC	70.7	90.0	96.1	98.3
IR_{pn} + Lasso DCE	72.0	89.9	95.3	98.1
LOMO + XQDA	86.3	**98.3**	**99.6**	**100.0**
$LOMO_{pn}$ + Lasso DNC	87.3	98.2	**99.6**	**100.0**
$LOMO_{pn}$ + Lasso DC	90.2	98.0	99.3	**100.0**
$LOMO_{pn}$ + Lasso DCE	**90.6**	97.9	99.2	**100.0**

Table 4.2 – Evaluation on closed world re-identification task. CMC value at rank $1, 5, 10$ and 20 for PRID2011 dataset.
Best results are in bold red.

FAR(%)	1	10	50	100
MDTS-DTW [47]	12.7	32.6	51.8	57.3
DVR [9]	**17.3**	29.1	49.9	57.8
IR + Eucl. Dist.	0.4	0.9	1.7	3.0
IR_n + Lasso DNC	0.1	0.8	1.7	3.3
IR_n + Lasso DC	0.1	0.8	2.2	3.0
IR_n + Lasso DCE	0.4	0.9	2.4	3.4
IR + XQDA[23]	0.6	2.2	8.4	14.9
IR_{pn} + Lasso DNC	2.8	8.9	27.4	35.7
IR_{pn} + Lasso DC	7.3	16.7	36.4	46.0
IR_{pn} + Lasso DCE	7.3	18.6	37.7	44.9
LOMO + XQDA[23]	5.1	15.2	45.3	59.1
$LOMO_{pn}$ + Lasso DNC	4.5	16.6	45.7	61.5
$LOMO_{pn}$ + Lasso DC	12.9	35.1	58.8	68.5
$LOMO_{pn}$ + Lasso DCE	17.2	**37.5**	**62.8**	**69.0**

Table 4.3 – Evaluation on open world re-identification task. DIR at first rank for several FAR values ($1\%, 10\%, 50\%$ and 100%) on iLIDS-VID dataset.
Best results are in bold red.

Relevance of sparse coding

For both datasets, for both closed and open world evaluation, using non collaborative sparse coding (Lasso DNC) for the matching step instead of XQDA metric learning leads to much better results in the case of IR features and to similar results in the case of LOMO features. Collaborative sparse coding approaches (Lasso DC and Lasso DCE) perform even better than the non collaborative approach. This shows the relevance of using sparse coding for the matching step of person re-identification.

Relevance of collaborative sparse coding

For closed world scenarios, on the iLIDS-VID dataset, collaborative sparse representation approaches (Lasso DC and Lasso DCE) bring almost a 10% improvement with a first rank recognition rate for LOMO features of 55.3% for XQDA and respectively 64.9% for Lasso DC and 65.4 for Lasso DCE. On PRID2011 dataset, the first rank recognition rate is already very high, 86.3% for XQDA, so the improvement is a little bit smaller, but still significant (+3.9% for Lasso DC and +4.3% for Lasso DCE). With the IRfeatures projected with XQDA, the gap is even more impressive with an increase of around +30% of the first rank recognition rate on the iLIDS-VID dataset, but XQDA results start lower with IRfeatures.

For open world scenarios, the improvement brought by collaborative sparse coding is also considerable. For both iLIDS-VID and PRID2011 datasets, when using the LOMO features, the first rank Detection and Identification Rate is more than doubled for the lowest non null False Acceptance Rate. On iLIDS-VID, we improve the DIR value at first rank by more than 7% and on PRID we raise it by more than 28%.

FAR(%)	1	10	50	100
MDTS-DTW [47]	42.7	55.2	70.5	72.8
DVR [9]	46.8	58.3	78.3	79.7
IR + Eucl. Dist.	2.2	5.5	9.8	19.3
IR_n + Lasso DNC	3.8	4.3	16.0	31.2
IR_n + Lasso DC	8.8	13.7	26.8	35.8
IR_n + Lasso DCE	8.7	14.5	27.7	34.5
IR + XQDA[23]	3.2	9.5	25.7	50.8
IR_{pn} + Lasso DNC	12.5	29.5	61.2	73.0
IR_{pn} + Lasso DC	21.0	44.8	66.2	75.2
IR_{pn} + Lasso DCE	28.0	46.8	69.5	76.3
LOMO + XQDA[23]	21.2	40.7	78.8	90.5
$LOMO_{pn}$ + Lasso DNC	21.2	39.2	74.5	90.2
$LOMO_{pn}$ + Lasso DC	49.8	69.3	88.2	**93.8**
$LOMO_{pn}$ + Lasso DCE	**55.7**	**71.0**	**90.2**	93.2

Table 4.4 – Evaluation on open world re-identification task. DIR at first rank for several FAR values (1%, 10%, 50% and 100%) on PRID2011 dataset.
Best results are in bold red.

Using collaborative sparse coding (Lasso DC or Lasso DCE) on normalized XQDA projected features greatly improves the performances compared to using XQDA only as a Mahalanobis metric. This is true for both LOMO and IRfeatures and for both closed world CMC evaluation and DIR vs FAR open world evaluation. An exception when collaborative sparse coding does not improve recognition rates, is on the iLIDS-VID dataset, with the non projected IRfeatures, where the performances are very low for both Euclidean distance matching and collaborative sparse representation matching.

Relevance of collaboration enhanced sparse coding for the open world re-id task

For both tested datasets, the performances of Lasso DC and Lasso DCE are similar for closed world evaluations. However, for open world cases, the DIR vs FAR value for Lasso DCE is about 5% higher than that of Lasso D. This is true for XQDA projected features (LOMO and NN) but once again there is the exception of non projected IRfeatures for which the performances of collaborative and collaboration enhanced sparse coding are comparable.

For open world scenarios, on the iLIDS-VID dataset, we obtain comparable results to state-of-the-art method DVR [9], with a DIR at rank 1 for a one percent FAR of around 17%. On the PRID2011 dataset, we outperform DVR [9] by far with a DIR at rank 1 for a one percent FAR of 55.7% compared to 46.8% for DVR.

The reasons why relying on residual errors of probe reconstructions using gallery dictionary or extended gallery dictionary does not have much effect on closed world performances is because CMC only evaluates the relevance of the ranking of gallery identities and adding additional elements to the collaborative gallery dictionary does

not influence the relative participation of the gallery dictionary elements in the probe reconstruction so the ranking of gallery identities remains stable.

In the open world case, DIR vs FAR is an hybrid evaluation which combines ranking and decision aspects. While for closed world re-id, having a small residual error for a wrong match does not matter, in the open world case, it can have an impact on the False Aceptance Rates, and as a consequence it modifies the Detection and Identification Rate which are reported for a few values of False Acceptance Rates. In the open world scenario, some probe person do not appear in the gallery and should therefore not be reconstructed by gallery elements. The additional dictionary's aim is to participate for those probe people so that it reduces the participation of gallery identities and avoids ranking gallery identities that are not similar to the probe person. Thanks to the additional dictionary, there are less false matches accepted for small values of residual errors. The residual errors corresponding to true matches are also bigger but to a lesser extent. This can be observed in Figure 4.6 where True Positive and True Negative rates are plotted when varying the decision threshold between 0 and 1. This explains how better Detection and re-Identification Rates can be obtained for the same False Acceptance Rate, even if the ranking evaluations are similar.

Relevance of the additional dictionary choice: Gallery vs Probe features

We recommended to use elements from the gallery camera as additional elements to avoid favoring features extracted from the probe camera while reconstructing query probe people. In Tables 4.5 and 4.6 are presented the open world DIR vs FAR results for collaborative sparse coding without additional dictionary (Lasso DC), with training probe features as additional dictionary (Lasso DCE probe) and with training gallery features as additional dictionary (Lasso DCE).

FAR(%)	1	10	50	100
IR_n + Lasso DC	0.1	0.8	2.2	3.0
IR_n + Lasso DCE probe	0.4	1.8	5.9	8.7
IR_n + Lasso DCE	0.4	0.9	2.4	3.4
IR_{pn} + Lasso DC	7.3	16.7	36.4	46.0
IR_{pn} + Lasso DCE probe	7.1	16.6	34.8	44.6
IR_{pn} + Lasso DCE	7.3	18.6	37.7	44.9
$LOMO_{pn}$ + Lasso DC	12.9	35.1	58.8	68.5
$LOMO_{pn}$ + Lasso DCE probe	16.9	36.9	61.1	**69.8**
$LOMO_{pn}$ + Lasso DCE	**17.2**	**37.5**	**62.8**	69.0

Table 4.5 – Evaluation on open world re-identification task. DIR at first rank for several FAR values (1%, 10%, 50% and 100%) on iLIDS-VID dataset.
Comparison of three variants of Lasso Direct direction: without additional dictionary, with an additional dictionnary composed of training probe features and with an additional dictionary composed of training gallery features. Best results are in bold red.

FAR(%)	1	10	50	100
IR_n + Lasso DC	8.8	13.7	26.8	35.8
IR_n + Lasso DCE probe	12.7	21.3	34.0	47.2
IR_n + Lasso DCE	8.7	14.5	27.7	34.5
IR_{pn} + Lasso DC	21.0	44.8	66.2	75.2
IR_{pn} + Lasso DCE probe	25.7	47.0	68.5	75.3
IR_{pn} + Lasso DCE	28.0	46.8	69.5	76.3
$LOMO_{pn}$ + Lasso DC	49.8	69.3	88.2	**93.8**
$LOMO_{pn}$ + Lasso DCE probe	54.8	**71.5**	89.3	93.2
$LOMO_{pn}$ + Lasso DCE	**55.7**	71.0	**90.2**	93.2

Table 4.6 – Evaluation on open world re-identification task. DIR at first rank for several FAR values ($1\%, 10\%, 50\%$ and 100%) on PRID2011.
Comparison of three variants of Lasso Direct direction: without additional dictionary, with an additional dictionnary composed of training probe features and with an additional dictionary composed of training gallery features. Best results are in bold red.

For the non projected IR features, using the training features from the probe camera gives better results than using features from the gallery camera as additional dictionary but the performance reached with non projected IR feature is really low. For XQDA projected features (LOMO and IR), forming the additional dictionary with features from the gallery camera gives slightlly better results than with features from the probe camera (+0.3 for iLIDS-VID and +0.9 for PRID2011). Regardless of the camera provenance of the additional dictionary features, collaboration enhanced sparse coding approaches perform better than the simple collaborative sparse coding approach.

4.4.4 Evaluation on the person verification task

In this section we report verification performances, ie. we evaluate the ability to distinguish positive pairs from negative pairs. We do not consider the ranking aspect anymore.

One way to compare re-identification methods for the verification task consists in comparing ROC curves where the x-axis corresponds to False Positive rate and the y-axis to True Positive rate. ROC curves are computed for a wide range of decision threshold values. In Figure 4.3 are presented the ROC curves for XQDA, Lasso DNC, Lasso DC and Lasso DCE obtained on iLIDS-VID and PRID2011 datasets. The ROC curves of the two collaborative sparse representation approaches (with and without collaboration enhancement) are very similar and they are clearly above the ROC curves of the XQDA and Lasso DNC approaches which are also similar.

A drawback of the ROC evaluation is that the decision threshold does not appear on the ROC curve. However, for real applications a decision rule is applied in order to tell apart positive pairs from negative pairs and it is important to know which threshold should be used, or at least in what range the relevant threshold is and what kind of results could be obtained for some given thresholds values. Except

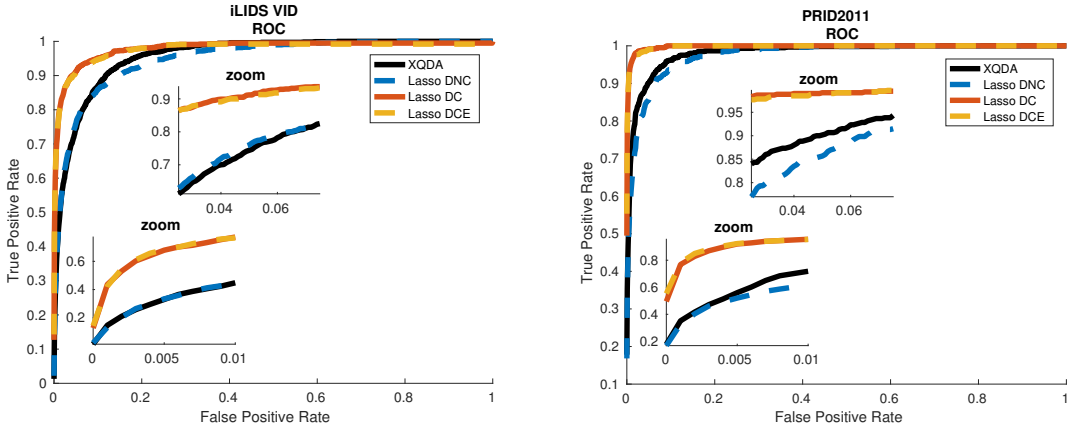


Figure 4.3 – ROC curve for XQDA, Lasso DNC, Lasso DC and Lasso DCE on iLIDS-VID (left) and PRID2011 (right) datasets. There is a zoom corresponding to the smallest values of false positive rates, and another zoom for slightly bigger false positive rates.
XQDA and the non collaborative sparse coding approach (Lasso DNC in blue) have similar ROC values. The two collaborative sparse coding approaches, Lasso DC and Lasso DCE also have similar ROC values. The collaborative sparse coding approaches perform much better than XQDA and the non collaborative approach on both datasets.

for the LADF approach [22] which proposes a local threshold as decision rule, the verification question has rarely been raised in the person re-identification literature, even in papers which cast this problem as a binary classification task.

Let us visualize in Figure 4.4, the distribution of dissimilarity scores of positive and negative pairs for XQDA, Lasso DNC, Lasso DC and Lasso DCE. The dissimilarity scores correspond to the distances in the case of XQDA, to the reconstruction errors in the case of Lasso DNC and to the residual errors in the case of Lasso DC and Lasso DCE. At first sight, we would group the distributions into the two same groups as those we found with the ROC results. Indeed, on one hand, the shape of the distributions of positive and negative pairs scores are similar for XQDA and Lasso DNC. Negative and positive pairs distributions overlap in a large interval compared to the range over which positive pairs scores spread and negative pairs scores spread over a larger interval than positive pairs scores. On the other hand, the distributions of positive and negative pairs scores are very similar for the two collaborative sparse coding approaches Lasso DC and Lasso DCE. Negative and positive pairs distributions overlap in a small interval compared to the range over which positive pairs scores spread and it is positive pairs scores that spread over a larger interval than negative pairs scores. Nonetheless, while XQDA and Lasso DNC lead to similar ROC curves with distributions of positive and negative score of similar shapes, we must point out that the interval over which the XQDA distances spread differs quite a lot between the two datasets (between 0 and 120 for iLIDS-VID and between 0 and 300 for PRID2011) while the reconstruction errors obtained with the Lasso DNC approach always stays between 0 and 1. Therefore, for the sparse coding approaches, collaborative or not, the range over which the dissimilarity scores spread is known and stays the same for every dataset. For XQDA however, since

the dissimilarity score spread over different intervals for every dataset, a validation step is necessary for each dataset to determine the range in which a good decision threshold could be found for a given application.

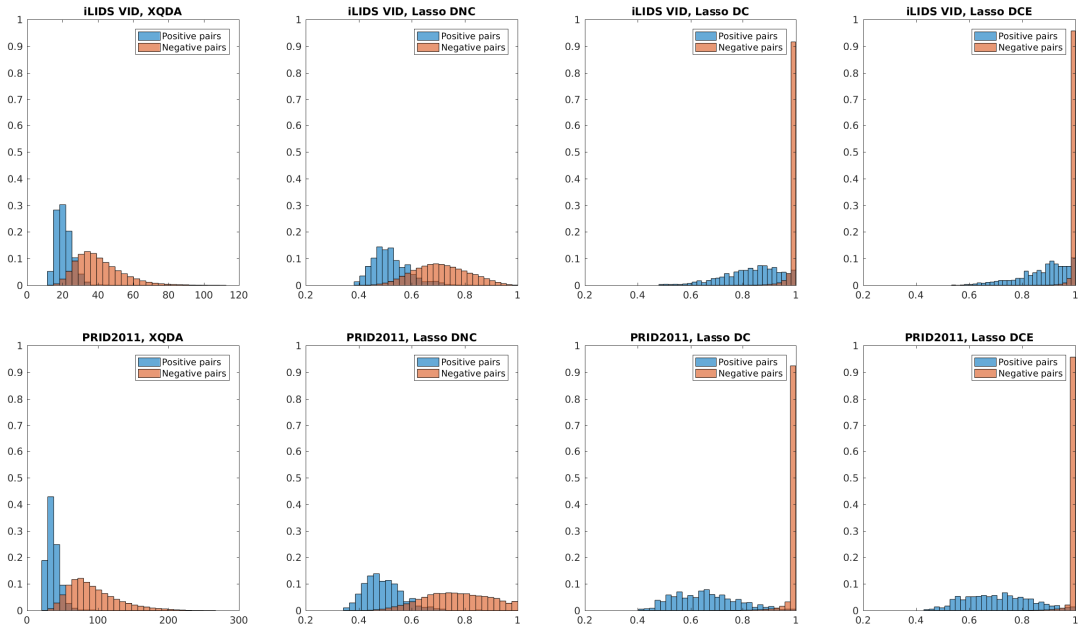


Figure 4.4 – Distribution of positive and negative pairs distances for XQDA and distribution of residual errors for Lasso DNC, Lasso DC and Lasso DCE on iLIDS-VID dataset (top) and PRID2011 dataset (bottom).
We can group these distributions in two groups. The first group contains the distributions obtained with the XQDA approach and the non collaborative Lasso DNC approach, where the scores (distance or reconstruction error) of negative pairs spread over a larger interval than the scores of positive pairs. The second group contains the distributions obtained with the collaborative sparse coding approaches Lasso DC and Lasso DCE where the scores of negative pairs span on a very small interval while the score of positive pairs spread over a much larger interval. We can notice that while the distances for XQDA spread over different intervals for iLIDS-VID and PRID2011, for all three sparse coding approaches, the reconstruction errors and residual errors stay between 0 and 1 for both datasets.

Instead of representing the distributions of positive and negative pairs scores, the same kind of information can be found by visualizing the TP rates and TN rates for varying thresholds. We plot for both iLIDS-VID and PRID2011, the TP rates and TN rates for varying threshold values for the XQDA method in Figure 4.5 and for the sparse coding methods (Lasso DNC, Lasso DC and Lasso DCE) in Figure 4.6. We can once again observe in Figure 4.5 that for XQDA, the distances that positive and negative pairs take are quite different on iLIDS-VID dataset and PRID2011 dataset and the decision threshold value which leads to equal TP rate and TN rate also differs a lot between datasets with a value smaller than 50 for iLIDS-VID and bigger than 50 for PRID2011. In Figure 4.6, we can also observe once more that the dissimilarity scores of sparse coding approaches, stay between 0 and 1. Although not in the same range, the shapes of TP rates and TN rates for varying thresholds of

XQDA and Lasso DNC are similar. Almost as soon as the TP rate starts increasing, the TN rate starts decreasing. For the collaborative sparse coding methods, there is a large interval in which the TN rate stays high and the TP rate increases. It is only near 1 that the TN rate drops.

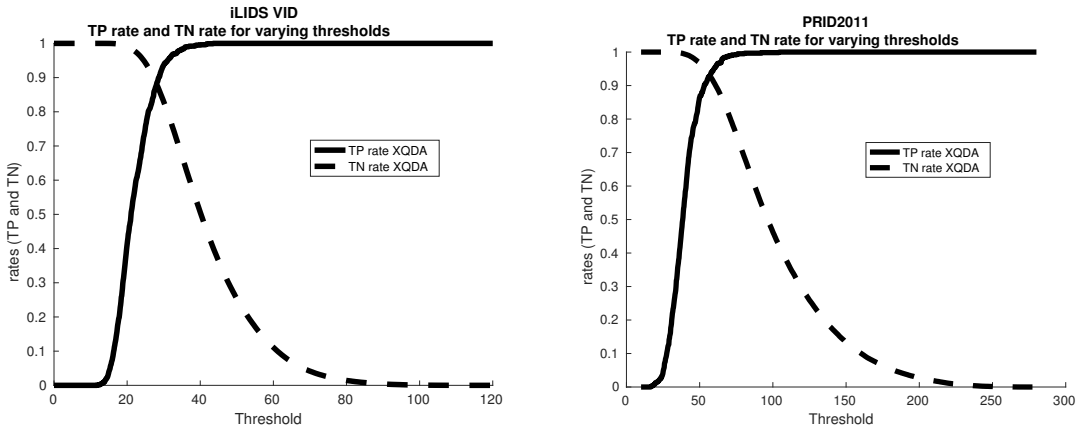


Figure 4.5 – TP rate and TN rate for varying thresholds for XQDA on iLIDS-VID and PRID2011 datasets.
The TP rate and TN rate have the same shape on both datasets. However, the distribution of positive and negative pairs do not span on the same intervals. The threshold value which lead to equal TP rate and TN rate is smaller than 50 for iLIDS-VID and bigger than 50 for PRID2011.

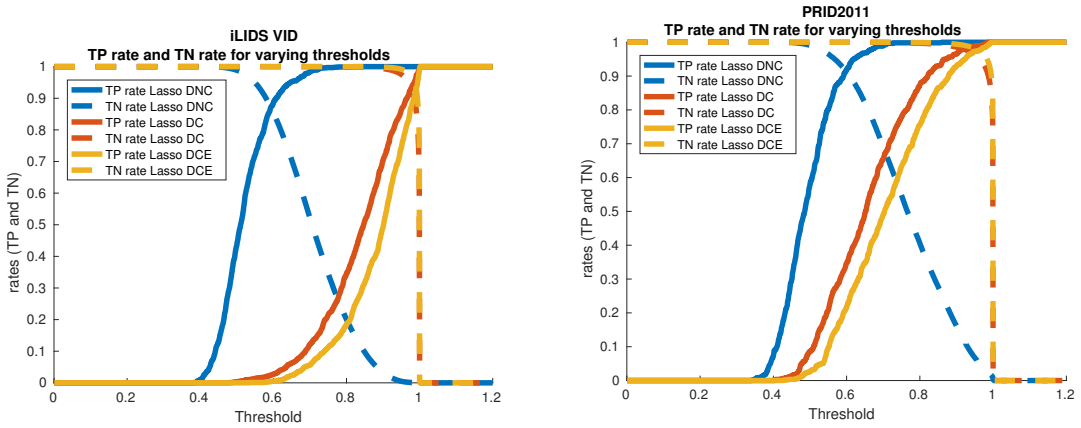


Figure 4.6 – TP rate and TN rate for varying thresholds sparse coding approaches (Lasso DNC, Lasso DC and Lasso DCE) on iLIDS and PRID datasets
The shape of TP rate and TN rate for the non collaborative approach Lasso DNC is clearly different from the shape of TP rate and TN rate of collaborative approaches Lasso DC and Lasso DCE. For the collaborative spasre coding approaches, the TN rate stays high in almost the whole interval and only drops quickly when the threshold gets close to 1.

Besides those observations, plotting the TP rate and the TN rate, allows for a more precise comparison of the distributions of positive and negative pairs scores and a better understanding of the impact of the additional dictionary. In Lasso DCE, since there are additional elements which can participate in the reconstruc-

tions of probe elements compared to Lasso DC, the elements that are in the initial collaborative gallery dictionary can only participate as much as it did without the additional dictionary, or less. Therefore the mean residual errors obtained with the additional dictionary are bigger, for both positive and negative pairs (the yellow curves are on the right of the red curves). However, what makes Collaboration Enhanced Lasso DCE perform better than Lasso DC in terms of DIR vs FAR for small values of FAR, is not the general increase of the residual errors values, it is the fact that for small decision threshold values, the false positive dissimilarity scores as pushed further away (to the right) than the true positives dissimilarity scores.

In the theoretical part, we argued that 1 was a natural decision threshold. However, some applications might prefer to retrieve more true positives at the cost of a smaller precision while other applications might prefer to retrieve less positive pairs but with a good precision. In Tables 4.8 and 4.7, we report the recall (true positive rate), specificity (true negative rate), classification rate ($\frac{TP+TN}{TP+TN+FP+FN}$) and precision ($\frac{TP}{TP+FP}$) for Lasso DCE for different values of decision threshold ($0.999, 0.950$ and 0.900) that are close to 1. For all three chosen decision thresholds, the TN rate is high and since there are much more negative pairs than positive pairs, the value of the classification rate is almost the same as the TN rate. For the same three decision thresholds, the TP rate varies more than the TN rate, especially on the iLIDS-VID dataset. Even if some of the precision values are low, knowing that there are much more negative pairs of identities than positive pairs of identities, the precision values are actually quite acceptable. In practice, for a chosen application, it shouldn't be difficult to find an appropriate decision threshold because the TN rate stays high in a large interval while the TP rate increases in a linear way in the interval of interest so one could slowly increase the decision threshold until the wanted precision and recall specifications are reached.

Threshold	TP rate	TN rate	classification rate	precision
0.999	97.8	81.6	81.7	3.4
0.950	73.8	99.2	99.0	38.3
0.900	48.8	99.9	99.5	70.3

Table 4.7 – Recall, Specificity, Classification rate and Precision values for 3 decision threshold values on the iLIDS-VID dataset.

Threshold	TP rate	TN rate	classification rate	precision
0.999	100.0	85.8	85.9	7.4
0.950	97.3	98.2	98.2	38.7
0.900	92.8	99.4	99.4	65.3

Table 4.8 – Recall, Specificity, Classification rate and Precision values for 3 decision threshold values on the PRID2011 dataset.

4.5 Conclusion

In this chapter we have compared sparse coding without collaboration and with collaboration for the person re-identification task. The competition induced between identities by collaborative sparse coding is what makes it relevant for ranking tasks or more precisely for relative ranking tasks. It enables to boost the ranking performances which lead to a significant gap in performances for closed and open world scenarios between the metric learning method XQDA and our collaborative sparse coding approach applied to $\mathcal{L}2$ norm normalized XQDA projected LOMO and IR features. Indeed, it lead for example for LOMO features to an increase of +10% on closed world iLIDS VID first rank recognition rate and to an increase of +28% on open world PRID2011 detection and re-identification rate at first rank when FAR equals 1%.

For open world re-identification, besides ranking abilities, a good method needs to be able to make a decision. We proposed the Lasso DCE method which stands for Direct Collaboration Enhanced Lasso approach. In addition to collaborative sparse coding adequacy for ranking tasks, the proposed collaboration enhancement makes collaborative sparse coding also fit for the detection task by better rejecting false matches. This lead to a further improvement of around 5% for DIR values at rank 1 when FAR equals 1% compared to collaborative sparse coding without additional dictionary (Lasso DC).

For the person verification task, depending on the application, ie. depending on whether it is more important to have a high precision, or to retrieve most of the true matches, the ideal decision threshold is to be determined by the user but we know the decision threshold should be close to 1. The shape of the positive and negative pairs distributions is definitely an advantage for easily finding that ideal threshold because the proportion of positive pairs increases almost linearly with the threshold while the proportion of negative pairs stays high in a large interval and only drops quickly near 1.

Chapter 5

Bidirectional Sparse Representations

In the previous chapter, we showed how for a given probe person, collaborative sparse coding could be used to reject wrong gallery matches and rank similar looking gallery people. In this chapter, we emphasize on the importance of a reciprocal relation. Instead of only focusing on ranking and rejecting gallery identities for a given probe person, we also take the gallery person point of view. Using once again sparse coding with an enhanced collaboration, for a given gallery person, we can reject wrong probe matches and also rank similar looking probe people if several of them are presented. By combining the results from both collaboration enhanced sparse coding approaches, we obtain more robust detection and ranking results.

In this chapter, we call direct direction the sparse coding of probe elements with gallery elements, and reverse direction the sparse coding of gallery elements with probe elements. Even if at first glance the reverse direction sparse coding approach seems to be quite symmetrical to the one presented in the previous chapter, several important aspects differ.

Therefore we first present the main differences between probe and gallery data which make the reverse direction sparse coding approach not so symmetrical to the direct direction sparse coding approach. Then we expose our reverse direction sparse coding approach. The third section focuses on the meaning of residual errors for direct and reverse directions. The fourth section explains how we combine direct and reverse direction sparse representations results. The fifth section deals with complexity issues. The last section presents extensive experiments on closed and open world re-identification tasks and on the verification task as well.

5.1 Difference between sparse coding of probe and gallery elements

5.1.1 Known and undetermined identities

As the reader should know by now, gallery elements are from known identities while the identity of probe elements is to be determined. Though obvious, this simple observation makes a crucial difference in the way the additional dictionaries are formed.

In the direct direction, the identity of the probe element is undetermined, he could be someone present in the gallery or not, so when reconstructing it with the gallery dictionary, it is important that the overall collaborative dictionary contains varied features that can reconstruct well anyone so that gallery identities do not participate to a non similar probe element's reconstruction only to obtain a balanced overall reconstruction error and sparsity term.

In the reverse direction, elements that are reconstructed are known gallery elements. A probe person would be considered similar to the gallery person if he participates to the reconstruction even if put in competition with elements that are similar to the gallery person. The participation of the probe person to the reconstruction of a gallery element is not meaningful if he is only put in competition with elements that are dissimilar to the gallery person.

5.1.2 Availability of gallery and probe data

In the direct direction, each probe element is approximated by a linear combination of gallery elements, where any gallery element is allowed to participate in the reconstruction. For a symmetric approach, in the reverse direction, we should reconstruct each gallery element with a linear combination of probe elements, where any probe element is also allowed to participate.

However, depending on the application, all probe elements might not all be available simultaneously at test time. Some applications might provide one test person's images at a time or a few identities images while other applications might provide simultaneously many people's images.

In the direct direction, the aim of the additional dictionary is to improve false match rejection in a sparse coding framework which is already collaborative. In the reverse direction, the additional dictionary is necessary when only one test person is presented at a time. Otherwise, with only one probe person's features, there is no collaboration, we fall into a non collaborative sparse coding method which is not so useful for our task.

5.1.3 Final goal

We must keep in mind that what we are looking for is the presence or the absence of a probe person in the gallery set, and his identity if relevant. Therefore, it is not important to be able to give a ranking of probe identities for a given gallery person, but it is important to be able to return a ranked list of gallery identities for a given probe person.

A ranked list of probe identities for given gallery identities can be very useful in a closed world setting to reinforce the robustness of the ranking. However in the open world case, even if identities in the gallery and the probe sets overlap, some gallery identities might not appear in the probe set (they are called distractors) and some probe identities might not be in the gallery set (they are the imposters who must be rejected), so such a ranked list looses a little bit of its interest even if it does not become useless.

5.2 Reverse direction: sparse coding of gallery elements

5.2.1 Sparse representation of gallery elements

Depending on the application, we might have more or less probe identities to re-identify at a given test time. In this chapter, $\mathcal{L}$ refers to the set of probe identities that we have to re-identify at a given test time. If we only have one probe person's images at a time, then $L = card(\mathcal{L}) = 1$. If we have several probe person's images at a time, then $L > 1$. Let $P = [P_{l_1}, \ldots, P_{l_L}]$ denote the concatenation of the features from all the available probe images.

For every gallery identity $k \in [1, K]$, we compute the sparse representation $A_{G_k,[P,D_k]}$ of their features G_k by optimizing the following Lasso problem:

$$A_{G_k,[P,D_k]} = \arg\min_A ||G_k - [P, D_k]A||_F^2 + \lambda||A||_1 \tag{5.1}$$

where the additional dictionary D_k is different for each gallery identity, and depends on each gallery identity's specific dictionary G_k. The way the dictionary D_k is computed is explained in the next subsection.

The dissimilarity score $s(l, k)$ between probe person l and gallery person k is defined by the mean residual reconstruction error of the reconstruction of gallery features G_k using only the elements from probe person l's dictionary P_l, when the sparse representation is the one computed with dictionary $[P, D_k]$:

$$s(l, k) = R_{G_k,[P,D_k],P_l} = \frac{||G_k - P_l A_{G_k,[P,D_k],P_l}||_F^2}{n_k} \tag{5.2}$$

The gallery identities are once again ranked by increasing dissimilarity score, and for a probe person l the best match is given by:

$$k^* = \arg\min_k R_{G_k,[P,D_k],P_l} \tag{5.3}$$

We call Lasso RCE this sparse coding approach where the sparse representations of a gallery person's features are computed using test probe features and gallery additional dictionary. R refers to Reverse direction, because instead of Figure 5.1 illustrates Figure 5.2 illustrates

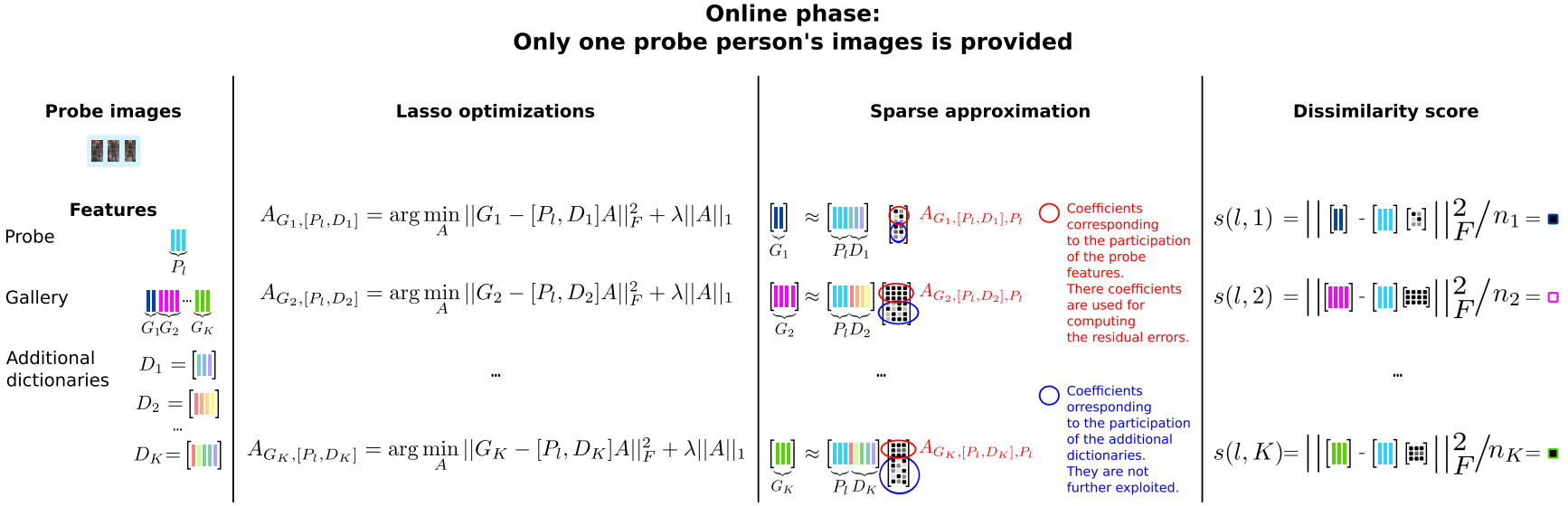


Figure 5.1 – Overview of the online part of the Lasso RCE (Reverse Collaboration Enhanced) approach in the case only one probe person is presented at a time.
In multi-shot datasets, probe and gallery identities are represented by several images and the number of images can vary with the person. n_k is the number of images of gallery person k. In the shown example, the gallery person 1 is represented by 1 image, the gallery person 2 by 4 images and the gallery person K by 3 images. A column feature is computed for each image. We have at disposal, the features of the probe person l whose identity we are looking for, the features of gallery people and the additional dictionary associated to each test gallery. For each gallery person, a Lasso problem is solved in order to find the sparse representation of the gallery features using a dictionary composed of the probe person's features and the additional dictionary corresponding to the considered gallery person. The sparse representations are presented by gray scale matrices where the darker the square, the smaller the value. The coefficients corresponding to the participation of the additional dictionaries (coefficients circled in blue) are discarded. The coefficients corresponding to the participation of the probe person features (coefficients circled in red) are used to compute the dissimilarity score (mean residual error) between the gallery person and the probe person. Once again, the dissimilarity score is displayed in gray scale, where smaller values are darker.

5.2.2 Choice of the additional dictionaries

In this reverse direction collaborative sparse coding approach, it is in order to find out whether the probe elements are similar to the gallery elements or not that we reconstruct gallery elements using probe instances. It is only if the probe person participates in the reconstruction of the gallery elements when collaborating with instances that are already considered similar to that gallery person that he will be considered a possible match. If the probe person participates more than non similar elements, it is not really informative.

Since the gallery elements are known, it is possible to find among the training samples, instances that are similar to the gallery person, or at least to select the most similar ones and use them as additional collaborative and competitive elements for the reconstruction.

Using the same argument as for direct direction collaboration enhanced sparse coding, we reckon that since the additional dictionary will come in complement to probe dictionaries, their elements should also be features corresponding to images captured by the probe camera. In order to select the training samples that will form the additional dictionaries, we compute the sparse representation of every gallery instance using the probe training dictionary T_p.

$$A^* = \arg\min_A ||G - T_p A||_F^2 + \lambda ||A||_1 \tag{5.4}$$

Each column of T_p corresponds to a training image feature. The identity associated to each column of T_p is not important. Therefore we do not decompose the matrix A^* according to the identity it is associated to in T_p (ie. by selecting blocs rows), but instead we decompose the matrix A^* by selecting blocs of columns, according to the decomposition of $G = [G_1, \ldots, G_K]$.

$$A^* = [A_1^*, \ldots, A_K^*] \tag{5.5}$$

For each gallery person k, the sparse representation of their features G_k is given by the sparse submatrix A_k^* of A^*. We form the additional dictionary D_k used in addition to probe elements for reconstructing the gallery person k by selecting the columns of T_p which corresponds to the rows of A_k^* which contain non zero values. Since A^* is sparse, A_k^* is also sparse, and the dictionary D_k is small. Put in words, D_k contains all the elements from the probe camera training set which participate in the reconstruction of at least one of the gallery person k's feature when put in competition with the other training elements.

The reason why we form an additional dictionary for each gallery person by selecting a few training samples rather than using the same global additional dictionary composed of all training samples is to reduce the computations while maintaining similar performances. Indeed, contrary to the direct direction where only one sparse representation was needed to rank gallery identities for a given probe person, in the reverse direction, we need to compute one sparse representation for each gallery element. Using a smaller dictionary reduces the computation and memory load.

It is particularly relevant to form the additional dictionaries by select training elements using collaboraive sparse coding because it is the way they will be used at test time, and training instances that have not been selected when reconstructing a given gallery element, will not be selected either if put in competition with even more elements at test time for reconstructing the exact same gallery instance.

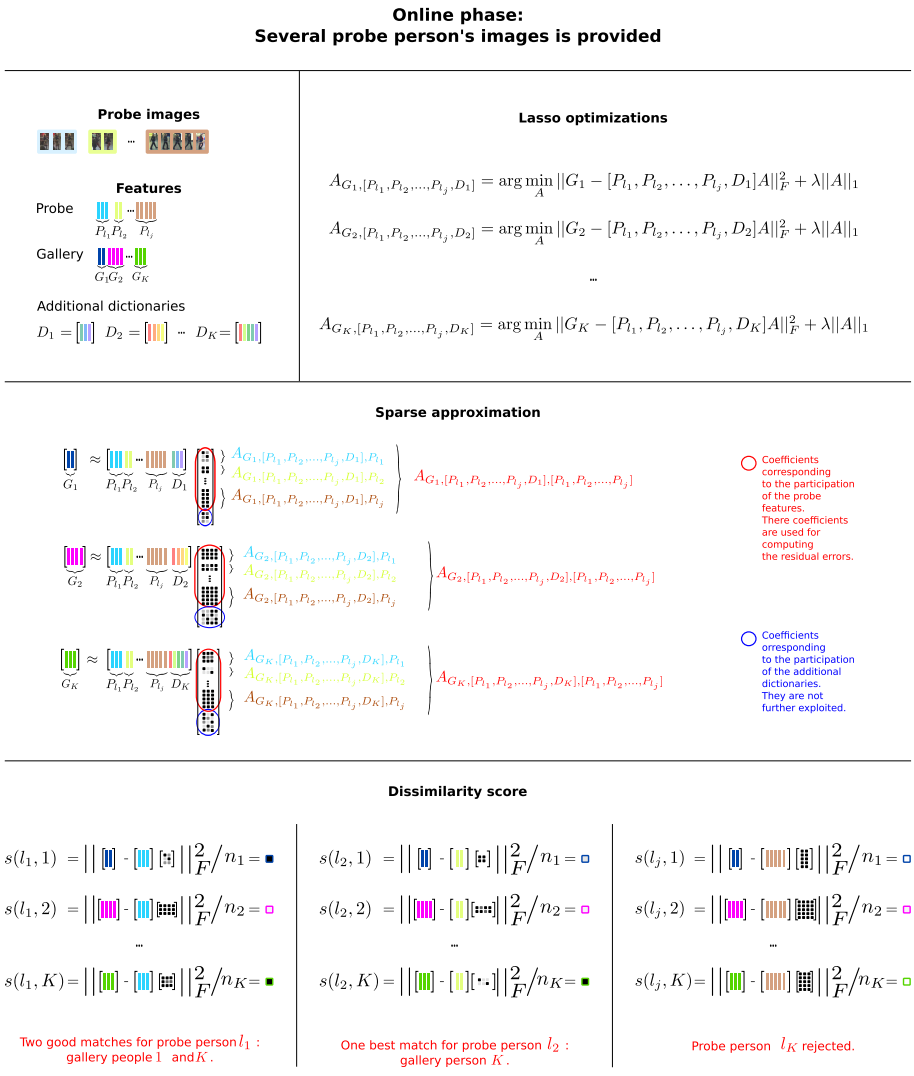


Figure 5.2 – Overview of the online part of the Lasso RCE (Reverse Collaboration Enhanced) approach when several probe person's images are simultaneously provided.
In multi-shot datasets, probe and gallery identities are represented by several images and the number of images can vary with the person. n_k is the number of images of gallery person k. In the shown example, the gallery person 1 is represented by 1 image, the gallery person 2 by 4 images and the gallery person K by 3 images. A column feature is computed for each image. We have at disposal, the features of several probe people $l_1, l_2, \ldots, l_j$ whose identity we are looking for, the features of gallery people and the additional dictionary associated to each test gallery. For each gallery person, a Lasso problem is solved in order to find the sparse representation of the gallery features using a dictionary composed of all the provided probe people's features and the additional dictionary corresponding to the considered gallery person. The sparse representations are presented by gray scale matrices where the darker the square, the smaller the value. The coefficients corresponding to the participation of the additional dictionaries (coefficients circled in blue) are discarded. The coefficients corresponding to the participation of each probe person features (coefficients circled in red) are used to compute the dissimilarity score (mean residual error) between the gallery person and the probe person. Once again, the dissimilarity score is displayed in gray scale, where smaller values are darker. If all dissimilarity score for a probe person are big (white in the figure), the probe person is considered as an imposter. Otherwise, gallery identities are ranked by increasing dissimilarity scores.

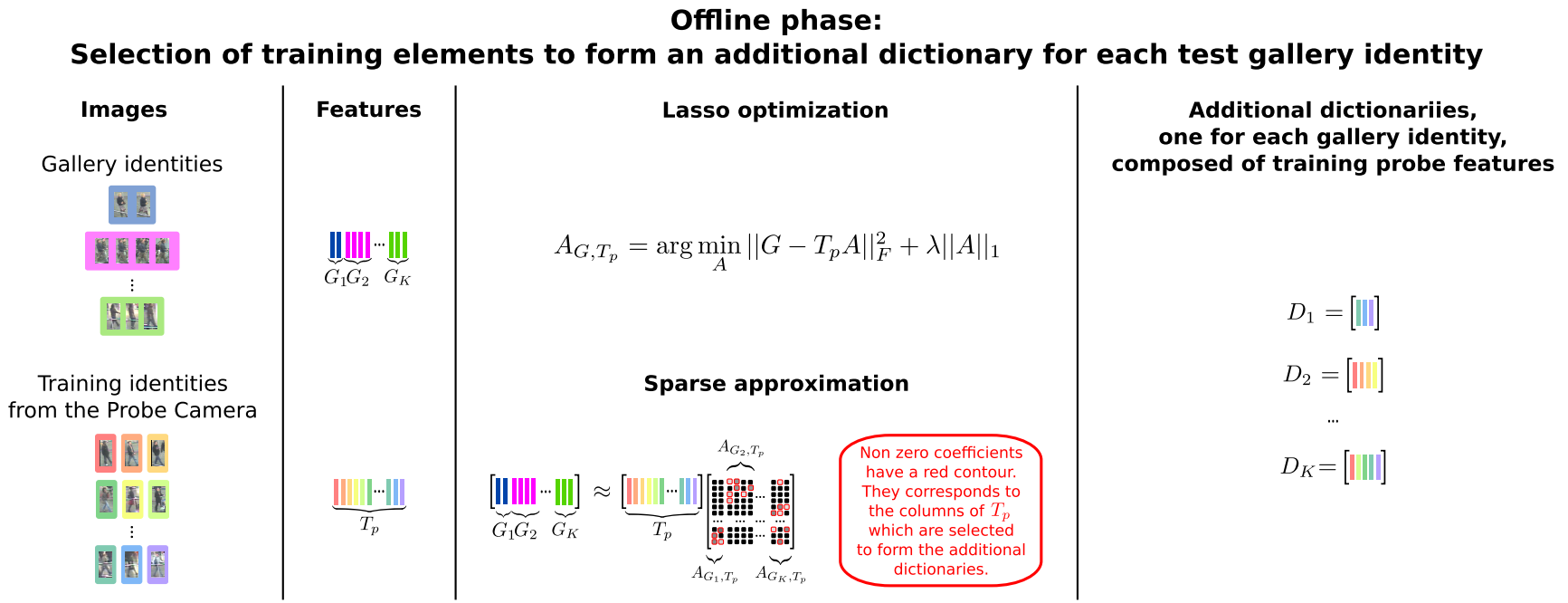


Figure 5.3 – Overview of the offline part of the Lasso RCE approach which aims at learning each gallery person's specific additional dictionary. The offline part of the Lasso RCE approach selects training elements to form an additional dictionary for each of the test gallery person. We have at disposal, the features corresponding to images from the training set, and the features corresponding to the known gallery people of the testing set. The number of images available for each gallery person vary. The size of the additional dictionary associated to a gallery person we aim at constructing is not fixed in advanced either and can also vary.
For each gallery person, a Lasso problem is solved to approximate the gallery person's features using elements from the training set (from the probe camera). The obtained sparse matrix is represented in gray scale where black dots correspond to zero values and gray dots (also circled in red) are non zero values. We form the additional dictionary for a given gallery person by selecting all the training instances which have participated in the reconstruction of that gallery person features (ie. which are associated to a non zero coefficients in the sparse matrix). Those elements should be those who are the most similar to the gallery features.

5.3 Ranking of gallery identities, meaning of the residual errors

In both the direct direction collaborative sparse coding approach, and the reverse direction we rely on the residual errors for ranking gallery identities for each probe person. However, the meaning behind the ranking based on the residual errors is different. Indeed, there is an asymmetry in our problem. In both directions, given a probe identity, the residual errors are used to rank gallery identities from the most similar ones to the most dissimilar ones, we never rank probe identities for a given gallery person.

Let's take a closer look at the residual errors involved in the ranking process. In the direct direction, gallery identities are ranked by increasing residual errors where the so called residual errors $\{R_{P_l,[G,D],G_k}\}_k$ are the residual errors of the reconstruction of a probe element using only the specific dictionary of each gallery person k. We therefore rank gallery identities based on how much each gallery person participated in the reconstruction of the probe elements, knowing that gallery identities were collaborating and competing against each other for the overall reconstruction. The best match is the gallery who participated the most.

In the reverse direction, gallery identities are still ranked by increasing residual errors but this time, the residual errors in question $\{R_{G_k,[P,D_k],P_l}\}_k$ are the residual errors of the reconstruction of each gallery features using the probe person's elements and an additional dictionary which is different for each gallery person. While in the direct direction the residual errors were computed from the same sparse representation, in the reverse direction, the residual errors we compare to rank gallery identities k seem to be unrelated to each other and it is legitimate to wonder whether these residual errors are comparable. Indeed, there is one additional dictionary per gallery person, so a given gallery person has an additional dictionary that is different from another gallery person and we do not control how well each gallery person's element's reconstructions could be using their associated additional dictionary. Depending on the content of the training set and that of the test gallery set, the additional dictionaries of some gallery people could be very similar to their associated gallery person, but could also be quite dissimilar for other cases. Is it thereby relevant to rank gallery identities based on these residual errors?

Let's take a step back and assume that instead of a different additional dictionaries per gallery identity, we use the whole probe camera training set as a common additional dictionary. In that case, ranking gallery identities based on the residual errors of their reconstruction using the probe person's elements and the common additional dictionary would mean looking for the gallery identity the probe person is the closest to, because it would be the gallery person for whom the probe person contributes the most when put in competition with the same set of elements. In this situation, it is understandable to compare the residual errors for ranking gallery identities.

Now, if we return to our previous case where the additional dictionary of each gallery person is selected by our proposed collaborative sparse coding approach 5.4, we should obtain similar results to the case when the same common additional dictionary is used because the elements we discarded to form the additional dictionary of a gallery person were already not participating in the reconstructions.

To sum up, in the direct direction, we rank the gallery identities according to which ones participated the most to the reconstruction of the probe person when they are all put together to collaborate. In the reverse direction, we look for the gallery person for whom the probe person participates the most when put into competition with the same elements.

In the reverse direction, in the case all the probe identities corresponding to the gallery identities are available at test time, there is a double comparison. Indeed, when computing the sparse representations gallery elements G_k, the features P_{l_i} of a given probe person l_i are in competition with other test probe identities features $\{P_{l'_i}\}$. However, when ranking gallery identities, we are not interested anymore in whether the probe person l_i participated more than probe person l'_i, but what we are interested in is to which gallery person's features reconstruction the probe person l_i participated the most. Therefore l_i might be the person who participated the most in the reconstruction of gallery person k's features, and not be the one who participated the most for gallery person k' but if the residual error for person k' is smaller than that of person k, then k will be ranked before k'. Exploiting the reconstruction errors from the reverse direction sparse reconstructions can be more powerful than exploiting the errors obtained with the direct direction because the reverse direction integrates different comparisons, but this is the case only if the reverse direction sparse representation involved the simultaneous participation of several probe identities that are also present in the gallery set.

5.4 Combination of both representations

We have presented two different collaborative sparse methods for the person re-identification which might have slightly different results because one is based on the probe person's point of view and the other one on the gallery people's point of view. A good match is a match where both parties agree, so combining the results from the two methods should lead to more a robust system.

We define the dissimilarity between the probe person l and the gallery person k as the sum of the dissimilarity score obtained with the direct direction method and the dissimilarity score obtained with the reverse direction method. Thus, the dissimilarity score is given by:

$$s(l,k) = R_{P_l,[G,D],G_k} + R_{G_k,[P,D_k],P_l} \tag{5.6}$$

Combining the dissimilarity scores from the two methods by a simple sum is relevant because the dissimilarity scores are in the same range (between 0 and 1).

Gallery identities are ranked by increasing dissimilarity value, and the best gallery match for probe id l is given by:

$$k^* = \arg\min_k s(l, k) \tag{5.7}$$

We have seen that for the direct direction sparse coding approach Lasso DCE, the multi-shot aspect of probe data is dealt with in the usual way by a simple average aggregation function and the multi-shot aspect of gallery data is well exploited thanks to the computation of residual errors by gallery identity rather than by gallery image.

In the reverse direction sparse coding approach Lasso RCE, it is the reverse. The multi-shot aspect of gallery data is dealt with in the usual way by a simple average aggregation function and the multi-shot aspect of probe data is fully exploited thanks to the computation of residual errors by probe identity.

Combining the direct and the reverse direction sparse coding approaches is therefore also a way to better exploit the availability of multiple images in both the probe set and the gallery set.

5.5 Complexity

This section studies the complexity of our method (direct and reverse direction) at test time for the task of ranking the K gallery identities when given $L \in \mathbb{N}^*$ probe identities to be re-identified. We address the differences between the cases when there is only one probe person's images available at a time, or when probe images are given by batches of u identities with $v = \frac{L}{u} \in \mathbb{N}$, or when all L probe identities images are given simultaneously. For simplicity, we consider that the computation of a sparse representation in the direct direction is given by the constant C_d and the computation of a sparse representation in the reverse direction is given by the constant C_r. We also count as one the computation of the sparse representations of all features of one given person, instead of counting it as the number of images available for that person.

Let's begin with re-identifying one probe person l. In the direct direction, we have a complexity of $C = C_d$ because probe person l's elements are approximated just once by a linear combination of all gallery elements. In the reverse direction, the complexity is of $C = KC_r$ because each gallery identity's elements must be reconstructed once using elements from the probe person l whom we wish to re-identify and elements from other identities. Therefore the reverse direction sparse representation approach is much more computationally expensive than using only the direct direction approach as it grows linearly with the number of gallery identities.

In the case we want to re-identify L probe people, why do we not just multiply the previous complexities by the number of probe people to be re-identified? It is because for the reverse direction some reconstructions can be re-used and only some

more residual errors need to be computed. Unfortunately this is not the case in the direct direction, so in the case we need to re-identify L probe person the complexity is $C = LC_d$. If we only have one probe person's image at a time, the complexity is also multiplied by L if we have to identify L people, so in that case the complexity of the reverse sparse coding method becomes $C = LKC_r$. In the case all probe images are available simultaneously, the complexity remains unchanged compared to the case we only wanted to re-identify one of the probe person and still equals $C = KC_r$. This is because the dictionary used for computing the sparse representation of gallery elements is $P = [P_{l_1}, \ldots, P_{l_L}]$ which already contains elements from every probe person. For re-identifying L probe people instead of just one, we simply need to compute the residual errors that were not needed previously. If probe identities are presented by batches of u identities, the concept is similar. For each batch, the complexity equals KC_r. Since there are $\frac{L}{u}$ batches, the overall complexity is $C = \frac{L}{u}KC_r$.

The table 5.1 summarizes the complexity of the different variants in the case we have one probe person to re-identify and in the case we have L probe person to re-identify. K is the number of gallery identities.

	LassoD	LassoR (all ids)	LassoR (u ids)	LassoR (1 id)
1 person to be re-id	C_d	KC_r	KC_r	KC_r
u people to be re-id	uC_d	KC_r	KC_r	uKC_r
L people to be re-id	LC_d	KC_r	$\frac{L}{u}KC_r$	LKC_r

Table 5.1 – Complexity in terms of number of sparse representations computations needed for the different proposed variants for the case there is one of K probe people to re-identify. LassoD refers to the Direct direction and LassoR to the Reverse direction. A distinction is made between the case when there is one or several probe identities's images simultaneously.

In the case only one probe person's images are presented at a time, the reverse direction sparse coding method is indeed much more computationally expensive than the direct direction approach, but if several identities's images are available simultaneously, it is not so much more expensive since the constant C_r is smaller than C_d because the additional dictionary is small.

5.6 Experimental results

The datasets, training and testing protocols are the same as those presented in section 3.3.3. The features used are $\mathcal{L}2$ norm normalized XQDA projected LOMO features [23].

5.6.1 Evaluation on closed and open world re-identification tasks

Depending on the applications, we might have at a time only one single probe person's images available, or several ones, or all of them. Not only it changes the complexity of the algorithm, but it also has an impact on its performance. Assuming that at the end, we have to re-identify L people, we tested our Lasso R approach

(R for reverse direction) for the following number or proportions of probe identities simultaneously available at test time:

- **Lasso RCEs**. There is only one identity provided at a time, **s** stands for single.
- **Lasso RCE$\frac{1}{4}$**. The images of a quarter of the total number L of probe people to be re-identified are simultaneously provided.
- **Lasso RCE$\frac{1}{4}$**. The images of half of the total number L of probe people are simultaneously provided.
- **Lasso RCEa**. All probe people's images are simultaneously available, **a** stands for all.

The closed world CMC results are reported in Tables 5.2 and 5.5 and the open world DIR vs FAR results are reported in Tables 5.2 and 5.5 for iLIDS-VID and PRID2011 datasets.

Rank	1	5	10	20
MDTS-DTW [47]	49.5	75.7	84.5	91.9
DVR [9]	51.1	75.7	83.9	90.5
XQDA[23]	55.3	83.1	90.3	96.3
Lasso DC	64.9	87.1	92.5	96.1
Lasso DCE	65.1	86.6	92.4	96.1
Lasso RCEs	65.4	88.3	93.9	96.8
Lasso RCE$\frac{1}{4}$	67.7	88.9	93.9	96.5
Lasso RCE$\frac{1}{2}$	69.1	89.3	93.7	**96.9**
Lasso RCEa	**69.9**	**89.8**	**94.2**	96.9
Lasso DCE+ RCEs	**68.1**	**88.9**	**93.7**	**96.7**
Lasso DCE+ RCE$\frac{1}{4}$	68.9	89.1	94.1	**96.9**
Lasso DCE+ RCE$\frac{1}{2}$	69.5	89.5	93.9	**96.9**
Lasso DCE+ RCEa	69.8	89.6	93.5	96.8

Table 5.2 – Closed world results on iLIDS-VID dataset.
CMC values at ranks $1, 5, 10, 20$ are reported. Best results are in bold red. Best results where only one probe person is provided at a time are in bold blue.

Relevance of reverse direction sparse coding

Even in the case we are given only one probe person's images at a time, the Lasso RCE approaches perform much better than XQDA alone for both closed and open world scenarios.

Moreover, even though the sparse coding phase of the reverse direction does not involve a direct competition neither between probe identities nor between gallery identities, in the closed world case, the results of Lasso DCE and Lasso RCEs are similar. In the open world case, Lasso RCEs performs better than Lasso DCE on PRID2011 dataset (+2.3%), but worse than Lasso DCE on iLIDS dataset (−3.3%).

Rank	1	5	10	20
MDTS-DTW [47]	69.6	89.4	94.3	97.9
DVR [9]	77.4	93.9	97.0	99.4
XQDA[23]	86.3	98.3	99.6	**100.0**
Lasso DC	90.2	98.0	99.3	**100.0**
Lasso DCE	90.6	97.9	99.2	**100.0**
Lasso RCEs	89.8	98.2	99.2	**100.0**
Lasso RCE$\frac{1}{4}$	90.6	98.2	99.6	**100.0**
Lasso RCE$\frac{1}{2}$	91.3	98.5	99.6	**100.0**
Lasso RCEa	**94.2**	98.5	99.2	**100.0**
Lasso DCE+ RCEs	**91.2**	**98.4**	**99.4**	**100.0**
Lasso DCE+ RCE$\frac{1}{4}$	91.5	98.4	**99.9**	**100.0**
Lasso DCE+ RCE$\frac{1}{2}$	92.5	**98.7**	99.8	**100.0**
Lasso DCE+ RCEa	93.8	98.3	99.6	**100.0**

Table 5.3 – Closed world results on PRID2011 dataset.
CMC values at ranks $1, 5, 10, 20$ are reported. Best results are in bold red. Best results where only one probe person is provided at a time are in bold blue.

In any case, it performs much better than XQDA.

Influence of the number of probe people's images simultaneously available

When several probe people's images are available at test time, the performances are improved. There more probe people's images are available, the better the results. In the closed world setting, for both iLIDS VID and PRID, there is an increase of around 4% between the case when there is only one person's images available at a time (Lasso RCEs) and the case when all probe images are simultaneously available (Lasso RCEa). The performances of intermediate cases (Lasso RCE$\frac{1}{4}$ and Lasso RCE$\frac{1}{2}$) are indeed in between the two extreme cases. In the open world case, the improvement induced by the simultaneous availability of more probe people's images at test time is more significant. In the case of iLIDs-VID, the increase is progressive with a DIR at first rank for a FAR value of 1% of 13.9% for Lasso RCEs, of 17.7% for Lasso RCE$\frac{1}{4}$, of 19.4% forLasso RCE$\frac{1}{2}$ and finally of 22.0% for Lasso RCEa. In the case of PRID, the increase is not significant between Lasso RCEs and Lasso RCE$\frac{1}{2}$, but there is a gap of +13.5% between Lasso RCE$\frac{1}{2}$ and Lasso RCEa. The increase is more progressive if we look at the results for FAR = 10%. The reason why there is such an improvement with the availability of more probe people's images is because some of those probe people indeed corresponds to some of the gallery identities.

Relevance of a reciprocal relation

Both direct and reverse directions sparse coding show great improvement compared to XQDA metric learning alone and lead to similar results, so we could wonder if they actually contain the same information.

FAR	1	10	50	100
MDTS-DTW [47]	12.7	32.6	51.8	57.3
DVR [9]	17.3	29.1	49.9	57.8
XQDA[23]	5.1	15.2	45.3	59.1
Lasso DC	12.9	35.1	58.8	68.5
Lasso DCE	17.2	37.5	62.8	69.0
Lasso RCEs	13.9	35.6	61.5	69.2
Lasso RCE$\frac{1}{4}$	17.7	40.3	63.7	70.7
Lasso RCE$\frac{1}{2}$	19.4	41.4	65.5	71.3
Lasso RCEa	22.0	44.7	67.4	72.7
Lasso DCE+RCEs	**18.0**	**40.5**	**66.5**	**71.6**
Lasso DCE+RCE$\frac{1}{4}$	20.7	41.9	67.0	71.6
Lasso DCE+RCE$\frac{1}{2}$	21.5	44.1	**68.2**	72.4
Lasso DCE+RCEa	**23.8**	**45.8**	68.1	**72.9**

Table 5.4 – Open world results on iLIDS-VID dataset.
DIR values at rank 1 for different values of FAR (1%, 10%,50% and 100%) are presented. Best results are in bold red. Best results where only one probe person is provided at a time are in bold blue.

For closed world scenarios, most often the combination of direct and reverse direction residual errors does improve the performances but it never improves the results for more than 3% for first rank recognition compared to the best of the two combined methods taken alone (Lasso DCE and Lasso RCE).

For open world scenarios, the increase in performances brought by the combination of the residual errors from the two directions is more significant. It is especially the case for PRID2011 where the improvement of the reverse direction approaches due only to the simultaneous availability of several probe identities's images (Lasso RCE$\frac{1}{4}$ and Lasso RCE$\frac{1}{2}$) wasn't so visible, but which becomes clear when combined with the direct direction approach. On one hand, there is a huge gap of around $13 - 14\%$ between the Lasso RCEa approach and the three variants Lasso RCEa, Lasso RCE$\frac{1}{4}$ and Lasso RCE$\frac{1}{2}$, on the other hand, the improvement is progressive from Lasso DCE+RCEs to Lasso DCE+RCEa.

These results support the idea that for a given person present in both the gallery and the probe set, the set of most similar identities in the other set are not exactly the same, and combining the results from both directions makes the results more robust by retrieving first the identities that appear in both directions list of most likely matches and thus leveraging the ambiguities that exist in each direction.

5.6.2 Influence of the choice of the additional dictionaries

In this section, we discuss different choices of additional dictionaries in order to validate our proposed sparse coding based selection of training elements to form the additional dictionaries. We consider the case when only one probe person's images are available at a time and we examine the following variants:

FAR	1	10	50	100
MDTS-DTW [47]	42.7	55.2	70.5	72.8
DVR [9]	46.8	58.3	78.3	79.7
XQDA[23]	21.2	40.7	78.8	90.5
Lasso DC	49.8	69.3	88.2	93.8
Lasso DCE	55.7	71.0	90.2	93.2
Lasso RCEs	58.0	69.7	91.0	93.0
Lasso RCE$\frac{1}{4}$	58.8	71.2	91.3	93.5
Lasso RCE$\frac{1}{2}$	59.7	79.0	92.0	94.0
Lasso RCEa	73.2	**87.3**	**95.3**	**96.0**
Lasso DCE+RCEs	**60.8**	**76.0**	**91.7**	**93.3**
Lasso DCE+RCE$\frac{1}{4}$	62.7	75.8	92.2	93.7
Lasso DCE+RCE$\frac{1}{2}$	66.7	78.8	93.2	94.7
Lasso DCE+RCEa	**73.3**	83.2	94.0	95.2

Table 5.5 – Open world results on PRID2011 dataset.
Open world DIR values at rank 1 for different values of FAR (1%, 10%,50% and 100%) are presented. Best results are in bold red. Best results where only one probe person is provided at a time are in bold blue.

- **Lasso RNC**. In this Lasso Reverse direction Non Collaborative approach, for each presented probe person, only this probe person's features are used to compute the sparse code of gallery features. There is no collaboration involved and reconstruction errors are used as dissimilarity scores.

- **Lasso RCEv1**. In this Lasso Reverse direction Collaboration Enhanced Variant 1 approach, the sparse representations of gallery features are computed using a collaborative dictionary composed of the probe person's features and an additional dictionary composed of features selected from probe images of the training set. Instead of selecting from the probe training features, elements that are similar to gallery features, ie. elements that participate in the reconstruction of gallery features, we form the additional dictionary by selecting a fixed number of elements that do not participate to any of the approximations of gallery features. We limit the size of this additional dictionary to the number of identities in the training set, ie. for 89 PRID2011 and 150 for iLIDS VID. Mean residual errors are used as dissimilarity scores.

- **Lasso RCEv2**. In this Lasso Reverse direction Collaboration Enhanced Variant 2 approach, the sparse representations of gallery features are computed using a collaborative dictionary composed of the probe person's features and an additional dictionary composed of one feature per probe training person. Mean residual errors are used as dissimilarity scores. The size of the additional dictionary is the number of identities in the training set.

- **Lasso RCEs**. The Lasso Reverse Collaboration Enhanced approach is the approach we propose in this thesis. The sparse code of gallery features are computed using a collaborative dictionary composed of probe features and an additional dictionary which differs for each gallery person. Mean residual errors are used as dissimilarity scores. The elements which compose the addi-

tional dictionary specific to a gallery person are the probe training features which participate in the reconstruction of that gallery person's features when all probe training features are put into competition. Therefore the size of the additional dictionary can vary depending on the gallery identity and ranges between around 30 and 100 in our experiments.

- **Lasso RCEv3**. In this Lasso Reverse direction Collaboration Enhanced Variant 3 approach, the additional dictionary is the same for every gallery person and is formed by all the training probe features T_p.

The closed world CMC evaluations are reported in Table 5.6 and the open world DIR vs FAR evaluations are reported in Table 5.7. The ROC curves are displayed in Figure 5.4. The distribution of the residual errors corresponding to positive and negative pairs are shown in Figure 5.5. The TP rate and TN rate are presented in Figure 5.6.

Dataset	PRID2011				iLIDS VID			
Rank	1	5	10	20	1	5	10	20
Lasso RNC	84.7	97.1	99.1	99.9	53.7	80.7	87.2	93.3
Lasso RCEv1 (dissimilar)	84.7	97.1	99.1	99.8	53.7	80.7	87.2	93.3
Lasso RCEv2 (1 image per train id.)	89.8	**98.7**	**99.9**	**100.0**	64.2	87.8	93.7	97.0
Lasso RCEs (similar)	89.8	98.2	99.2	**100.0**	65.4	**88.3**	**93.9**	**96.8**
Lasso RCEv3 (all probe train images)	**89.9**	98.2	99.2	**100.0**	**65.5**	**88.3**	**93.9**	**96.8**

Table 5.6 – Closed world results. Comparison of 5 reverse direction sparse coding approaches. CMC values at rank $1, 5, 10$ and 20 are reported on PRID2011 and iLIDS VID. Best results are in bold blue because we consider here the case when only one probe person's images are available at a time.

Dataset	PRID2011				iLIDS VID			
FAR(%)	1	10	20	100	1	10	20	100
Lasso RNC	20.0	36.3	76.3	89.7	4.4	19.0	46.4	58.4
Lasso RCEv1 (dissimilar)	20.3	35.5	75.0	88.5	4.4	19.0	46.4	58.3
Lasso RCEv2 (1 image per train id.)	43.8	61.3	88.0	92.5	9.9	31.6	59.6	68.3
Lasso RCEs (similar)	**58.0**	69.7	**91.0**	**93.0**	13.9	35.6	61.5	69.2
Lasso RCEv3 (all probe train images)	**58.0**	**70.0**	90.8	**93.0**	**14.1**	**35.8**	**61.8**	**69.3**

Table 5.7 – Open world results. Comparison of 5 reverse direction sparse coding approaches. We report for PRID2011 and iLIDS-VID, DIR at first rank versus FAR, when FAR takes values $1\%, 10\%, 20\%$ and 100%. Best results are in bold blue because we consider here the case when only one probe person's images are available at a time.

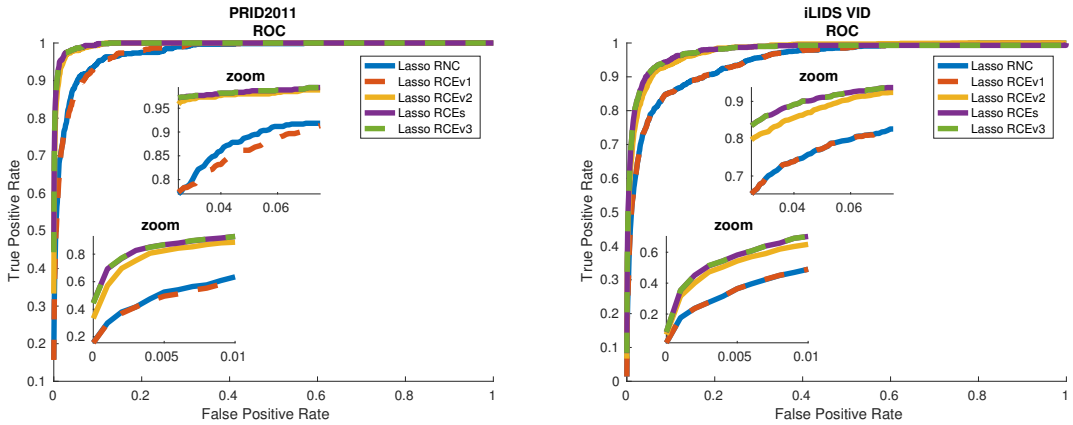


Figure 5.4 – Comparison of the ROC curves of 5 reverse direction sparse coding approaches (Lasso RNC, Lasso RCEv1, Lasso RCEv2, Lasso RCEs, Lasso RCEv3) on PRID2011 and iLIDS VID datasets.

A quick look at Figures 5.4, 5.5, and 5.6 and Tables 5.6 and 5.7 enables to group the five variants into three groups. The first group is composed of the approaches Lasso RNC and Lasso RCEv1. The second group is formed of only one approach, Lasso RCEv2. The third group contains Lasso RCEs and RCEv3. If we were to group them into only two groups, we would merge the second and third groups.

Lasso RNC, Lasso RCEv1: no collaboration and useless collaboration

In every aspects, the results obtained for the collaboration enhanced reverse sparse coding approach Lasso RCEv1 where the additional dictionary is composed of probe training features that do not participate to any of the gallery features's reconstruction are quasi identical to the results obtained with the non collaborative reverse sparse coding approach Lasso RNC. Except in Figure 5.6 where we can observe for the PRID2011 dataset that the TP rate and TN rate of the collaboration enhanced Lasso RCEv1 approach (in red) are pushed towards the right compared to those of the non collaborative approach Lasso RNC (in blue), ie. the residual errors obtained with Lasso RCEv1 are bigger than the reconstruction errors obtained with Lasso RNC, everything else is similar. The distribution of the dissimilarity scores overlap over a large interval and the distribution of negative pairs score spread over a larger interval than the distribution of positive pairs score, similarly to the direct direction non collaborative sparse coding approach Lasso DNC. The ROC curves of Lasso RNC and Lasso RCEv1 are much lower than those obtained with other choices of additional dictionaries. The closed world CMC values and the open world DIR vs FAR performances are also much below the other approaches for both PRID2011 and iLIDS-VID, with for example a difference of around 30% for DIR at first rank on PRID2011 dataset. These results confirm that it is useless to force a collaboration between test probe features and an additional dictionary composed of train features that are unlikely to participate in the reconstruction of gallery elements. The results are similar to the case when no collaboration is involved.

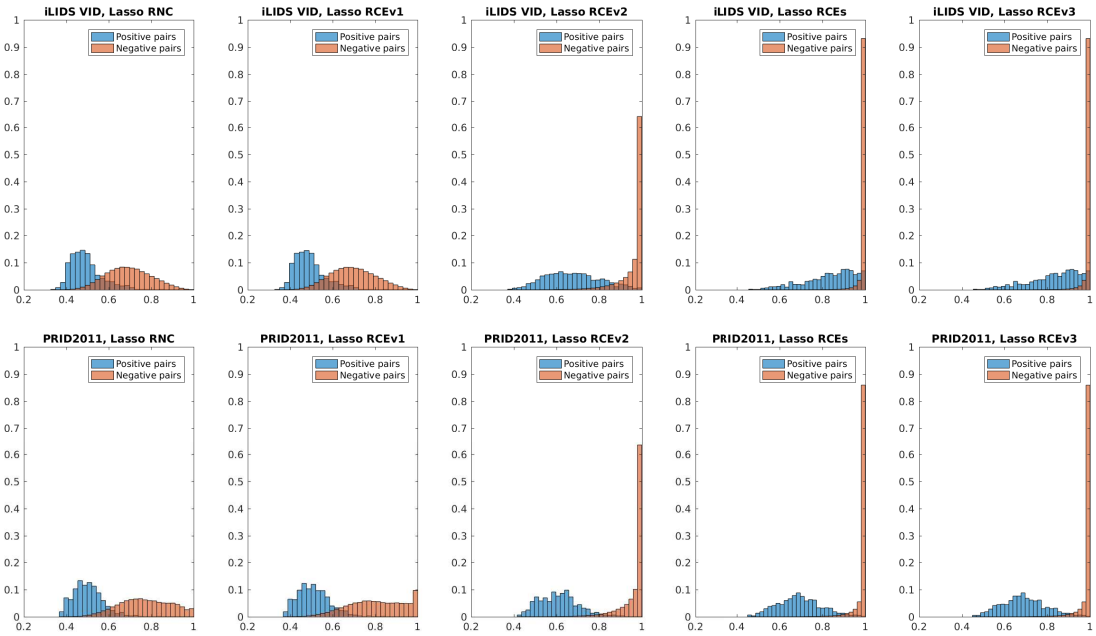


Figure 5.5 – Comparison of the distributions of positive and negative pairs dissimilarity score of 5 reverse direction sparse coding approaches (Lasso RNC, Lasso RCEv1, Lasso RCEv2, Lasso RCEs, Lasso RCEv3) on PRID2011 (top) and iLIDS VID (bottom) datasets.

Lasso RCEv3, Lasso RCEs: full collaboration and useful collaboration

In section 5.3, we discussed the meaning of residual errors and their use for ranking tasks. We evoked the necessity of a common additional dictionary for every gallery identity instead of a specific additional dictionary associated to each gallery identity so that the residual errors associated to a given probe person can be compared and have an interpretable meaning. However, the results shown in Figures 5.4, 5.5, and 5.6 and Tables 5.6 and 5.7 confirm that the way we form the additional dictionaries specific for each gallery person (Lasso RCEs) give similar results to the case when all training features corresponding to probe camera images are put into a commun additional dictionary used for the reconstruction of all gallery people (Lasso RCEv3). The results are not completely identical but this is due to very minor differences. Using the whole training probe features as additional dictionaries gives similar or slightly better results, with a difference of recognition rate of at most 0.3% in the open world case and at most 0.1% in the closed world case. These differences are not even observable on the ROC curve in Figure 5.4 nor on the positive and negative pairs distributions in Figure 5.6, even on the zoomed parts. Forming the additional dictionary of a gallery person by selecting from the training set the most similar elements using sparse coding (Lasso RCEs) leads to open world results that are much better than the non collaborative reverse direction sparse coding approach Lasso RNC with an improvement in the open world case of +28% for PRID2011 and +9.7% for iLIDS. In Lasso RCEv3, the features of the whole training probe set are used (around 10000 elements) whereas in Lasso RCEs , the additional dictionary size is often smaller than 100 due to the sparsity of the selection. Yet the results

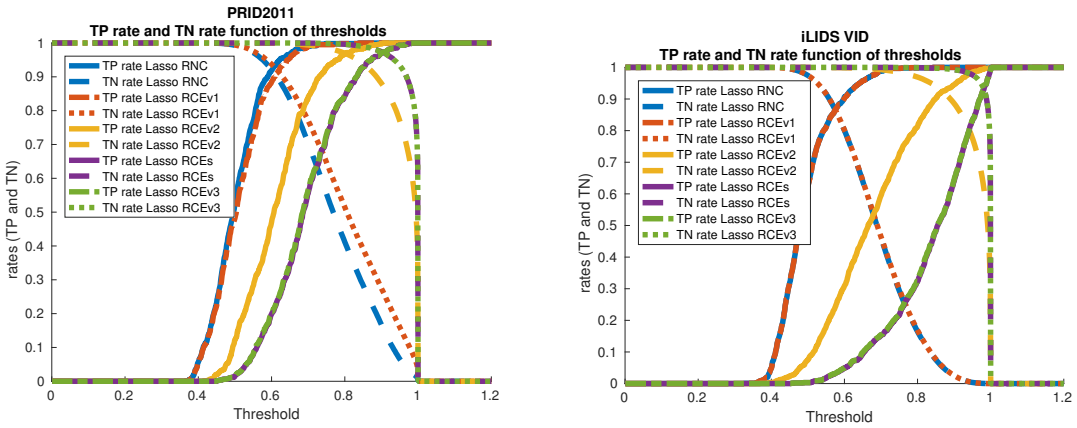


Figure 5.6 – Comparison of th TP rates and TN rates of 5 reverse direction sparse coding approaches (Lasso RNC, Lasso RCEv1, Lasso RCEv2, Lasso RCEs, Lasso RCEv3) on PRID2011 (top) and iLIDS VID (bottom) datasets.

of Lasso RCEs and RCEv3 are quasi identical in every aspect. This demonstrates the relevance of the selection of training elements we proposed. There is no need to use the whole training set as additional dictionary and selecting for each gallery person the most similar elements enables to force useful collaboration during the online re-identification process.

Lasso RCEv2: an intermediate case

The Lasso RCEv2 is a variant of the Lasso RCE approach in which the additional collaborative dictionary is also composed of elements selected from the training probe set but which does not require to compute the sparse code of gallery elements using training probe elements. The additional collaborative dictionary is commun to every gallery person and is formed by selecting one image per probe training identity.

In terms of CMC, Lasso RCEs and the Lasso RCEv2 variant are similar, but in terms of DIR vs FAR the results differ more significantly (difference of 14% on PRID2011 and of 4% on iLIDS-VID in the open world case). We already came across a similar situation when comparing Lasso DC and Lasso DCE, and the same reason applies to this case as well. Both Lasso RCEs and Lasso RCEv2 are collaboration enhanced sparse coding approaches, where the additional collaborative dictionary contains elements that are likely to be used in the reconstruction of gallery elements. The additional dictionary of Lasso RCEs contains more of these useful elements because they have been specifically selected due to their similarity to the gallery features while the Lasso RCEv2 contains one element randomly selected from each training identity. Therefore the residual errors of Lasso RCEs are in general bigger than the residual errors of Lasso RCEv2 (cf. Figure 5.6) but this does not influence the relative ranking of gallery identities, ie. it does not have any significant impact on the CMC evaluation.

In Lasso RCEv2, the additional dictionary is composed of as many elements as there are identities in the training set, ie. 150 for iLIDS-VID and 89 for PRID2011. In Lasso RCEs, the additional dictionaries result from a sparse coding selection approach, so their size vary and are most of the time much smaller than 100. Even though the additional collaborative dictionaries in Lasso RCEv2 are bigger than in Lasso RCEs, the residual errors are smaller for positive and negative pairs. This is because for each gallery person only the elements in the collaborative dictionary which participate in their reconstruction of the gallery features matter. Overall, for each gallery person, this number of elements is smaller in the case of Lasso RCEv2 than in the case of Lasso RCEs, even if the size of the collaborative dictionary is bigger. The results obtained with Lasso RCEs are better than with Lasso RCEv2, but Lasso RCEv2 also performs well, much better than XQDA and non collaborative sparse coding Lasso RCN.

5.6.3 Influence of the number of probe identities simultaneously available

In Figure 5.7, are plotted the TP rates and TN rates for the reverse direction collaborative sparse coding for different proportions of test probe identities available at a time at test time. The TP and TN rates are shifted to the right when the collaborative dictionary gets bigger. This is because when more elements are allowed to participate in the reconstructions, the contributions of the probe people we consider are bound to participate at most as much as they would participate without the presence of the additional elements, and most of the time their participation will be reduced, and the residual errors will thus be bigger.

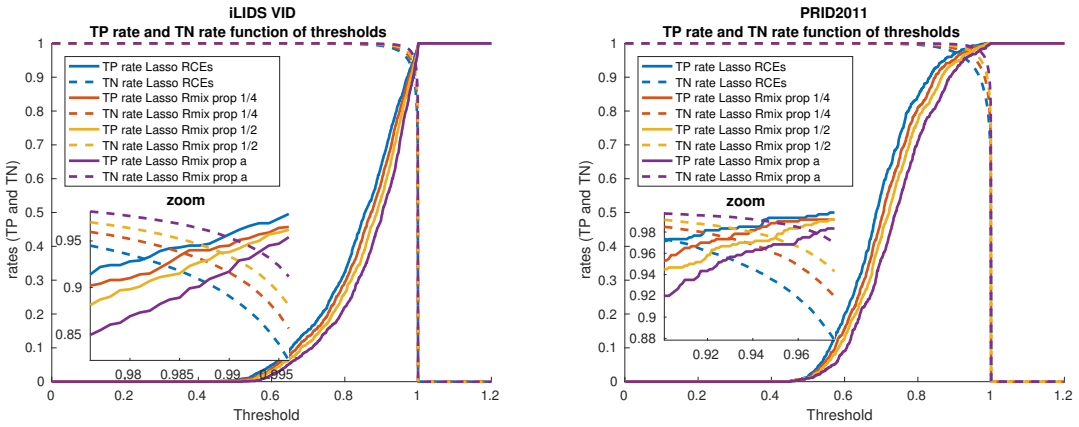


Figure 5.7 – TP rate and TN rate for on iLIDS and PRID datasets. Comparison of Lasso RCE approach for different proportions of simultaneously available probe people's images.

The ROC curves in Figure 5.8 show that in terms of proportion of well retrieved True Positive Pairs for given proportions of wrongly retrieved False Positive Pairs, there is no big difference between the different cases of applications (more or less probe people's images simultaneously available) except for low values of FAR for which having more probe people's images at disposal lead of higher TP rates. What

makes the difference between these cases is the presence among all available probe images, of probe identities that are also in the gallery and which enables to reduce the amount of False Positive that have very low residual errors.

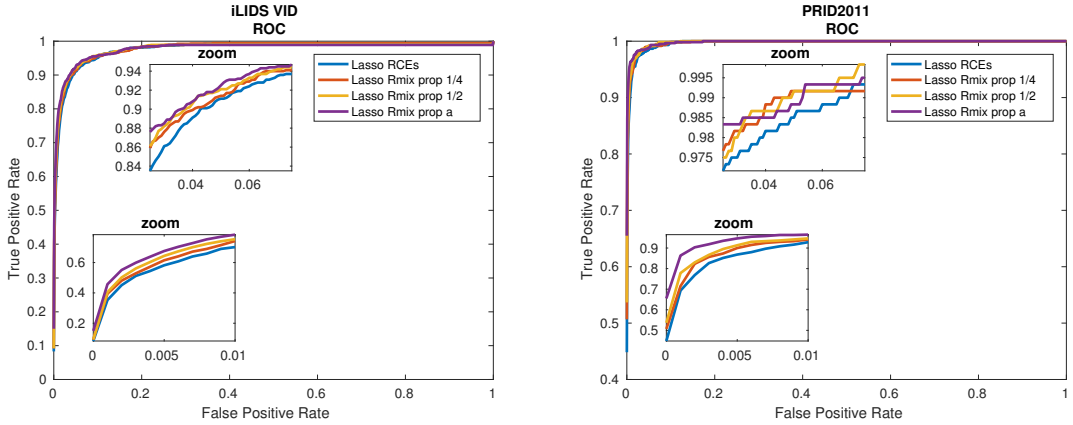


Figure 5.8 – ROC curves for collaboration enhanced sparse representation in the reverse direction on iLIDS and PRID datasets. The variants correspond to different practical cases when only one probe person's images are available at a time, or when the images are simultaneously available for a quarter, a half or for all of the probe people to be re-identified.

5.6.4 Evaluation on the person verification task

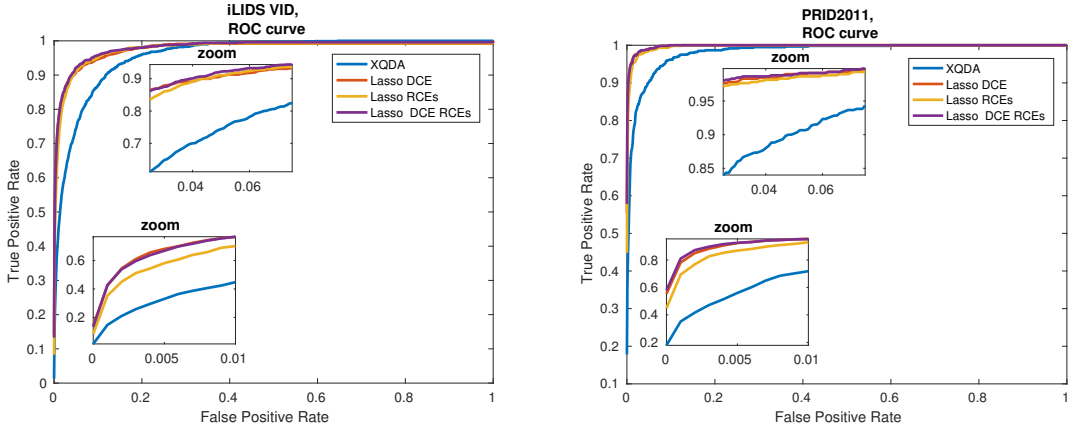


Figure 5.9 – ROC curves for XQDA, Lasso DCE, Lasso RCEs and Lasso DCE+RCEs on iLIDS VID (left) and PRID2011 (right) for the open world scenario.

The Figure 5.9 presents for iLIDS VID and PRID2011, the ROC curves for XQDA, for the direct direction approach Lasso DCE, for the reverse direction approach Lasso RCEs and for the bidirectional approach Lasso DCE+RCEs. Similarly to Lasso DCE, we can observe that the ROC curve of Lasso RCEs is high above the ROC curve of XQDA meaning that we retrieve more true positive pairs for the same false positive rates. The ROC curves of the direct direction approach Lasso DCE and the reverse direction approach Lasso RCEs are actually really similar. We can

see that the ROC curve of the reverse direction approach is slightly below that of the direct direction approach. When zooming, we can observe that the ROC value corresponding to the combination of the two approaches is imperceptibly above the ROC curve of each approach taken separately.

In Figure 5.9, we observed that the TP rates for small FP rates are bigger for Lasso DCE than for Lasso RCEs for both iLIDS VID and PRID2011. We have also seen in the re-identification task section 5.6.1 that the DIR at first rank when FAR equals 1% is higher for Lasso RCEs than for Lasso DCE on the PRID2011 dataset. This might seem contradictory but it is possible because for a probe imposter, FAR only takes into account the score (distance/reconstruction error/residual error...) of the first wrongly matched gallery identity. The other wrong matches of a probe imposter are not taken into account. Moreover wrong matches corresponding to non imposters probe people are not considered either when reporting DIR at first rank. Therefore in our case, even if the first rank detection and re-identification rate is higher for a given FAR value for Lasso RCEs than for Lasso DCE, when we consider the pairs of probe-gallery identities which have small residual errors, there is relatively more wrong matches in the Lasso RCE approach than in the Lasso DCE approach.

Let's now take a look at the TP rate and TN rate presented in Figure 5.10 corresponding to the Lasso DCE, Lasso RCEs and Lasso DCE+RCEs approaches. For a better visualization, we also zoomed on the TP and TN rates when the threshold gets close to 1.
The shape of the TP rates and TN rates of all three collaborative sparse coding approaches (Lasso DCE, Lasso RCEs and Lasso DCE+RCEs) are similar, especially with the TN rates which stay high in almost the whole interval $[0, 1]$ and only starts to drop when the threshold gets closer to 1. On the iLIDS VID plot, we observe that the residual errors obtained with the reverse direction approach (yellow curves) are smaller than those obtained with the direct direction approach (red curves). For the same value of threshold, the TP rate (solid lines) of Lasso RCE (yellow) is above that of Lasso DCE (red). On the zooms, the TN rate (dash lines) of Lasso RCE (yellow) are below that of Lasso DCE (red), ie. for the same value of threshold, there are more False Positives for Lasso RCE than for Lasso DCE. For a same decision threshold, the reverse direction sparse coding approach Lasso RCEs rejects less wrong matches than the direct direction collaborative sparse coding approach. Lasso DCE. Since the Lasso DCE+RCEs combined the direct and reverse direction approaches by simply adding the residual errors, for the same decision threshold, The number of pairs detected as false matches (TN+FN) is an intermediate value between that of Lasso DCE and Lasso RCEs.

In Tables 5.8 and 5.9 are reported the recall, specificity and classification rate and precision values for the Lasso DCE, Lasso RCE and Lasso DCE+RCEs approaches for three choices of decision threshold $(0.999, 0.950$ and $0.900)$. As we have just mentioned when studying the TP and TN rates, for the same decision threshold, the proportion of found TP pairs and found TN pairs differs. When the decision

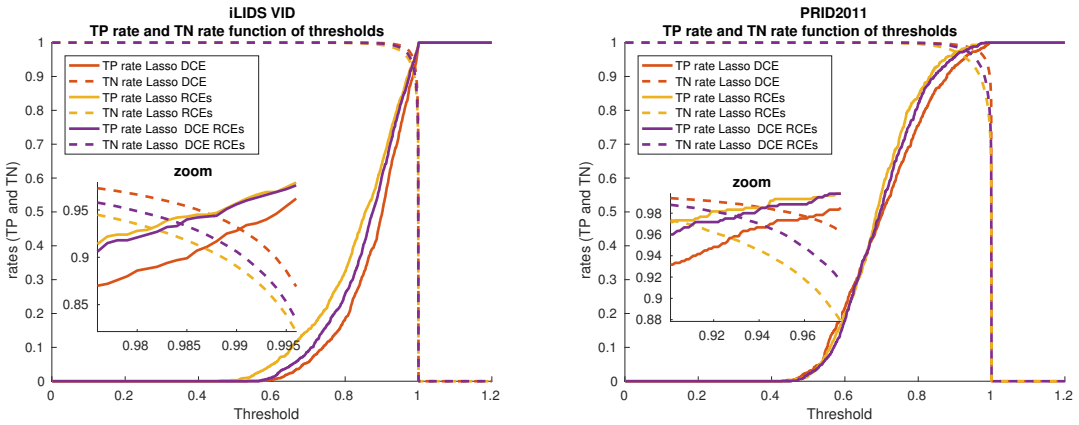


Figure 5.10 – TPrate and TN rate for Lasso DCE, Lasso RCEs and Lasso DCE+RCEs on iLIDS VID (left) and PRID2011 (right) for the open world scenario.

threshold is bigger, we find more right matches, but we also wrongly accept more false matches so the precision decreases. If we look at the results for a decision threshold equal to 0.999 on the PRID2011 dataset, it seems that the direct direction Lasso DCE approach is better fitted for the verification task than the reverse direction approach Lasso RCEs or the bidirectional approach Lasso DCE+RCEs.

Thres	Method	TP rate	TN rate	classif. rate	precision
0.999	DCE	97.8	81.6	81.7	3.4
	DCE+RCEs	98.4	77.4	77.6	2.8
	RCEs	98.5	76.6	76.8	2.8
0.950	DCE	73.8	99.2	99.0	38.3
	DCE+RCEs	80.8	98.6	98.5	27.9
	RCEs	83.1	97.7	97.6	19.7
0.900	DCE	48.8	99.9	99.5	70.3
	DCE+RCEs	60.0	99.7	99.4	58.5
	RCEs	63.8	99.3	99.1	38.4

Table 5.8 – Recall, Specificity, Classification rate and Precision values for 3 choices of threshold on the iLIDS VID dataset.

Thres	Method	TP rate	TN rate	classif. rate	precision
0.999	DCE	100.0	85.8	85.9	7.4
	DCE+RCEs	100.0	72.0	72.3	3.9
	RCEs	100.0	68.2	68.5	3.5
0.950	DCE	97.3	98.2	98.2	38.7
	DCE+RCEs	98.8	96.2	96.3	23.2
	RCEs	99.3	93.5	93.5	15.0
0.900	DCE	92.8	99.4	99.4	65.3
	DCE+RCEs	95.7	98.8	98.8	48.7
	RCEs	96.8	97.5	97.5	30.6

Table 5.9 – Recall, Specificity, Classification rate and Precision values for 3 choices of threshold on the PRID2011 dataset

5.7 Conclusion

In this chapter we have presented another open world person re-identification method based on collaborative sparse coding but we adopted the probe's point of view rather than the gallery's point of view. Though the general idea of our collaboration enhancement is the same as in the previous chapter, the asymmetry of the re-identification problem where gallery identities are known while probe people can be anyone in an open world setting brought to light some interesting elements that allowed us to propose an appropriate selection of training elements to form additional collaborative dictionaries adapted to each gallery person.

Extensive experiments on iLIDS VID and PRID2011, have demonstrated the relevance of our reverse direction sparse coding approach, where we compute the sparse code of gallery elements rather than that of probe instances, as well as the relevance of our construction of additional collaborative dictionaries for each gallery person for both closed and open world re-identification tasks. Contrary to the direct direction case where the additional dictionary acts as a collaboration enhancement, in the reverse direction, the additional dictionaries are required because since probe identities can be provided one by one and not simultaneously there is no "natural" collaboration between probe identities.

In order to take into account both the probe person and the gallery identities's point of view and make sure that there is a strong reciprocal similarity between the probe person and the top ranked gallery identities, we propose to combine the two approaches (direct and reverse direction). Because the two sets of residual errors are in the same range and present similar distributions, combining the residual errors from the direct and the reverse direction by a simple sum is relevant and gives the same importance to the probe and the gallery point of view. The experiments show that combining the two approaches into one bidirectional approach improves the performances for both closed and open world scenarios.

Depending on the applications, one or several probe people's images can be simultaneously available. If only one person's images is provided, the proposed method shows good performances in the closed settings and it already outperforms other

state-of-the-art methods in the open world settings (+14% for PRID2011). When available, the images of the several probe people are jointly used in the collaborative dictionary. The more identities are simultaneously available, the less computations for a same number of people identified and the better the results.

Chapter 6

Conclusion and Perspectives

6.1 Conclusion

The person re-identification task essentially based on surveillance camera images started as a subtask of multi-camera tracking ten years ago and has tremendously evolved since then. At first, the person re-identification task only involved two cameras, with identities who always appeared once in each camera. Today several generalization of this problem are tackled by the person re-identification community. In this thesis, we aimed at well exploiting the multiple images available for each identity but our main focus was the open world aspect of person re-identification.

The open world person re-identification task we tackled is defined by two subtasks, the detection subtask and the re-identification subtask. In the detection task, the goal is to determine whether a probe person is likely to be one of the gallery people. The object of the re-identification subtask is to rank the gallery identities that have been considered to be possible matches for the presented probe person. Contrary to the closed world case where a relative ranking of gallery identities from the most similar to the most dissimilar one was enough, the open world re-id task requires a decision rule to separate possible matches from unlikely matches.

Considering that when the verification task is completely solved, ie. when we are able to tell apart positive pairs from negative pairs, then the re-identification task (ranking task) should consequently also be resolved, we propose a first method called COPReV which casts the open world re-identification task as a binary classification task. Through the optimization of an objective function which balances the penalization on positive and negative pairs and encourages pairs distances to be far from a predefined threshold, we learn a linear projection of the features that enables to separate positive pairs and negative pairs by a simple threshold rule. After applying our COPReV subspace learning approach to LOMO [23] and Inception-Resnet-v2 [117] features projected with XQDA [23], the distributions of positive pairs and negative pairs distance overlapped less. This resulted in a gain in performances for both closed and open world person re-ie tasks. Moreover, COPReV can be used for person verification and there is no need to look for the right threshold for the decision rule. The threshold used during the training phase can also be used as

decision rule for the verification task and it leads to almost balanced true positive rate and true negative rate. However, for the re-identitication tasks, the gain in performance is not so significant with initially already good features. Moreover, for the open world case, the results are still far from state-of-the-art methods DVR [9] and MDTS-DTWt [47]. We explain these mitigated results by the fact that the verification task is not perfectly solved and therefore the improvement brought to re-identifications tasks though real are limited. Assuming that binary classification constraints alone could enable to tackle re-identification and verification tasks might have been too presumptuous.

In the light of this observation, our second contribution is based on collaborative sparse coding which is a tool that has good ranking capabilities and which we adapted so that it also became a decision tool in our re-id framework. The computation of the sparse code of probe elements using a collaborative dictionary composed of all the gallery identities puts gallery identities into competition against each other and only the most similar elements are selected. Since elements that are more similar participate more, collaborative sparse coding is a good ranking tool. It is also already partially a decision tool because most dissimilar elements do not participate in the reconstructions. However, in every case, even when no elements are similar to the element to be reconstructed, some elements will be selected. Moreover if there is one gallery person who is very different from the others and that the probe person has some similarity with him, the whole participation will be reported on that one gallery person, making him a very likely match even if they are not so similar. Our Lasso DCE method enhances the collaboration of the collaborative gallery dictionary by adding to it elements from the training set and this lead to significant improvement for the open world person re-identification task.

Our third contribution stems from the idea that a good match involves two elements each of which considers the other element as a good match. In the Lasso DCE approach, we only compared gallery identities among themselves and looked for the ones that were the most similar to the probe person. In this Lasso RCE approach, we propose to check for a gallery person whether the presented probe person is similar to him or not by approximating him using the probe person's features put in concurrence with elements that are selected from the training set and which are already the gallery person's most similar elements but which describe other identities. The probe person is assumed to be a possible match if he participates when he is put in concurrence with those elements that are already similar to the gallery person. Indeed, if he participates, it means he is as similar or more similar to the gallery person than elements that are already alike. The Lasso RCE approach alone has similar performances to the Lasso DCE approach. The combination of the two further boosts the results and we achieve remarkable detection and recognition rates for open world experiments. Moreover, the residual errors are bounded (between 0 and 1) and while the true positive rate increases almost linearly with the threshold, the true negative rate stays high for a wide range of threshold and only drops near 1, so one can easily find an appropriate decision threshold for his person verification application.

6.2 Perspectives

6.2.1 Design adapted features for sparse coding approaches

Developing a feature learning method specifically designed for the sparse representation approaches we proposed for matching is one of the research axes which should be privileged in a future work. Indeed, in this thesis, we mainly focused on the open world aspect of re-identification and the question of the feature description has only been evoked to show the effectiveness of our approaches for different features. However, these features were not specifically designed for a use in a sparse coding matching framework. The Inception-Resnet-v2 features were not even learnt for the person re-identification task but for classification tasks. As for the XQDA metric learning approach which we use as a projection, if applied without our bidirectional sparse coding approach, it performs remarkably well in the closed world setting, but when it comes to open world re-identification, the performances completely collapse. Even if we manage to obtain results which outperform other existing methods for the open world re-identification task using normalized XQDA projected LOMO features with our bidirectional sparse coding matching scheme, the use of adapted features could give even better results.

One possibility could be to use training data to jointly learn a linear projection of the features along with computing the sparse code and residual errors. A more interesting variant would be to use an existing neural network architecture designed for computing features for the re-identification task and replace the cost function of the network by a new cost function adapted for a sparse coding based matching step.

Moreover, instead of keeping the distinction between training images coming from the gallery or the probe camera, the training stage should enable to lessen the disparity between images coming from different cameras and allow for tackling multiple cameras (more than 2) scenarios.

6.2.2 Adapt the sparse coding framework to multi-camera scenarios

If some differences between camera views can be lessen by well designed features, it is likely that there will still be features corresponding to images from different people taken from the same camera which will be more similar than features corresponding to images from the same person captured by different cameras. For a multi-camera (more than 2) scenario, it is therefore important to tackle the issue of the camera provenance. When presented a probe images captured by camera A, should we look for the possible matches separately for each gallery camera B, C, D , etc., or should we put into competition the gallery identities appearing in all the gallery cameras? On one hand comparing the images from all the cameras together might favor camera views that are more similar to the probe camera view and consequently favor gallery identities which do appear in those gallery cameras, but on the other

hand, isn't the competition between identities and the sparsity constraints what made the collaborative sparse coding approach so performant for re-identification? Wouldn't it be a shame to find possible gallery matches per gallery camera and then use an ad hoc way of ranking these gallery identities selected from different cameras? These are important questions which need to be considered because in the real world, there are not just two cameras. People are captured by numerous cameras, in various places and tackling into account the differences due to the camera view is crucial. In that regards, the fact that the residual errors are always in the same range (between 0 and 1) is definitely an asset but it is not good enough. Depending on the pair of cameras considered, the best verification threshold might still vary.

6.2.3 Learn the additional dictionaries in the reverse direction sparse coding

For now, for the reverse direction sparse coding, we proposed to select elements from the training set to form one additional dictionary for each gallery identity. An issue with such a way of forming the additional dictionaries is that some gallery identities are better approximated by their associated additional dictionary than other gallery identities. We have no control over how similar the selected training elements are to the test gallery identities. This can lead to a disparity in the range value that the residual errors take and we have seen that while the range value of similarity or dissimilarity score do not matter in the closed world setting, for open world re-identification and person verification tasks it plays an important role. In other words, for a given gallery person, we compared a probe person to the people in the training set who are the most similar to the gallery person, but the similarity between those people in the training set and the gallery person is not quantified. It would be better to compare a probe person to instances whose similarity to the gallery person is somehow quantified. Therefore the next step to improve the reverse direct Lasso RCE approach we proposed in this thesis, is to find a way to create additional dictionaries which would enable to obtain the same residual errors value if presented images with the same level of similarity to a gallery identity. While we selected instances from the training set, a learning phase could also be based on training data, but it might also be possible to learn dictionaries directly from the test gallery set.

6.2.4 Learn the additional dictionary in the direct direction sparse coding

Contrary to the reverse direction where we have shown the relevance of the chosen selection process for the additional dictionaries, for the direct direction, we simply proposed to use the whole training data available as additional dictionary. Though the results show that it does improve significantly the performances for the open world re-identification task, there is room for improvement. Indeed, even if there exist methods which aim at solving Lasso problem efficiently even with large dictionaries, having smaller dictionaries from the beginning can reduce the computation time and memory load. An in-depth study should be conducted to assess the impact of the use of certain types of elements when forming the additional dictionary. In

particular, we could distinguish training elements that are similar to some gallery identities from training elements which are dissimilar to every gallery elements. Once the type of elements needed for forming a useful additional dictionary is clear, we could consider learning an additional dictionary rather than selecting elements from the training set to form it.

6.2.5 Complexity and speed considerations for huge datasets

In this thesis, we used small dimension features (dimension 1536 for Inception) and projected features especially when the dimension of the features were too big (dimension 26960 for LOMO features). This was of course for performance reasons but also for time consumption problem as well. For these small dimension features, there were no memory nor speed issues for the datasets we perform tests on. Indeed, even if some tracklets are composed of several hundred images, iLIDS-VID and PRID2011 datasets only contain a few hundred identities. Actually even if existing person re-identification datasets have increased in size, they are still quite small with only a few thousand identities for the biggest datasets. This roughly corresponds to the number of students in a Parisian high school. Nonetheless, when deployed in the real word, it isn't just a few thousand people that the algorithm will have to deal with, it could be million. The question of the scalability becomes crucial. Reducing the size of the dictionaries with a better management of multi-shot data could be part of the solution. Multi-shot data gives richer and more complete descriptions but a good selection of these images might enable to keep only the important information. The tracklets could be preprocessed so that redundant images, images with occluded people and images where people are not well centered are removed.

6.2.6 A better use of simultaneously appearing people

In the reverse sparse coding approach, we have observed that when several probe identities' images were simultaneously available and were therefore used together for computing the sparse representation of gallery features, the performances were improved. Yet, we did not fully exploit the information that we had several distinct probe people's images. Similarly to some re-ranking approaches, if a gallery person is already the best match for one of the probe people, he should not also the best match for another distinct probe person.

6.2.7 Generalize even more the re-identification task: dynamic set of identities

The main focus of the thesis was the open world issue with the decision and ranking aspects to take into account. While this task is already much more realistic than the closed world re-identification task, the gallery set is still static. Even when new identities are detected, they are not added to the base of known identities, so if these people are observed once more, they will still be rejected as unknown people. Many new issues will arise with dynamic sets of identities re-identification tasks. Indeed, if identities are detected as yet unknown, they should be given a new label and be added into the set of known identities. However, there can be mistakes. Known

people might be detected as unknown and be added under a new identity label into the set of known identities. Algorithms will have to dynamically question decisions made about previously seen people. Some people might have to be fused into a single one while others will be splitted into two distinct people.

Just as closed world re-identification task was assessed by the CMC measure and the open world re-identification task we dealt with by the DIR versus FAR measure, new evaluation measures will have to be defined for new scenarios. The definition of the evaluation measure is critical because it will dictate the orientation of the approaches which will be developed to tackle the new tasks.

Chapter 7

Résumé en français

7.1 Introduction

Dans l'esprit collectif, "ré-identification" rime souvent avec "reconnaissance faciale". Or si ces domaines de la vision par ordinateur ont tous les deux pour but de trouver l'identité d'une personne, la reconnaissance faciale s'appuie sur des images hautes résolutions du visage de personnes prises dans des conditions particulières d'illumination et de pose qui requièrent la coopération de ces personnes, alors que la ré-identification de personnes s'appuie sur des images comportant les silhouettes de personnes apparaissant sur des images de vidéo surveillance ayant souvent une faible résolution et qui sont acquises sans soumettre les personnes à des contraintes spécifiques. Cette absence de contraintes a un prix. En effet, la reconnaissance faciale permet une identification des personnes à long terme. La ré-identification de personnes à partir d'images de vidéo surveillance quant à elle ne permet de retrouver l'identité d'une personne que dans un court lapse de temps, souvent d'une journée si l'on considère que la personne ne change pas de vêtements au cours de la journée.

C'est d'ailleurs de cette absence de contraintes que découle la plupart des difficultés associées à la ré-identification de personnes. Les personnes, les caméras et l'environnement sont autant de paramètres qui varient et qui mènent parfois à l'obtention d'images très différentes représentant une seule et même personne ou à l'acquisition d'images qui semblent très similaires alors qu'elles correspondent à plusieurs personnes différentes. En effet, selon le rendu couleur des caméras et l'illumination de la scène, les couleurs d'un même vêtement peuvent sembler très différentes. De plus une personne n'est pas un solide indéformable, ainsi on peut observer une grande variabilité dans la pose d'une personne. A cela s'ajoute les variations dues au positionnement de la caméra par rapport à la personne. Cette dernière peut être capturée de face, de profil ou de dos, par des caméras placées à différentes hauteurs ce qui mène à des problèmes d'alignement des images. Ces problèmes d'alignement sont parfois amplifiés par l'utilisation d'un mauvais détecteur de personnes. L'environnement joue également un rôle, avec en particulier des problématiques liées à l'arrière plan et les occlusions.

Le problème de la ré-identification de personne est un problème en constante

évolution. Littéralement, ré-identifier quelqu'un suppose que cette personne a déjà été identifiée auparavant, et il s'agit de retrouver son identité parmi l'ensemble des personnes déjà connues. On appelle galerie l'ensemble des personnes connues. On appelle personne requête une personne dont on souhaite retrouver l'identité. Dans la plupart des bases d'images utilisées pour la recherche en ré-identification de personnes, les images sont capturées par deux caméras. Les images des personnes connues proviennent d'une des caméras, tandis que les images des personnes requêtes proviennent de l'autre caméra. Les quelques dernières années, plusieurs nouvelles bases d'images pour la ré-identification ont été assemblées avec des images prises par plus de deux caméras. D'autre part, pendant longtemps, on comparait les performances des méthodes de ré-identification sur une base d'images où chaque personne n'était représentée que par une image par caméra (single-shot). De nos jours, nous disposons pour de bases d'images où plusieurs images par personnes par caméra sont disponibles (multi-shot) avec parfois des trajectoires complètes composées de plusieurs centaines d'images. Ces deux évolutions restent dans le cadre d'une base d'identité fermée. L'ensemble des identités, précédemment identifiées (galerie) et à ré-identifier (requête), est un ensemble fixe connu. Toutes les personnes à ré-identifier sont supposées être présentes dans la galerie. Cependant dans la vie réelle, il n'y a aucune garantie qu'une personne dont l'identité est à déterminer a déjà été identifiée auparavant et apparait déjà dans la galerie.

Dans cette thèse, nous nous intéresserons non seulement au problème de la base d'identités fermée, mais aussi et surtout au problème de la base d'identités ouverte, et ce dans le cadre de bases d'identités multi-shot. Lorsque la base d'identités est ouverte, cela signifie que la personne dont on souhaite connaître l'identité ne fait potentiellement pas partie de la base d'identités connues. Bien qu'il ne soit pas possible de ré-identifier une personne qui n'a pas été précédemment identifiée, on parlera tout de même de ré-identification en base ouverte pour évoquer le problème suivant composé de deux sous tâches, la détection et la ré-identification (classement). Comme pour le cas de la base fermée, nous disposons d'un ensemble d'identités connues, la galerie. La détection consiste à déterminer si la personne requête appartient ou non à l'ensemble des identités de la galerie. La ré-identification consiste à ranger les identités de la galerie selon leur similarité avec la personne requête.

7.2 Etat de l'art

La problématique de la base ouverte n'est apparue que très récemment et la plupart des approches de ré-identification de personnes ont donc été développées pour répondre aux exigences de la ré-identification en base fermée. Ainsi la littérature concernant la ré-identification en base ouverte est assez réduite avec à notre connaissance, moins de dix papiers traitant de cette problématique [1, 2, 3, 13, 46, 48, 9, 47]. Quant à la ré-identification de personnes en base fermée, la littérature est très vaste. On peut diviser les approches de ré-identification en trois grands groupes: les méthodes visant à développer des descripteurs discriminatifs [118, 17, 9], les

méthodes d'apprentissage de métrique [75, 20, 23], et les méthodes se basant sur l'apprentissage de réseaux de neurones [12, 39, 93].

Historiquement, les premières méthodes proposées pour la ré-identification de personnes sont des méthodes qui cherchent surtout à définir des descripteurs qui sont capables de différencier les personnes. Dans le premières méthodes, telles que SDALF [118] et Pictorial Structures [17], ces descripteurs sont définis de manière ad hoc et sont basés sur des idées intuitives. Les méthodes SDALF [118] et PS [17] proposent de détecter des parties du corps (tête, torse, jambes, etc.) afin d'extraire des informations couleurs et textures de ces parties. La comparaison de deux images consiste alors à comparer pour les différentes parties, les descripteurs de chacune des deux images. Puis des méthodes supervisées intégrant une phase d'apprentissage sont apparues. Au lieu d'utiliser des descripteurs ad hoc, certaines approches suggèrent d'utiliser tout un ensemble de descripteurs de base et d'apprendre lors de la phase d'apprentissage une manière de les combiner de manière optimale avec des méthodes de boosting [18] ou encore des forêts de classification [19]. Enfin, avec le développement des bases d'identités où chaque personne est représentée non plus par une seule image mais par une trajectoire, quelques méthodes récentes [9, 47] intégrent des informations temporelles en plus de la simple apparence.

Décrire une image est une étape indispensable pour la ré-identification, mais l'étape de comparaison de deux images ou de deux ensembles d'images est tout aussi importante. Les méthodes d'apprentissage de métrique ont pour but justement de développer de nouvelles métriques qui permettent de mieux discriminer les personnes qu'une simple distance euclidienne. Cet apprentissage de métrique s'effectue la plupart du temps en optimisant une fonction objective avec un terme de pénalisation des paires positives (même identité) et un terme de pénalisation des paires négatives (identités distinctes) [75, 20]. D'autres approches cherchent des solutions à des problèmes aux valeurs propres généralisés [21, 23]. Si ces méthodes d'apprentissage de métriques nécessitent des données d'apprentissage, elles sont également à l'origine d'un fort progrès en terme de performances.

L'arrivée en 2014 [12] de méthodes utilisant des réseaux de neurones a permis un nouveau bond des performances et depuis lors, de plus en plus de nouvelles approches sont basées sur les réseau de neurones. Des réseaux de neurones sont ainsi proposés en tant qu'outil pour extraire de nouveaux descripteurs [44, 45] ou encore comme méthode globale effectuant à la fois l'extraction de descripteurs et la comparaison de deux images [12, 39, 93]. L'inconvénient de telles approches est que la phase d'apprentissage nécessite un important volume de données sans lesquelles il est impossible d'apprendre les paramètres du réseau sans complètement sur-apprendre. C'est d'ailleurs la raison pour laquelle ce n'est que ces dernières années que de nouveaux datasets assez larges comprenant de nombreuses images ont fait leur apparition (CUHK [12], Market [10], DUKE [49]).

En ce qui concerne le problème de la base ouverte, on peut distinguer plusieurs

types de scénarios. Dans [1], non seulement l'hypothèse selon laquelle on connait toutes les personnes de la galerie est relâchée, mais cela est poussé encore plus loin. Il n'y a plus de galerie ni de regroupement des images d'une même personne apparaissant sur la même caméra. Le but est d'associer une identité à chaque image. Cela ressemble à un problème de clustering non supervisé où les différentes classes sont inconnues et leur nombre aussi. Dans [2, 3] un petit nombre d'identités est supposé connus et forme l'ensemble cible. Lorsqu'une personne requête est présentée, il faut déterminer si cette personne fait parti de l'ensemble cible ou non. Trouver l'identité précise de la personne requête n'est pas nécessaire. Dans [13, 46, 9, 47], la ré-identification de personnes en base d'identité ouverte est définie comme un problème avec deux sous-problèmes. Comme dans le cadre de la base fermée, nous disposons d'une galerie qui contient les identités connues. La détection permet de déterminer si une personne requête fait possiblement parti de la galerie ou non. Le classement permet de ranger par ordre de similarité les personnes de la galerie qui sont considérées comme de potentiels bons appariements. Enfin le papier [48] propose des scénarios encore plus généraux dans le cadre non plus de caméras fixes mais de caméras se déplaçant sur des drones. Il s'agit donc non plus seulement de reconnaître des personnes lorsqu'elles apparaissent sur des caméras distinctes mais également d'être capable de reconnaître une même personne lorsqu'elle est à nouveau détectée par la même caméra située sur un drone qui a changé de position, d'orientation, etc...

Les problèmes de ré-identification de personnes en base d'identités fermée et en base d'identités ouverte diffèrent en terme de scénario et donc en terme d'évaluation également. En base fermée, seul le rang auquel la bonne personne galerie est retrouvée est pris en compte dans l'évaluation CMC (Cumulative Match Curve). Pour différentes valeur de rang r, la CMC donne la proportion de personnes requête ré-identifiées dans les r premiers rangs, et ce peu importe la valeur du score de dissimilarité ou de similarité. En base d'identité ouverte, chaque scénario est associé à une mesure d'évaluation différente. Le scénario que nous étudions ici est la tâche de ré-identification en base ouverte décomposable en deux sous-tâches, la détection et le classement. L'évaluation associée consiste à fournir les valeurs de DIR (Detection and Identification Rate) pour différentes valeurs de FAR (False Acceptance Rate). Pour différents taux FAR de personnes requête n'apparaissant pas dans la galerie mais ayant été acceptées par erreur comme en en faisant partie, DIR fournit le taux de personnes requête faisant effectivement partie de la galerie et pour lesquelles le bon appariement est au premier rang.

7.3 COPReV

7.3.1 Présentation de la méthode

Dans cette section nous allons présenter notre méthode nommée COPReV pour "Closed and Open world Person Re-identification and Verification". L'idée de cette méthode est simple. Dans le cadre d'une base d'identités fermée, ranger une liste de personnes galerie par ordre de similarité permet d'obtenir une CMC parfaite, ie.

avec un taux de ré-identification de 100% dès le premier rang, dès lors que la bonne identité galerie est au premier rang pour chacune des personnes requête. Dans le cadre d'une base d'identités ouverte, cela n'est pas suffisant. En effet, s'il existe des personnes requêtes ne faisant pas partie de la galerie, mais pour lesquelles il existe une personne de la galerie pour laquelle le score de dissimilarité est plus petit (ou dont le score de similarité est plus grand) que celui d'une personne requête faisant partie de la galerie avec l'identité galerie correspondance, alors on aura une faible valeur de DIR pour des valeurs fixées de FAR. Par contre, si le score de dissimilarité des paires positives est toujours inférieur à un certain seuil, et que le score de dissimilarité des paires négatives est toujours supérieur à ce même seuil, alors le score de dissimilarité de toutes les paires positives sera plus petit que celui de n'importe quel paire négative. Ainsi les bons appariements seront toujours rangés avant les paires négatives. Cela permet donc de résoudre à la fois le problème de la ré-identification en base d'identités ouverte et en base d'identités fermée.

Fort de ces observations, la méthode COPReV consiste à apprendre une projection des descripteurs telle qu'avec les descripteurs projetés, la distance entre les descripteurs de deux images d'une même personne est inférieure à un certain seuil, et que la distance entre les descripteurs correspondant à deux personnes distinctes est supérieure à ce même seuil. Pour ce faire, nous proposons d'optimiser une fonction objective contenant deux termes, un terme pénalisant les paires positives ayant une distance supérieure au seuil fixé et un terme pénalisant les paires négatives ayant une distance inférieure au seuil fixé. Nous avons fait le choix d'une fonction logistique généralisée comme fonction de perte associée à chacun des termes de pénalisation. Une telle fonction a une forme en S et agit donc comme une fonction de comptage qui compte le nombre de paires mal classées. Afin de ne pas favoriser les paires négatives dont le nombre est bien supérieur au nombre de paires positives, et afin d'éviter de donner davantage d'importance aux identités qui sont représentées par davantage d'images, des pondérations sont appliquées aux deux termes de pénalisation. Des contraintes supplémentaires sur le comportement de la fonction de perte sont détaillées dans la partie écrite en anglais et permettent d'obtenir une fonction de perte n'ayant plus que deux paramètres.

Pour résumer, nous cherchons la matrice de projection L qui minimise la fonction objective suivante:

$$E(L) = \sum_{i \in \mathcal{I}} \left[\frac{1}{m_{ii}} \sum_{y \in D_{ii}} \mathcal{L}_+ \left(\|Ly\|_2^2 - \tau \right) + \frac{1}{K-1} \sum_{j \in \mathcal{I} \setminus i} \left(\frac{1}{m_{ij}} \sum_{y \in D_{ij}} \mathcal{L}_- \left(\tau - \|Ly\|_2^2 \right) \right) \right] \tag{7.1}$$

où $\mathcal{I}$ désigne l'ensemble des identités de l'ensemble d'apprentissage et K son cardinal. i et j font référence à des identités. D_{ii} représente l'ensemble des différence des descripteurs de deux images de la personne i, D_{ij} représente l'ensemble des différence des descripteurs de deux images, l'une représentant la personne i, l'autre représentant la personne j. τ est le seuil de décision qui permet de distinguer les paires positives et négatives. Les fonctions $\mathcal{L}_+$ et $\mathcal{L}_-$ sont les fonctions de perte

qui s'appliquent respectivement aux paires positives et aux paires négatives. Elles s'écrivent sous la forme suivante:

$$\mathcal{L}(z) = 1 - \frac{1}{(1+\nu e^{-\lambda z})^{\frac{1}{\nu}}} \quad \text{où} \quad \lambda < 0 \quad \text{et} \quad \nu > 1 \tag{7.2}$$

7.3.2 Résultats expérimentaux

Nous avons testé COPReV sur deux datasets, PRID2011 [11] et iLIDS VID [9]. Afin de montrer l'apport de notre approche, nous l'avons testé avec deux descripteurs: les descripteurs LOMO [23] et les descripteurs Inception-Resnet-v2 [117] abrégé par IR. Chacun de ces deux descripteurs ont au préalable subit une projection en utilisant la méthode d'apprentissage de métrique XQDA [23] considérée comme une projection plutôt qu'une distance.

Les résultats en base fermée sont présentés dans les tableaux 7.1 et 7.2.

Dataset	PRID2011			
Rang	1	5	10	20
MDTS-DTW [47]	69.6	89.4	94.3	97.9
DVR [9]	77.4	93.9	97.0	99.4
XQDA+IR	41.2	68.9	79.9	90.2
COPReV+IR	53.0	80.8	91.5	98.1
XQDA+LOMO[23]	**86.4**	**98.3**	**99.6**	**100.0**
COPReV+LOMO	82.8	97.8	**99.6**	**100.0**

Table 7.1 – Evaluation en base fermée sur PRID2011.
Les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 20. Les meilleurs résultats sont en rouge et en gras.

Dataset	iLIDS-VID			
Rang	1	5	10	20
MDTS-DTW [47]	49.5	75.7	84.5	91.9
DVR [9]	51.1	75.7	83.9	90.5
XQDA+IR	11.1	29.0	39.6	51.5
COPReV+IR	21.9	51.2	66.9	81.3
XQDA+LOMO[23]	**55.9**	**83.4**	90.5	96.1
COPReV+LOMO	53.9	**83.4**	**91.6**	**97.9**

Table 7.2 – Evaluation en base fermée sur iLIDS VID.
Les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 20. Les meilleurs résultats sont en rouge et en gras.

Pour PRID et iLIDS VID, l'amélioration entre XQDA et COPReV est assez importante pour les descripteurs IR avec plus de 10% de différence dans le taux de ré-identification sur quasiment tous les rangs. Avec les descripteurs LOMO, les résultats obtenus avec XQDA et COPReV sont assez similaires mais très légèrement

meilleurs pour les rangs plus élevés. En comparaison avec l'état de l'art, COPReV appliqué aux descripteurs LOMO donne de meilleurs résultats.

Les résultats en base ouverte sont présentés dans les tableaux 7.3 et 7.4.

Dataset	PRID2011			
FAR(%)	1	10	50	100
MDTS-DTW [47]	42.7	55.2	70.5	72.8
DVR [9]	**46.8**	**58.3**	78.3	79.7
XQDA+IR	3.0	8.7	24.7	47.7
COPReV+IR	8.3	15.8	40.0	60.5
XQDA+LOMO [23]	21.0	40.5	80.3	**90.3**
COPReV+LOMO	26.5	43.5	**81.0**	87.5

Table 7.3 – Evaluation en base ouverte sur PRID.
Les valeurs de DIR au rang 1 sont fournies pour les valeurs suivants de FAR: 1%, 10%,50% et 100%. Les meilleurs résultats sont en rouge et en gras..

Dataset	iLIDS-VID			
FAR(%)	1	10	50	100
MDTS-DTW [47]	12.7	**32.6**	**51.8**	57.3
DVR [9]	**17.3**	29.1	49.9	57.8
XQDA+IR	0.6	2.0	8.6	13.7
COPReV+IR	1.2	5.7	17.4	25.8
XQDA+LOMO [23]	5.6	15.4	45.8	**59.9**
COPReV+LOMO	3.9	21.0	47.9	59.1

Table 7.4 – Evaluation en base ouverte sur iLIDS VID.
Les valeurs de DIR au rang 1 sont fournies pour les valeurs suivants de FAR: 1%, 10%,50% et 100%. Les meilleurs résultats sont en rouge et en gras.

Dans le cadre de la base ouverte, avec les descripteurs IR, l'apport de COPReV comparé à XQDA est visible pour les différentes valeurs de FAR. Avec les descripteurs LOMO, l'amélioration est visible pour PRID et iLIDS pour des valeurs de FAR moyennes, égales à 10% et 50% mais aussi lorsque $FAR = 1\%$ sur PRID. Cependant, ces performances sont encore largement en dessous de celles d'autres méthodes de l'état de l'art.

7.3.3 Conclusion

COPReV s'appuie uniquement sur des contraintes de vérification pour apprendre une projection des descripteurs. Les performances sont assez mitigées ce qui suggère que des contraintes de rangement sont également nécessaires.

7.4 Représentations parcimonieuses avec une collaboration élargie

7.4.1 Présentation de la méthode

Dans le cas idéal où une méthode de vérification donne des résultats parfaits, les problèmes de ré-identification en base ouverte et en base fermée sont également résolus. Cependant si la méthode de vérification n'est pas parfaite, ce qui est le cas dans la pratique, les performances en ré-identification en base ouverte et fermée ne sont absolument plus garanties non plus. Au lieu d'imposer uniquement des contraintes sur l'appartenance ou non de deux images à une même classe (identité), nous proposons dans cette section une méthode qui prend en compte à la fois l'aspect vérification (décision), et l'aspect ré-identification (classement). Notre approche est basée sur les représentations parcimonieuses collaboratives. Nous proposons de renforcer l'aspect collaboratif de ces méthodes afin de mieux gérer le problème de la base ouverte et en particulier la possibilité de rejeter une personne requête lorsqu'elle ne fait pas partie de la galerie.

Plusieurs méthodes de ré-identification de personnes qui exploitent les représentations parcimonieuses collaboratives existent déjà [109, 102]. Pour comprendre l'intérêt de la représentation parcimonieuse collaborative, commençons par expliquer ce qu'est une représentation parcimonieuse. Etant donné un vecteur colonne x de dimension d et un dictionnaire D de dimension $d \times c$, une représentation parcimonieuse de x est un vecteur colonne $a_{x,D}$ de dimension c tel que $Da_{x,D}$ est une approximation de x et que $a_{x,D}$ est parcimonieux. Une telle représentation parcimonieuse peut être obtenue en optimisant la fonction suivante:

$$a_{x,D} = \arg\min_{a} ||x - Da||_2^2 + \lambda||a||_1 \tag{7.3}$$

où λ est un paramètre compris entre 0 et 1, $||.||_2$ est la norme euclidienne et $||.||_1$ est la norme 1. Deux termes interviennent: l'erreur de reconstruction ($||x - Da||_2^2$) et le terme de pénalisation pour obtenir un vecteur parcimonieux ($||a||_1$).

Dans le cadre de la ré-identification de personnes, le dictionnaire D est composé de la concaténation des dictionnaires de chacune des K personnes de la galerie $D = [D_1, D_2, \ldots, D_K]$ et la représentation parcimonieuse du descripteur de la personne requête est calculée à partir de ce dictionnaire collaboratif D. Plus qu'une simple collaboration, les différents éléments de D sont mis en concurrence et seuls les éléments les plus similaires au descripteur de la personne requête sont sélectionnés pour participer à la combinaison linéaire parcimonieuse de colonnes de D qui approxime ce descripteur de la personne requête. Ainsi, certaines identités galerie ne participent pas du tout à la reconstruction du descripteur de la personne requête. Ces identités galerie sont exclues du classement des identités galerie semblables à la personne requête. D'autre part, les identités galerie qui participent à la reconstruction sont rangées par ordre de similarité en considérant que plus le dictionnaire d'une personne galerie participe à la reconstruction, plus elle est similaire à la personne requête.

Dans le cas de la base ouverte, l'aspect collaboratif de la représentation qui met en compétition les identités galerie permet comme en base fermée de classer les identités galerie. Par contre, la combinaison de l'aspect collaboratif et de l'aspect parcimonieux qui permet de ne représenter l'élément requête qu'à partir du dictionnaire de peu d'identités galerie n'est pas suffisant pour rejeter complètement une personne requête comme ne faisant par partie de la galerie. En effet, la représentation parcimonieuse est parcimonieuse et non pas complètement nulle. Or si la représentation parcimonieuse correspondant au dictionnaire D était entièrement nulle, cela signifierai qu'aucune personne galerie ne doit être associée à la personne requête qui est alors rejetée. Nous proposons donc d'élargir le dictionnaire avec lequel est calculé la représentation parcimonieuse de l'élément requête de telle sorte qu'il est envisageable d'obtenir une participation nulle pour chaque personne galerie, même si la représentation parcimonieuse n'est quant à elle par totalement nulle. Les éléments non nuls correspondent au dictionnaire additionnel que nous concaténons au dictionnaire D afin de calculer la représentation parcimonieuse de l'élément requête.

Pour résumer, pour identifier une personne requête représentée par le descripteur x, nous calculons la représentation parcimonieuse de x en utilisant comme dictionnaire collaboratif le dictionnaire $[D_1, D_2, \ldots, D_K, T]$, concaténation des dictionnaires de chaque identité galerie et des éléments de l'ensemble d'apprentissage T en optimisant la fonction suivante:

$$a_{x,[D_1,D_2,\ldots,D_K,T]} = \arg\min_a ||x - [D_1, D_2, \ldots, D_K, T]a||_2^2 + \lambda||a||_1 \qquad (7.4)$$

Les identités galerie dont les dictionnaires n'interviennent pas dans la reconstruction ne sont pas classées, les autres identités galerie sont classées selon leur participation plus ou moins importante à la reconstruction.

7.4.2 Résultats expérimentaux

Nous avons testé l'approche parcimonieuse collaborative avec collaboration élargie sur deux datasets, PRID2011 [11] et iLIDS VID [9]. Afin de montrer l'apport de notre approche, nous l'avons testé avec deux descripteurs: les descripteurs LOMO [23] et les descripteurs Inception-Resnet-v2 [117] abrégé par IR. Chacun de ces deux descripteurs ont au préalable subit une projection en utilisant la méthode d'apprentissage de métrique XQDA [23] considérée comme une projection plutôt qu'une distance, ainsi qu'une étape de normalisation afin d'obtenir des descripteurs de norme unité.

Les résultats en base fermée sont présentés dans les tableaux 7.5 et 7.6.

En base fermée, pour les deux descripteurs IR et LOMO, il y a une amélioration nette (de 4% au minimum jusqu'à 29% du taux de ré-identification au premier rang) entre l'approche d'apprentissage de métrique XQDA seul et l'utilisation d'une approche parcimonieuse collaborative. Pour iLIDS VID et PRID, pour IR et LOMO, l'utilisation d'une collaboration élargie plutôt qu'une collaboration usuelle n'a pas d'effet sur les performances. Cela s'explique par le fait que l'ajout d'éléments

	Rang 1	Rang 5	Rang 10	Rang 20
MDTS-DTW [47]	49.5	75.7	84.5	91.9
DVR [9]	51.1	75.7	83.9	90.5
IR + XQDA	11.7	30.9	41.5	53.6
IR + collaboratif	40.7	66.6	77.1	85.9
IR + collaboratif élargi	40.1	65.1	76.2	85.9
LOMO + XQDA	55.3	83.1	90.3	**96.3**
LOMO + collaboratif	64.9	**87.1**	**92.5**	96.1
LOMO + collaboratif élargi	**65.1**	86.6	92.4	96.1

Table 7.5 – Evaluation en base fermée sur iLIDS VID.
Les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 20. Les meilleurs résultats sont en rouge et en gras. Best results are in bold red.

	Rang 1	Rang 5	Rang 10	Rang 20
MDTS-DTW [47]	69.6	89.4	94.3	97.9
DVR [9]	77.4	93.9	97.0	99.4
IR + XQDA	43.4	71.9	82.4	91.6
IR + collaboratif	70.7	90.0	96.1	98.3
IR + collaboratif élargi	72.0	89.9	95.3	98.1
LOMO + XQDA	86.3	**98.3**	**99.6**	**100.0**
LOMO + collaboratif	90.2	98.0	99.3	**100.0**
LOMO + collaboratif élargi	**90.6**	97.9	99.2	**100.0**

Table 7.6 – Evaluation en base fermée sur PRID.
Les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 20. Les meilleurs résultats sont en rouge et en gras.

supplémentaires ne modifie pas le classement des identités galerie, il modifie seulement la valeur de dissimilarité, mais celle dernière n'est pas prise en compte par l'évaluation CMC.

Les résultats en base ouverte sont présentés dans les tableaux 7.7 et 7.8.

En base ouverte, à nouveau, la différence est nette entre XQDA et l'approche parcimonieuse collaborative. Ce qui diffère par rapport au cas base fermée est l'apport de l'approche collaborative avec une collaboration élargie qui permet d'améliorer le taux de DIR de 5% lorsque FAR est égal à 1% par rapport à l'approche collaborative usuelle. Ainsi nous atteignons voire même surpassons les performances à l'état de l'art.

7.4.3 Conclusion

Afin de mieux rejeter les mauvais appariements dans le cadre de la base ouverte, nous avons proposé une méthode parcimonieuse collaborative avec une collaboration élargie qui permet d'améliorer les performances en base ouverte et ainsi égaler ou dépasser les autres méthodes de l'état de l'art avec pour FAR=1%, un DIR égal à 17.2% sur iLIDS VID et 55.7% sur PRID2011.

FAR(%)	1	10	50	100
MDTS-DTW [47]	12.7	32.6	51.8	57.3
DVR [9]	**17.3**	29.1	49.9	57.8
IR + XQDA[23]	0.6	2.2	8.4	14.9
IR + collaboratif	7.3	16.7	36.4	46.0
IR + collaboratif élargi	7.3	18.6	37.7	44.9
LOMO + XQDA[23]	5.1	15.2	45.3	59.1
LOMO + collaboratif	12.9	35.1	58.8	68.5
LOMO + collaboratif élargi	17.2	**37.5**	**62.8**	**69.0**

Table 7.7 – Evaluation en base ouverte sur iLIDS VID.
Les valeurs de DIR au rang 1 sont fournies pour les valeurs suivants de FAR: 1%, 10%,50% et 100%. Les meilleurs résultats sont en rouge et en gras.

FAR(%)	1	10	50	100
MDTS-DTW [47]	42.7	55.2	70.5	72.8
DVR [9]	46.8	58.3	78.3	79.7
IR + XQDA[23]	3.2	9.5	25.7	50.8
IR + collaboratif	21.0	44.8	66.2	75.2
IR + collaboratif élargi	28.0	46.8	69.5	76.3
LOMO + XQDA[23]	21.2	40.7	78.8	90.5
LOMO + collaboratif	49.8	69.3	88.2	**93.8**
LOMO + collaboratif élargi	**55.7**	**71.0**	**90.2**	93.2

Table 7.8 – Evaluation en base ouverte sur PRID.
Les valeurs de DIR au rang 1 sont fournies pour les valeurs suivants de FAR: 1%, 10%,50% et 100%. Les meilleurs résultats sont en rouge et en gras.

7.5 Représentation collaborative bidirectionnelle

7.5.1 Présentation de la méthode

L'approche parcimonieuse collaborative avec collaboration élargie a permis dans la section précédente d'obtenir des résultats dépassant déjà l'état de l'art. Dans cette section, nous renforçons cette approche en insistant sur l'importance d'une relation réciproque afin d'obtenir une correspondance encore plus robuste. Au lieu de ne s'intéresser qu'au classement des identités galerie similaires à la personne requête, nous évaluons également le problème du point de vue de chacune des personnes de galerie. Nous nommerons méthode inverse cette approche où les rôles de la requête et de la galerie sont inversés. De même qu'un score de dissimilarité peut être calculé à partir de la représentation parcimonieuse d'un élément requête en utilisant le dictionnaire de la galerie, la représentation parcimonieuse d'un élément de la galerie en utilisant des éléments requête fournit également un score de dissimilarité. La combinaison de ces deux scores par une simple somme forme une nouvelle méthode bidirectionnelle.

Il est important de noter que le problème n'est pas tout à fait symétrique. En effet, ce sont toujours les identités galerie que l'on souhaite classer en fonction de

leur similarité avec la personne requête. De plus, si les personnes galeries sont toujours toutes disponibles puisqu'elles sont connues, les images des identités requêtes peuvent être fournies une par une (on parlera de méthode inverse 1 personne), ou toutes simultanément (on parlera de méthode inverse toutes personnes).

7.5.2 Résultats expérimentaux

Nous avons testé l'approche parcimonieuse collaborative avec collaboration élargie sur deux datasets, PRID2011 [11] et iLIDS VID [9]. Nous avons distingué le cas où les personnes requête sont présentées une par une (cas "inverse 1 pers.") et le cas où toutes les personnes sont présentées simultanément (cas "inverse toutes pers"). Afin de montrer l'apport de notre approche, nous l'avons testé avec deux descripteurs: les descripteurs LOMO [23] et les descripteurs Inception-Resnet-v2 [117] abrégé par IR. Chacun de ces deux descripteurs ont au préalable subit une projection en utilisant la méthode d'apprentissage de métrique XQDA [23] considérée comme une projection plutôt qu'une distance, ainsi qu'une étape de normalisation afin d'obtenir des descripteurs de norme unité.

Les résultats en base fermée sont présentés dans les tableaux 7.9 et 7.10.

Rang	1	5	10	20
MDTS-DTW [47]	49.5	75.7	84.5	91.9
DVR [9]	51.1	75.7	83.9	90.5
XQDA[23]	55.3	83.1	90.3	96.3
collaboratif	64.9	87.1	92.5	96.1
collaboratif élargi	65.1	86.6	92.4	96.1
inverse 1 pers.	65.4	88.3	93.9	96.8
inverse toutes pers.	**69.9**	**89.8**	**94.2**	**96.9**
collaboratif élargi+ inverse 1 pers.	68.1	88.9	93.7	96.7
collaboratif élargi+ inverse toutes pers.	69.8	89.6	93.5	96.8

Table 7.9 – Evaluation en base fermée sur iLIDS VID.
Les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 20. Les meilleurs résultats sont en rouge et en gras.

En base fermée, les performances de l'approche inverse lorsqu'une seule personne requête est présentée à la fois sont similaires à l'approche parcimonieuse collaborative avec collaboration élargie présentée dans la section précédente et sont donc nettement supérieures à XQDA seul. Lorsque toutes les personnes requêtes sont présentées de manière simultanée, nous observons à nouveau un gain de 4% en terme de taux ré-identification au premier rang. Quant à la combinaison des approches directes et inverses, cela améliore les performances lorsqu'une seule requête est présentée à la fois, mais cela n'apporte pas de gain en performances lorsque toutes les personnes requêtes sont présentées simultanément.

Les résultats en base ouverte sont présentés dans les tableaux 7.11 et 7.12.

En base ouverte, les performances de l'approche inverse lorsqu'une seule personne requête est présentée à la fois sont similaires voire même meilleurs que l'approche

Rang	1	5	10	20
MDTS-DTW [47]	69.6	89.4	94.3	97.9
DVR [9]	77.4	93.9	97.0	99.4
XQDA[23]	86.3	98.3	99.6	**100.0**
collaboratif	90.2	98.0	99.3	**100.0**
collaboratif élargi	90.6	97.9	99.2	**100.0**
inverse 1 pers.	89.8	98.2	99.2	**100.0**
inverse toutes pers.	**94.2**	98.5	99.2	**100.0**
collaboratif élargi+ inverse 1 pers.	91.2	98.4	99.4	**100.0**
collaboratif élargi+ inverse toutes pers.	93.8	**98.3**	**99.6**	**100.0**

Table 7.10 – Evaluation en base fermée sur PRID.
Les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 20. Les meilleurs résultats sont en rouge et en gras.

FAR	1	10	50	100
MDTS-DTW [47]	12.7	32.6	51.8	57.3
DVR [9]	17.3	29.1	49.9	57.8
XQDA[23]	5.1	15.2	45.3	59.1
collaboratif	12.9	35.1	58.8	68.5
collaboratif élargi	17.2	37.5	62.8	69.0
inverse 1 pers.	13.9	35.6	61.5	69.2
inverse toutes pers.	22.0	44.7	67.4	72.7
collaboratif élargi+inverse 1 pers.	18.0	40.5	66.5	71.6
collaboratif élargi+inverse toutes pers.	**23.8**	**45.8**	**68.1**	**72.9**

Table 7.11 – Evaluation en base ouverte sur iLIDS.
Les valeurs de DIR au rang 1 sont fournies pour les valeurs suivants de FAR: 1%, 10%,50% et 100%. Les meilleurs résultats sont en rouge et en gras.

parcimonieuse collaborative avec collaboration élargie présentée dans la section précédente et sont donc nettement supérieures à XQDA seul. Lorsque toutes les personnes requêtes sont présentées de manière simultanée, nous observons à nouveau un gain de 9% jusqu'à 15% en terme de DIR lorsque FAR est égal à 1%. Quant à la combinaison des approches directes et inverses, cela améliore surtout les performances lorsqu'une seule requête est présentée à la fois, mais cela n'apporte pas systématiquement un gain en performances lorsque toutes les personnes requêtes sont présentées simultanément.

7.5.3 Conclusion

Dans cette section nous avons présenté une approche qui se veut symétrique à l'approche présentée dans la section précédente où les rôles de la galerie et de la requête sont inversés. Cette approche inverse présente des résultats meilleurs que l'approche directe que ce soit en base fermée ou ouverte. De plus la combinaison des deux approches permet d'améliorer les performances et d'obtenir des résultats meilleurs que l'état de l'art lorsqu'une seule personne requête est présentée à la

FAR	1	10	50	100
MDTS-DTW [47]	42.7	55.2	70.5	72.8
DVR [9]	46.8	58.3	78.3	79.7
XQDA[23]	21.2	40.7	78.8	90.5
collaboratif	49.8	69.3	88.2	93.8
collaboratif élargi	55.7	71.0	90.2	93.2
inverse 1 pers.	58.0	69.7	91.0	93.0
inverse toutes pers.	73.2	**87.3**	**95.3**	**96.0**
collaboratif élargi+inverse 1 pers.	60.8	76.0	91.7	93.3
collaboratif élargi+inverse toutes pers.	**73.3**	83.2	94.0	95.2

Table 7.12 – Evaluation en base ouverte sur PRID.
Les valeurs de DIR au rang 1 sont fournies pour les valeurs suivants de FAR: 1%, 10%,50% et 100%. Les meilleurs résultats sont en rouge et en gras.

fois. Lorsque les personnes requêtes sont présentées simultanément, les résultats sont meilleurs que lorsqu'une seule personne requête est présentée à la fois.

7.6 Conclusion et perspectives

7.6.1 Conclusion

D'abord considéré comme un sous problème du problème de suivi multi-caméras, le problème de la ré-identification de personnes est devenue une tâche à part entière de vision par ordinateur en 2007. Depuis lors, cette problématique n'a eu de cesse d'évoluer en prenant en compte des données de plus en plus riches et en considérant des scénarios de plus en plus réalistes. Ainsi dans cette thèse, nous avons tenu à proposer des méthodes adaptées à l'exploitation de données multi-shot. Néanmoins, la recherche a été surtout axée sur la problématique de la ré-identification en base ouverte où contrairement à la base fermée, la galerie ne contient pas toutes les identités possibles et la personne requête doit pouvoir être rejetée si elle ne correspond à aucune personne de la galerie.

Sachant que lorsque le problème de la vérification de personnes est parfaitement résolu (ie. lorsqu'on est capable de déterminer si deux ensemble d'images correspondent à une seule personne ou à deux personnes distinctes), les problèmes de ré-identification en base fermée (classement) et en base ouverte (détection et classement) sont également résolus, nous avons proposé une méthode appelée COPReV (Closed and Open World Person Re-identification and Verification) qui utilise exclusivement des contraintes de vérification pour apprendre une projection des descripteurs telle qu'après projection, les distances des descripteurs de paires positives sont inférieures à un certain seuil et celles des paires négatives sont supérieures à ce même seuil. Si cette méthode permet globalement une amélioration des performances pour les descripteurs testés dans le cadre de bases d'identités fermée et ouverte, l'amélioration est variable selon le descripteur utilisé et les résultats en base ouverte restent largement inférieurs à ceux de l'état de l'art.

Ainsi nous avons proposé un deuxième type d'approches basées sur les représentations parcimonieuses collaboratives. Les représentations parcimonieuses collaboratives permettent d'une part de mettre en concurrence les identités de la galerie quant à leur similarité avec la personne requête (aspect collaboratif) et d'autre part d'exclure du classement de nombreuses identités galerie, celles qui sont les moins semblables à la personne requête. Nous avons proposé une approche avec une collaboration élargie afin de mieux traiter la problématique de la base ouverte tout en conservant sur la problématique de la base fermée, des performances équivalentes à une approche parcimonieuse collaborative standard. Grâce à la collaboration élargie, le nombre de mauvais appariements de personnes requêtes n'existant pas dans la galerie avec des personnes de la galerie est réduit. Afin de renforcer l'approche collaborative élargie proposée, nous avons développé une seconde approche collaborative dans laquelle les rôles des personnes de la galerie et des personnes requêtes sont inversées. Les deux approches ne sont néanmoins pas complètement symétriques car au final ce sont toujours les personnes galeries que l'on doit ranger pas ordre de similarité avec la personne requête. En combinant les deux approches collaboratives élargies, nous obtenons une approche collaborative bidirectionnelle qui montre des résultats meilleurs que ceux de l'état de l'art.

7.6.2 Perspectives

Tout au long de la thèse, nous nous sommes essentiellement intéressés a l'appariement des personnes requête et galerie. Nous avons utilises deux descripteurs afin de montrer la pertinence de la méthode d'appariement. Si les méthodes proposées améliorent les performances pour les deux descripteurs testés, les résultats diffèrent grandement selon les descripteurs choisis. Des travaux futurs pourraient être conduits afin de développer des descripteurs directement adaptés a un appariement avec les approches parcimonieuses proposées. L'apprentissage de ces nouveaux descripteurs pourraient notamment se faire avec une approche de type réseaux de neurones.

Certaines étapes des méthodes proposées peuvent également être améliorées. En particulier d'autres manières de former les dictionnaires collaboratifs pourraient être proposées, avec par exemple un apprentissage de dictionnaire plutôt qu'une sélection d'éléments de l'ensemble d'apprentissage.

Des considérations sur les temps de calcul et les besoins mémoire devrons également être pris en compte afin de permettre l'utilisation de nos méthodes dans le cadre de bases d'images avec toujours plus d'identités et plus d'images par identités.

Enfin, des scénarios encore plus complexes devrons faire l'objet de futurs travaux car si le problème de ré-identification en base ouverte étudié dans cette thèse se rapproche davantage des problématiques rencontrées en pratique comparé au problème de la base fermée, les nouvelles identités requêtes non présentes dans la galerie ne sont pas pour le moment ajoutée dans la galerie pour de futures ré-identifications. L'ensemble des identités de la galerie devrait être un ensemble dynamique qui

s'agrandit au fur et a mesure de la détection de nouvelles personnes auparavant inconnues.

Bibliography

[1] Brais Cancela, Tim Hospedales, and Shaogang Gong. Open-world person re-identification by multi-label assignment inference. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[2] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2650–2657. IEEE, 2012.

[3] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):591–606, 2016.

[4] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2650–2657. IEEE, 2012.

[5] Yasutomo Kawanishi, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, volume 5, page 6, 2014.

[6] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, 2007.

[7] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *International Conference on Computer Vision (ICCV'07)*, October 2007.

[8] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1988–1995. IEEE, 2009.

[9] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014.

[10] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.

[11] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[13] Shengcai Liao, Zhipeng Mo, Jianqing Zhu, Yang Hu, and Stan Z Li. Open-set person re-identification. *arXiv preprint arXiv:1408.0872*, 2014.

[14] Solène Chan-Lang, Quoc Cuong Pham, and Catherine Achard. Bidirectional sparse representations for multi-shot person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, pages 263–270. IEEE, 2016.

[15] Solène Chan-Lang, Quoc Cuong Pham, and Catherine Achard. Closed and open-world person re-identification and verification. In *Digital Image Computing Techniques and Applications (DICTA), 2017 International Conference on*. IEEE, 2017.

[16] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. *The IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[17] Michele Stoppa Loris Bazzani Dong Seon Cheng, Marco Cristani and Vittorio Murino. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference*, pages 68.1–68.11. BMVA Press, 2011. http://dx.doi.org/10.5244/C.25.68.

[18] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Computer Vision–ECCV 2008*, pages 262–275. Springer, 2008.

[19] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: what features are important? In *European Conference on Computer Vision Workshops and Demonstrations*, 2012.

[20] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2666–2672. IEEE, 2012.

[21] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.

[22] Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, and John R. Smith. Learning locally-adaptive decision functions for person verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference*, pages 21.1–21.11. BMVA Press, 2010. doi:10.5244/C.24.21.

[25] Andy J. Ma, Pong C. Yuen, and Jiawei Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[26] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, and Jun Chen. Ranking optimization for person re-identification via similarity and dissimilarity. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1239–1242. ACM, 2015.

[27] Mang Ye, Jun Chen, Qingming Leng, Chao Liang, Zheng Wang, and Kaimin Sun. Coupled-view based ranking optimization for person re-identification. In *International Conference on Multimedia Modeling*, pages 105–117. Springer, 2015.

[28] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016.

[29] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. 2017.

[30] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[31] Jorge García, Niki Martinel, Alfredo Gardel, Ignacio Bravo, Gian Luca Foresti, and Christian Micheloni. Discriminant context information analysis for post-ranking person re-identification. *IEEE Transactions on Image Processing*, 26(4):1650–1665, 2017.

[32] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *The IEEE International Conference on Computer Vision*, 2015.

[33] Andy Jinhua Ma and Ping Li. Query based adaptive re-ranking for person re-identification. In *Asian Conference on Computer Vision*, pages 397–412. Springer, 2014.

[34] Vu-Hoang Nguyen, Thanh Due Ngo, Khang MTT Nguyen, Due Anh Duong, Kien Nguyen, and Duy-Dinh Le. Re-ranking for person re-identification. In

Soft Computing and Pattern Recognition (SoCPaR), 2013 International Conference of, pages 304–308. IEEE, 2013.

[35] Wei Li, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. Common-near-neighbor analysis for person re-identification. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1621–1624. IEEE, 2012.

[36] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[37] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *Computer Vision and Pattern Recognition*, 2015.

[38] Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang. Transferring a semantic representation for person re-identification and search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[39] Dong Yi, Zhen Lei, and Stan Z Li. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*, 2014.

[40] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE, 2014.

[41] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *arXiv preprint arXiv:1512.05300*, 2015.

[42] Ejaz Ahmed, Michael Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[43] 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.

[44] 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.

[45] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[46] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong. Towards unsupervised open-set person re-identification. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 769–773. IEEE, 2016.

[47] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[48] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Investigating open-world person re-identification using a drone. In *Computer Vision-ECCV 2014 Workshops*, pages 225–240. Springer, 2014.

[49] Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J. Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[50] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, volume 2, 2009.

[51] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011.

[52] Alina Bialkowski, Simon Denman, Sridha Sridharan, Clinton Fookes, and Patrick Lucey. A database for person re-identification in multi-camera surveillance networks. In *Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on*, pages 1–8. IEEE, 2012.

[53] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[54] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[55] Athira Nambiar, Matteo Taiana, Dario Figueira, Jacinto C Nascimento, and Alexandre Bernardino. A multi-camera video dataset for research on high-definition surveillance. *International Journal of Machine Intelligence and Sensory Signal Processing*, 1(3):267–286, 2014.

[56] Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino. The hda+ data set for research on fully automated re-identification systems. In *European Conference on Computer Vision*, pages 241–255. Springer, 2014.

[57] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *European Conference on Computer Vision*, pages 330–345. Springer, 2014.

[58] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2016.

[59] Chongyang Zhang, Bingbing Ni, Li Song, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Best: benchmark and evaluation of surveillance task. In *Asian Conference on Computer Vision*, pages 393–407. Springer, 2016.

[60] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, and Qi Tian. Person re-identification in the wild. *arXiv preprint arXiv:1604.02531*, 2016.

[61] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[62] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[63] Slawomir Bak, Etienne Corvee, Francois Brémond, and Monique Thonnat. 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.

[64] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *The IEEE International Conference on Computer Vision*, 2013.

[65] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[66] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *Computer Vision and Pattern Recognition , 2014 IEEE Conference on*, 2014.

[67] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[68] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1629–1642, 2015.

[69] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing*, 6(7):965–976, 1997.

[70] Cheng-Hao Kuo, Sameh Khamis, and Vinay Shet. Person re-identification using semantic color names and rankboost. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 281–287. IEEE, 2013.

[71] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *European Conference on Computer Vision*, pages 536–551. Springer, 2014.

[72] Ryan Layne, Tim Hospedales, and Shaogang Gong. Person re-identification by attributes. In *Proceedings of the British Machine Vision Conference*, pages 24.1–24.11. BMVA Press, 2012.

[73] Ryan Layne, Tim Hospedales, and Shaogang Gong. Re-id: Hunting attributes in the wild. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[74] Xiao Liu, Mingli Song, Qi Zhao, Dacheng Tao, Chun Chen, and Jiajun Bu. Attribute-restricted latent topic model for person re-identification. *Pattern recognition*, 45(12):4204–4213, 2012.

[75] Mert Dikmen, Emre Akbas, Thomas Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. *Computer Vision–ACCV 2010*, pages 501–512, 2011.

[76] Martin Hirzer, Peter M Roth, and Horst Bischof. Person re-identification by efficient impostor-based metric learning. In *The IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2012.

[77] 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.

[78] Cijo Jose and François Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *European Conference on Computer Vision*, pages 875–890. Springer, 2016.

[79] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, 2012.

[80] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[81] Xiaokai Liu, Hongyu Wang, Jie Wang, and Xiaorui Ma. Person re-identification by multiple instance metric learning with impostor rejection. *Pattern Recognition*, 67:287–298, 2017.

[82] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014.

[83] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[84] Shengcai Liao and Stan Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[85] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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

[87] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 498–505. IEEE, 2009.

[88] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[89] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014.

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

[91] 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.

[92] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[93] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[94] 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.

[95] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[96] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[97] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. *arXiv preprint arXiv:1607.05369*, 2016.

[98] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016.

[99] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[100] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

[101] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[102] Srikrishna Karanam, Yang Li, and Richard Radke. Sparse re-id: Block sparsity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2015.

[103] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with block sparse recovery. *Image and Vision Computing*, 60:75–90, 2017.

[104] Mohamed Ibn Khedher, Mounim A El Yacoubi, and Bernadette Dorizzi. Multi-shot surf-based person re-identification via sparse representation. In *The IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2013.

[105] Srikrishna Karanam, Yang Li, and Richard J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[106] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 44.1–44.12. BMVA Press, September 2015.

[107] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised\ ell _1 graph learning. In *European Conference on Computer Vision*, pages 178–195. Springer, 2016.

[108] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[109] Le An, Xiaojing Chen, Songfan Yang, and Bir Bhanu. Sparse representation matching for person re-identification. *Information Sciences*, pages –, 2016.

[110] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[111] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[112] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan. Sample-specific svm learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[113] Xiang Li, Ancong Wu, Mei Cao, Jinjie You, and Wei-Shi Zheng. Towards more reliable matching for person re-identification. In *Identity, Security and Behavior Analysis (ISBA), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[114] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[115] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *European Conference on Computer Vision*, pages 405–422. Springer, 2016.

[116] Svebor Karaman and Andrew D Bagdanov. Identity inference: generalizing person re-identification scenarios. In *Computer Vision–ECCV 2012. Workshops and demonstrations*, pages 443–452. Springer, 2012.

[117] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[118] 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.