Person re-identification in images with deep learning

Yiqiang Chen

N°d'ordre NNT : 2018LYSEI074

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'INSA LYON

Ecole Doctorale N° 512
INFORMATIQUE ET MATHEMATIQUES

Spécialité de doctorat : Informatique

Soutenue publiquement le 12/10/2018, par :
Yiqiang CHEN

Person Re-identification in Images with Deep Learning

Devant le jury composé de :

BENOIS-PINEAU, Jenny	Prof.	Université de Bordeaux	Rapporteure
BREMOND, François	DR	INRIA	Rapporteur
THOME, Nicolas	Prof.	CNAM	Examinateur
ACHARD, Catherine	MCF-HDR	Sorbonne Université	Examinatrice
DUFOUR, Jean-Yves	Dr. Ing.	Thales ThereSIS Lab	Examinateur
BASKURT, Atilla	Prof.	INSA-LYON	Directeur de thèse
DUFFNER, Stefan	MCF	INSA-LYON	Co-directeur de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr** INSA : R. GOURDON	M. Stéphane DANIELE Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec. : M.C. HAVGOUDOUKIAN **ecole-doctorale.eea@ec-lyon.fr**	M. Gérard SCORLETTI École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 **gerard.scorletti@ec-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES **secretariat.e2m2@univ-lyon1.fr**	M. Philippe NORMAND UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://www.ediss-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE **secretariat.ediss@univ-lyon1.fr**	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 Avenue Jean CAPELLE INSA de Lyon 69 621 Villeurbanne Tél : 04.72.68.49.09 Fax : 04.72.68.49.16 **emmanuelle.canet@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 Fax : 04.72.43.16.87 **infomaths@univ-lyon1.fr**	M. Luca ZAMBONI Bât. Braconnier 43 Boulevard du 11 novembre 1918 69 622 Villeurbanne CEDEX Tél : 04.26.23.45.52 **zamboni@maths.univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Marion COMBE Tél : 04.72.43.71.70 Fax : 04.72.43.87.12 Bât. Direction **ed.materiaux@insa-lyon.fr**	M. Jean-Yves BUFFIÈRE INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 **jean-yves.buffiere@insa-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Marion COMBE Tél : 04.72.43.71.70 Fax : 04.72.43.87.12 Bât. Direction **mega@insa-lyon.fr**	M. Jocelyn BONJOUR INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **http://ed483.univ-lyon2.fr** Sec. : Viviane POLSINELLI Brigitte DUBOIS INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 **viviane.polsinelli@univ-lyon2.fr**	M. Christian MONTES Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Person Re-identification in Images with Deep Learning

Yiqiang CHEN

Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS)
INSA-Lyon

This dissertation is submitted for the degree of
Doctor of Philosophy

INSA-Lyon October 2018

献给我亲爱的爷爷奶奶

Je voudrais dédier cette thèse à mes chers grand-parents.

不管风吹浪打，我自闲庭信步。

—毛泽东《水调歌头-游泳》

Malgré le vent et la vague, je marche avec la sérénité.

- Mao Zedong, “Shui Diao Ge Tao - Natation”

Remerciements

Tout d'abord, je souhaite remercier à toutes les personnes qui m'ont aidé et soutenu tout au long de ma thèse. Avec tous mes respects, je remercie grandement mes encadrants Atilla Baskurt et Stefan Duffner qui m'ont dirigé ces trois années de thèse. Merci à Atilla de m'avoir donné l'opportunité de travailler avec vous et m'avoir fait confiance, d'être toujours présent, patient et attentif durant ma thèse, et de m'avoir encouragé et soutenu lors que j'avais des difficultés. Merci à Stefan d'être tout le temps disponible pour répondre à mes questions, de m'avoir donné des conseilles et de m'avoir montré la rigueur scientifique, ainsi pour les relectures attentive et corrections précises sur les article et la thèse. Je suis vraiment ravi d'avoir travaillé avec vous. J'ai beaucoup appris à vos cotés.

Je remercie également à Jean-Yves Dufour et Andrei Stoian d'avoir suivi tout au long de ma thèse et me donner des conseils et remercie à Thales pour le soutien financier. Mes remerciements se dirigent ensuite vers tous les membres du Jury, en particulier à Monsieur François Bremond et Madame Jenny Benois-pineau pour le temps qu'ils ont pu accorder à relire mes travaux et leurs remarques très intéressantes et constructives.

Je remercie à nouveau Atilla et ainsi Christophe Garcia, Khalid Idrissi, Cristian Wolf. Le cours de traitement d'image au département TC que vous avez organisé m'a ouvert la porte de la vision d'ordinateur et l'apprentissage automatique. Vos présentations intéressantes dans ce cours, m'a fait découvrir ce domaine fantastique et m'a fait décider à poursuivre mes études et même mon carrière dans cette direction.

J'adresse toute ma gratitude à tous mes amis et l'ensemble des membres de l'équipe IMAGINE pour toute la bienveillance durant ma thèse, et également pour les nombreux bons moments passés ensembles. Entre autre Yuqi, Yongzhe, Yongli, Jinzhe, Abderrahmane, Quentin, Fabien, Riyadh ...

Enfin, les mots les plus simples étant les plus forts, j'adresse toute mon affection à ma famille, et en particulier à ma adorable femme Jiaxin qui m'a supporté l'idée de poursuivre mes études en thèse. Sa compagnie et son encouragement sont les piliers fondateurs de ce que j'ai achevé. Sans elle, je n'aurais pas pu faire cette thèse. Je remercie à mes grands-parents et à mes parents de me supporter de faire les études en France et pour leur confiance et tendresse depuis toujours.

Abstract

Video surveillance systems are of great value for public safety. A major difficulty in such systems concerns person re-identification which is defined as the problem of identifying people across images that have been captured by different surveillance cameras without overlapping fields of view. With the increasing need for automated video analysis, this task is receiving increasing attention. And it underpins many critical applications such as cross-camera tracking, multi-camera behavior analysis and forensic search. However, this problem is challenging due to the large variations of lighting, pose, viewpoint and background.

To tackle these different difficulties, in this thesis, we propose several deep learning based approaches to obtain a better person re-identification performance in different ways. In the first proposed approach, we use pedestrian attributes to enhance the person re-identification. The attributes are defined as semantic mid-level descriptions of persons, such as gender, accessories, clothing etc. They are helpful to extract characteristics that are invariant to the pose and viewpoint variations. In order to make use of the attributes, we propose a Convolutional Neural Network (CNN)-based person re-identification framework composed of an identity classification branch and of an attribute recognition branch. At a later stage, these two cues are combined to perform person re-identification.

Secondly, among the challenges, one of the most difficult is the variation under different viewpoints. To deal with this issue, we consider that the images under various orientations are from different domains. We propose an orientation-specific CNN. This framework performs body orientation regression in a gating branch, and in another branch learns separate orientation-specific layers as local experts. The combined orientation-specific CNN feature representations are used for the person re-identification task.

Learning a similarity metric for person images is a crucial aspect of person re-identification. As the third contribution, we propose a novel listwise loss function taking into account the order in the ranking of gallery images with respect to different probe images. Further, an evaluation gain-based weighting is introduced in the loss function to optimize directly the evaluation measures of person re-identification.

Finally, as the last contribution, we proposed a deep learning based method using group context to improve person re-identification task. This method aims to reduce the ambiguity caused by the similar clothing among a large gallery set. Location-invariance of members in the group is achieved by global max-pooling to better associate groups. And we propose to combine the single person appearance and the group context to enhance the re-identification performance. This method can be extended to any CNN-based person re-identification method to exploit group context.

For all the four contributions of this thesis, we carry out extensive experiments on popular benchmarks and datasets to demonstrate the effectiveness of the proposed systems.

Résumé

La vidéosurveillance est d'une grande valeur pour la sécurité publique. La ré-identification de personnes, étant un des aspects les plus importants des systèmes de vidéosurveillance, est définie comme le problème de l'identification d'individus dans des images captées par différentes caméras de surveillance à champs non recouvrants. Avec l'augmentation du besoin de l'analyse automatique de vidéo, la ré-identification reçoit de plus en plus d'attention. De plus, il sert pour de nombreuses d'applications importantes telles que le suivie de personnes à travers plusieurs caméras, l'analyse de comportement et la recherche contextuelle. Cependant, le tâche est difficile à cause d'une série de défis sur l'apparence de la personne, tels que des variations de poses, du point de vue et de l'éclairage etc.

Pour régler ces différents problèmes, dans cette thèse, nous proposons plusieurs approches basées sur l'apprentissage profond de sorte d'améliorer la performance sur la ré-identification de différentes manières. Dans la première approche, nous utilisons les attributs des piétons pour cette amélioration. Les attributs sont définis comme la description sémantique de niveau intermédiaire sur les personnes telles que genre, accessoires, vêtements etc. Ils sont des descripteurs robustes aux différentes variations. Pour détecter et exploiter les attributs, nous proposons un système basé sur un CNN (Convolutional Neural Network) qui est composé d'une branche de classification d'identité et d'une autre branche de reconnaissance d'attributs. Les sorties de ces deux branches sont fusionnées ensuite pour la ré-identification de personnes.

Deuxièmement, le plus grand défi dans ce problème est la variation de point du vue. Nous considérons que les images de piétons avec les différentes orientations sont de différents domaines, et nous proposons une architecture spécifique à ces domaines. Le système fait une régression de l'orientation dans une branche et dans une autre branche apprend les couches qui sont spécifiques à différentes orientations. Les caractéristiques de différentes orientations sont ensuite combinées pour avoir une représentation robuste au changement de point de vue.

L'apprentissage de similarité pour les images de personnes est l'une de plus importantes étapes de la ré-identification de personnes. Comme troisième contribution de cette thèse, nous proposons une nouvelle fonction de coût basée sur des listes d'exemples. Cette fonction de coût correspond au niveau du désordre de classement d'une liste d'images par rapport à différentes images requêtes. Une pondération basée sur l'amélioration de l'évaluation du classement est introduite pour optimiser directement les mesures d'évaluation.

Enfin, une dernière contribution consiste à utiliser le contexte dans l'image pour ré-identifier des personnes. Pour ce faire, nous utilisons non seulement l'image de la personne à identifier mais aussi le groupe de personnes qui l'entoure. Nous proposons d'extraire une représentation de caractéristiques visuelles invariante à la position d'un individu dans une image de group. Cette prise en compte de contexte de groupe réduit l'ambigüité de ré-identification et améliore ainsi la performance.

Pour chacune de ces contributions de cette thèse, nous effectuons de nombreuses expériences sur différentes bases de données pour montrer l'efficacité des approches proposées et pour les comparer à l'état de l'art.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Video Surveillance

1.1.1 Context

Video surveillance systems monitor the behavior, activities, or other changing information of people by means of electronic equipment. Today a huge amount of video surveillance or closed-circuit television (CCTV) cameras are installed throughout the world. Based on some historical camera shipments and the predicted life spans of these devices, the count is estimated to be 245 million video surveillance cameras globally in 2014.

These cameras occur in various domains ranging from rather small home surveillance applications in private areas, to medium-sized and large installations for monitoring public areas, e.g. shopping centers, airports, train stations, public transportation, sports centers and so on.

The system was very quickly popularized in the world. In France, only 60 cities (communes) installed at least one camera in 1999. In 2014, the number was 2384 and the demand is on the rise. Almost 130 million surveillance cameras have been are shipped in the world in 2017.

1.1.2 Importance of video surveillance

The reason for this enormous increase is probably twofold: firstly, cameras have become relatively cheap and secondly they are an effective tool for protecting people and property day and night. The stated goal of the installation of CCTV cameras is to reduce crime and increase public safety. The CCTV cameras get the images and send them to a video

surveillance center where all scenes are viewed (real-time mode) or stored (posteriori mode) as shown in Fig. 1.1.

It is an efficient way to prevent crimes. Surveillance cameras highly increase the cost of the commission of a crime and have been found to reduce the overall crime rate by approximately 25 percent at subway stations after the installation, due to their deterrent effect on potential offenders.

It can be used to prevent crimes in case of a real-time mode. These installations are helpful in keeping an eye on the people entering a place and their activities in case any kind of threat is sensed at that place. The police or the security agents could allow rapid interventions to preempt incidents through real time detections of suspicious behaviors.

It is also effective for finding evidence after an accident or an attack in case of a posteriori mode. The recorded footage can be used to identify suspects, witnesses, or vehicles involved in the crime, as well as to establish the timeline of an incident or a crime. In addition, it can also be used for the purpose of traffic monitoring, industrial processes etc.

1.1.3 Automatic video analysis

Most existing video surveillance systems only provide the infrastructure to capture, store and distribute video, while leaving the task of threat detection exclusively to human operators. Human monitoring of surveillance video is a very labor-intensive task. It is generally agreed that watching video feeds requires a bigger level of visual attention than most every day tasks. Specifically, vigilance, the ability to focus and to react to rarely occurring events is extremely demanding and prone to error due to lapses in attention.

Moreover, millions hours of video data generated by a large number of cameras require more and more operators for the task. It's almost infeasible to achieve real-time prevention due to the high cost, thereby severely reducing the effectiveness of surveillance. With the proliferation of digital cameras and the advent of powerful computing ressources, automatic video analysis has become possible and more and more common in video surveillance applications, thus reducing considerably this cost. For the real-time mode, the goal of automatic video analysis for security and surveillance is to automatically detect events and situations that require the attention of security personnel. Automated analysis of large amounts of video data can not only process the data faster but significantly improve the ability to preempt incidents. Augmenting security staff with automatic processing thus increases their efficiency and effectiveness.

For the posteriori mode, searching for a given person of interest in thousands hours of recorded videos provided by many cameras, requires to assign a large number of enforcement officers to this task, and requires a lot of time to be performed. Automated content-based

video retrieval reproducing and assisting human analysis on the recorded videos largely enhances forensic capabilities.

For these reasons, analyzing and understanding video content is becoming a critical field of research. This research domain covers many tasks like object detection and recognition, object tracking, gesture recognition, behavior analysis and understanding, etc. All these tasks are used in many domains like robotics, entertainment, but also, to a large extent, in security and video surveillance. They are difficult problems due to the large variability of the underlying signals and acquisition conditions.

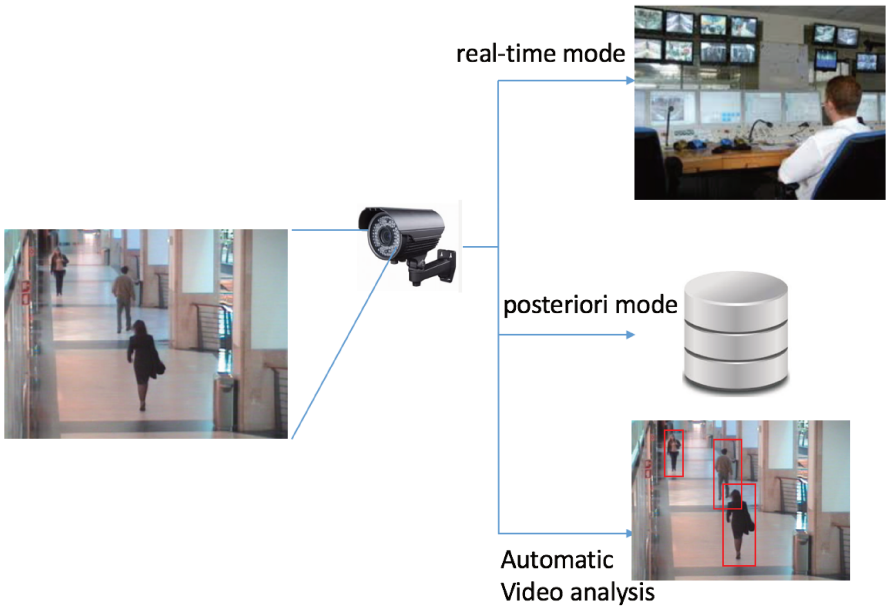


Fig. 1.1 Different operating mode of video surveillance systems

1.2 Person Re-identification

1.2.1 Definition

Person re-identification consists in matching images of a particular person captured in a network of cameras with non-overlapping fields of view. The task is different from the classic identification and detection tasks. The identification consists in determining the identity of a person in an image and the detection consists in discriminating people from the background without knowing the identity. Re-identification answers the question whether a given image belongs to the same person as a query image. The identification task helps us to know who it is and the detection task indicates whether it is a person. But the re-identification tells when

and where this person appeared with respect to a given camera and, using several cameras, potentially allows for the estimation of his/her trajectory over a short period of time.

The general pipeline of person re-identification is shown in Fig. 1.2. The person re-identification is based on the pedestrian detection task. In the first step, we form a pedestrian gallery set by collecting the cropped pedestrian images or extracted pedestrian image signature from each camera scene in the network. Then, we measure the similarity or distance to the query image. Finally, we show the best matched images according to the measured similarity.

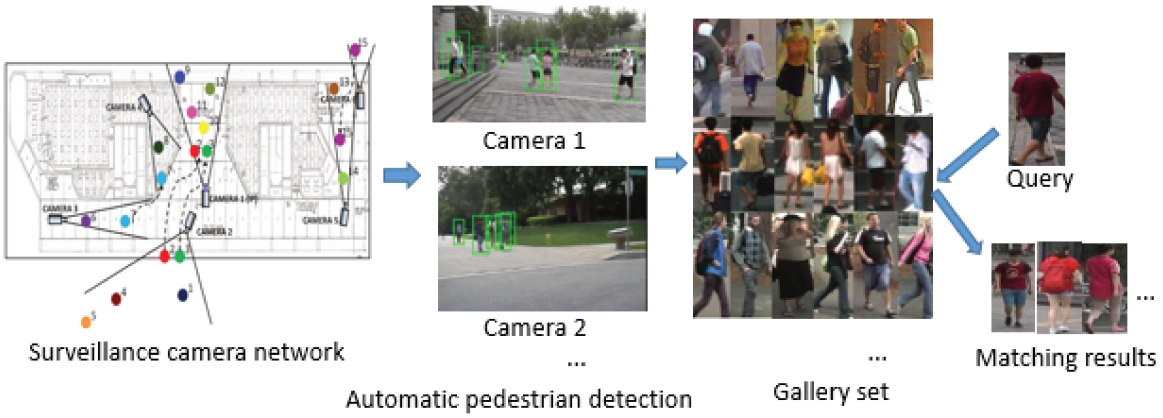


Fig. 1.2 Person re-identification system pipeline.

An assumption that is generally made is that individuals keep the same clothing in different scenes. This works best in a scenario with a relatively short time period, guaranteeing the constraint of a similar visual appearance. In reality, re-identification cannot be applied to find similarities among people after several days due to likely alterations in their visual appearance. The reason for this hypothesis is that more accurate biometrics like faces are not always available in videos in "far-sight" surveillance settings, which are the most common in practice. In that case, the re-identification algorithms rely mainly on the overall appearance of the individual.

The surveillance cameras generally span large geospatial areas and have non-overlapping fields of views to provide an extensive coverage. An individual leaving out of one view would need to be matched in one or more other views at different physical locations over a period of time, and be distinguished from numerous visually similar but different candidates in those views. Thus, re-identification becomes a suitable approach for providing data association when different images of people are captured without a sufficient temporal or spatial continuity. This can be used for long term activity and behavior characterization of people and enabling various applications.

1.2.2 Applications of person re-identifications

Person re-identification methods bear an enormous potential for a wide range of practical applications, ranging from security and surveillance to retail and health care.

- **Cross-camera person tracking.** Understanding a scene through computer vision requires the ability to track people across multiple cameras, to perform crowd movement analysis and to recognise activities. As a person moves from one camera's scene into another one, person re-identification is used to establish correspondence between disconnected tracks to accomplish tracking across the multiple cameras. This allows for the reconstruction of the trajectory of a person across the larger scene.

- **Tracking by detection.** Even in the single-camera tracking setting, the person re-identification could be helpful. Tracking multiple people is not a trivial task, especially in complex and crowded scenarios with frequent occlusions and interaction of individuals. The task itself consists in modeling a vast variety of data present in videos that may include long-term occlusion, and a varying number of people. Recently, some tracking methods named tracking-by-detection make use of human detection techniques to perform the tracking task. The main idea is to detect the persons, estimate their motion patterns and associate detections in different frames. This linking step, called data association, is in fact a kind of re-identification.

- **Person retrieval.** In this case, re-identification is associated with a recognition task. The specific query with a target person is provided and all the corresponding instances are searched in a large database. The re-identification task is thus employed for image retrieval and usually provides ranked lists, similarly related items, and so on.

- **Human-machine interaction.** In a robotic scenario, solving the re-identification problem can be considered as "non-cooperative target recognition", where the identity of the interlocutor is maintained, allowing the robot to be continuously aware of the surrounding people.

- **Long-term human behavior and activity analysis.** For example, analyzing customer shopping trends by observing them touching, surveying, and trying products in stores under different surveillance cameras. Another example, geriatric health care analysis explores the elder people's long-term behavior to assist doctors to make more accurate diagnoses.

1.2.3 Challenges

Solving the person re-identification problem is inherently challenging. To match a person across different scenes, it has to deal with intra-class variation, i.e. the same individual under different views may undergo large appearance changes, and to overcome the inter-class confusion, i.e. different persons can look alike across camera views. Some challenging examples are shown in fig. 1.3. The challenging factors and their effects are explained in the following.

Fig. 1.3 Examples of some person re-identification challenges. Each pair of images shows the same person except (g). (a) viewpoint variation, (b) pose variations, (c) illumination changes, (d) partial occlusion, (e) inaccurate pedestrian detection, (f) accessory change (the person has a back bag in the first image, but not in the second), (g) low resolution, (h) different people with similar clothing.

- **Illumination variation.** Illumination conditions can vary in different camera scenes or during the day. The same person observed under different lighting conditions can have a color difference on the appearance. This increases the intra-class variation.
- **Camera viewpoint variation.** Since the height of the cameras, the distance between the person and the camera and the direction in which the people are facing are varying, different shapes or sizes of pedestrians can be observed under different viewing angles. A person can not be viewed from 360 degrees in a single image. Each view contains

in fact partial information about the person's appearance. Some parts are not visible in one viewpoint, but could be observed in another. In terms of shape, person images from different people from the same viewpoint may look more similar than two images from the same person from different views. The viewpoint variation is one of the most challenging problems which increases at the same time the intra-class variation and the inter-class confusion.

- **Pose variation.** The articulation of the human body leads to deformations in appearances of the same individual observed. A learned model on standing pose will probably fail to detect a running, crouching or a sitting person. Pose variations imply that the body part localization and visibility changes within a given bounding box. and is difficult to predict in the resulting images, which are most often of relatively low resolution and quality.

- **Low resolution.** In most realistic settings, the cost of the required number of cameras in all zones could be very high, so that the coverage is rather sparse, leaving "blind gaps". So cameras are usually installed in high places on walls, and pedestrians are thus usually far away from the camera. Even for high resolution cameras, for a given person, the image could still be of relative low resolution.

- **Inaccurate pedestrian detection.** In an automatic video analysis context, person re-identification methods usually operate on cropped pedestrian images returned by a person detector. However, the performance of existing pedestrian detection algorithms is not that accurate for the re-identification purpose, i.e. detections include too much background or contain only part of the person. Human body regions are therefore not be well aligned across images, which has a serious impact on the re-identification performance of most existing methods.

- **Large number of candidates in gallery set.** A camera network may cover a large public space, like a train station or a campus. Thus there can be a huge amount of candidate for a given re-identification query, and the number of candidate increase over time. The computation for matching with a large gallery set becomes expensive. To alleviate this problem, some temporal reasoning and the spatial layout of the different cameras can be used for pruning the set of candidate matches.

- **Similar clothing** In a large gallery set, there is a high probability that people have similar clothing. Most people in public spaces wear dark clothes in winter, many people wear almost the same blue jeans. This increases the ambiguity and uncertainty

in the matching process. In this case, it's more difficult to find the discriminative signature from the visual appearance of different people.

- **Partial occlusion.** Sometimes people are partially or completely occluded by overlaps with other people or by structures in the environment. If some important or discriminative parts are not visible, the matching fails probably. It happens when people are walking in group or in a crowded public space.

- **Real-time constraint.** In some emergency situations, we should find the location of a suspect immediately. It's important to have a real-time low latency implementation for processing numerous input video streams, and returning query results promptly. The search space for person matching can be extremely large with numerous potential candidates to be discriminated. Thus the searching time is crucial.

- **Clothing or accessories change.** As discussed in section 2.1, we suppose the appearance constancy in the person re-identification problem. But in realistic setting, this hypothesis could be easily violated. The longer the time and space separation between views, the greater the chance that people may appear with some changes of clothes or carried objects in different camera views. For example, taking the bagpack from back in hand or taking off a hat.

- **Camera setting.** The same object acquired by different cameras shows color dissimilarities. The same person with the same clothes can be rendered in different ways. There may also be some geometric differences. For example, the shape of a person may be observed with varying aspect ratios.

- **Small number of images per identity for training.** Since one person may appear very limited times in a camera network, it's difficult to collect much data of one single person. Thus, usually data is insufficient to learn a good model of each specific person's intra-class variability.

- **Data labeling.** This is a common difficulty in the computer vision field. Training a good model robust to all variations in a supervised way couldn't be done without a sufficient amount of annotated data. For a large camera network, manually collecting and annotating amount of data from every camera would be prohibitively expensive.

1.3 Thesis Outline and Contributions

In the following chapter, we will first briefly resume some of the most important and representative person re-identification methods in the literature, categorizing them into four main classes. Additionally, we will present existing public person re-identification datasets.

Besides the literature presentation, our objective is to propose the innovative person re-identification approaches obtaining a competitive or better results compared to the state of the art.

In this thesis, we concentrated on single-image approaches for person re-identification due to constraints from the given application scenario. Using multi-shot or video-based methods may further improve our proposed approaches, e.g. analysing gait of walking pedestrians, and this would be an interesting future research direction. However, the benefit may be limited due to the high inherent redundancy in the data. Also, exploiting videos may be computationally expensive.

As mentioned in the previous section, the intra-class variation and inter-class confusion are two main problems in person re-identification. And to our knowledge, the most important factors causing these two problems are the viewpoint variations and the similar clothing of different people. On one hand, we consider to use more extra cue to help person re-identification task. To overcome the viewpoint variations, we come up to use the pedestrian attribute and orientation information. For the ambiguity of similar appearance, we consider to use the group context information of a pedestrian. On other hand a robust distance metric is indispensable to at same time deal with intra and inter class variance, thus we proposed an improved variation of triplet loss for deep metric learning.

We have four main contributions corresponding to these four different aspects to overcome the challenges and improve the person re-identification performance. These contributions are each presented in the following chapters:

- **Chapter 3** presents a method using pedestrian attributes to enhance re-identification. The attributes are defined as semantic mid-level descriptions of persons, such as gender, accessories, clothing and so on. They could be considered as soft-biometrics and are helpful to cope with the pose and viewpoint variations. Our contribution is two-fold. First, we propose to combine low-level handcrafted features and CNN features to perform attribute recognition. Second, we propose a CNN based person re-identification framework composed of an identity classification branch and of an attribute recognition branch. These two cues are combined at a later stage in the framework to perform person re-identification.

- **Chapter 4** presents an orientation-specific CNN for person re-identification. The method is designed to deal with large viewpoint variations. We consider that the images under different orientations are from different domains. So our contribution is a mixture-of-expert deep CNN to model the multi-domain pedestrian images for person re-identification. The proposed framework performs body orientation regression in a gating branch, and in another branch separate orientation-specific layers are learned as local experts. The combined orientation-specific CNN feature representations are used for the person re-identification task.

- **Chapter 5** presents a novel listwise loss function corresponding to the ranking disorders of the gallery as the third contribution. Learning a similarity metric for person images is one of the most crucial aspects of person re-identification. The proposed method is based on a list of instances and performs an "online" ranking to calculate an evaluation gain. A gain-based weighting is introduced in the loss function to optimize directly the evaluation measures of person re-identification.

- **Chapter 6** presents a novel method to extract group feature representations that are invariant to the relative displacements of individuals within a group. Then we use this group feature representation to perform group association under non-overlapping cameras. Since people often walk in groups and even tend to walk alongside strangers, we propose a neural network framework to combine the group cue with the single person feature representation to improve the person re-identification performance by reducing ambiguities from similar clothing.

Finally, chapter 7 concludes this work with a short summary over the different contributions and some perspectives on future research directions.

Chapter 2

Literature review

2.1 Overview

As previously presented, this thesis proposes four different person re-identification approaches based on deep learning. So in this chapter, we present the literature of two topics. The first section introduces the deep learning techniques. We start from introducing the classical neural networks, then we present Convolutional Neural Networks (CNN), and some of its complex variants as well as the Siamese Neural Networks which are often applied in person re-identification. The second section concerns the state-of-the-art in person re-identification. We will introduce some representative approaches from different families based on feature extraction, metric learning, deep learning etc.

2.2 Deep Learning Introduction

In recent years, neural network-based deep learning algorithms become a popular branch of machine learning. Deep learning algorithms attempt to model high-level abstractions in data by using multiple processing layers with complex structures, or otherwise composed of multiple non-linear transformations. They have shown to be able to outperform state-of-the-art methods in many tasks in the fields of computer vision, natural language processing, robotics, just to name a few. In this section, we will introduce the main concept of deep learning and discuss why it is effective.

2.2.1 Neural networks

As their name indicates, neural networks are inspired from biology and the human brain. Human behaviors are controlled by a nervous system composed of nerve cells or neurons. In the human brain, there are around 85 billion neurons. As shown in Fig. 2.1, a neuron reacts with other neurons to pass the information. When the dendrites of a neuron receive some excitatory input, the membrane potential of the neuron increases gradually. If the membrane voltage reaches a specific threshold, an action potential is initiated and propagated along its axon to post-synaptic neurons.

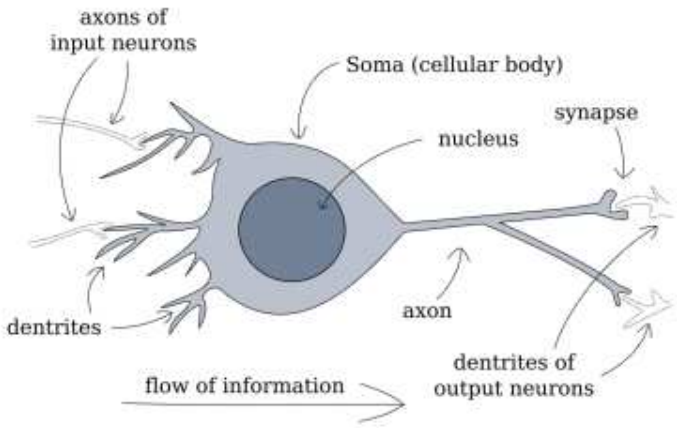


Fig. 2.1 Structure of a biological neuron

Mathematically, we model an artificial neuron as a function which calculates a weighted sum of the input vector x with a weight vector w, adds a bias term b and transforms the sum with a usually non-linear function σ called the activation function:

$$y = \sigma(\sum_i w_i x_i + b) = \sigma(w^T x + b) \tag{2.1}$$

We have several choices for activation function (see Fig. 2.2). The thresholding function is firstly proposed, but not much used due to its non-derivability. The most common choices are the sigmoid function, like the hyperbolic tangent or the logistic function, or the Rectified Linear Unit (ReLU) function.

Artificial neurons can only make a simple computation by themselves and act as a weak classifier. However, they can be arranged in neural networks to perform more complex operations. Although neurons can in theory be arranged quite arbitrarily, in practice they are often arranged in an acyclic graph which means that the input of a neuron does not depend on its output, even indirectly. Neural networks organized with such a topology are referred to as feed-forward neural networks because the activations can be propagated forward in the

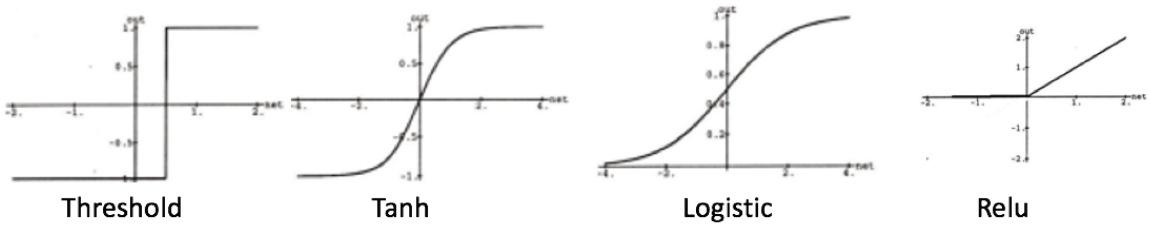


Fig. 2.2 Different neural activation functions.

network. By contrast, recurrent neural networks can contain cyclic connections. Recurrent neural networks are potentially better at modeling dynamical systems, but the presence of cycles makes training much more difficult.

Multi-Layer Perceptrons (MLP) are a popular feed-forward neural networks. The classic MLP contains 2 layers (in addition to the input layer): an output layer and one hidden layer. Each neuron has connections to neurons of the preceeding layer. When each neuron is connected to all neurons of the preceeding layer (as it is the case with most recent deep neural networks), we speak of a fully connected layer (see Fig. 2.3).

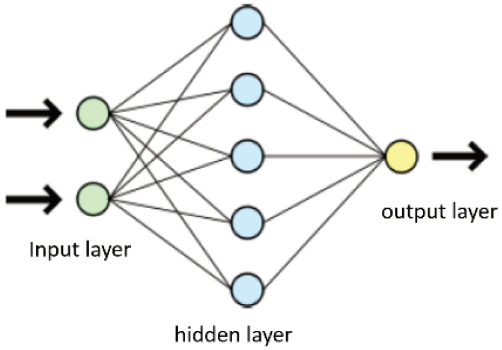


Fig. 2.3 Illustration of a Multi-Layer Perceptron

The basic idea of neural network is to combine the elementary simple function to fit whichever function by toning the parameters of the network. The universal approximation theorem [40] states that a feed-forward network with a single hidden layer containing a finite number of neurons can approximate any continuous function on compact subsets of R^n, under mild assumptions on the activation function.

2.2.2 Backpropagation

The Backpropagation algorithm [93] is the most common and maybe the most universal training algorithm for feed-forward neural networks. It is an application of the gradient

descent optimization algorithm in the context of neural network. The word "backpropagation" is an abbreviation, for "backward propagation of errors". The Backpropagation algorithm can be divided into two phases: (1) a forward phase from the input layer to the output layer to compute the neural activations; (2) a backward phase from the output layer to the input layer to compute the errors, backpropagate the gradients and update the weights.

A cost function E is calculated between the computed outputs and their desired targets for a batch of training samples. We note also h_j^l the output of the l^{th} neuron of the j^{th} layer of the network, g_j^l the input for the activation function of the l^{th} neuron of the j^{th} layer. $W_{i,j}^{k,l}$ represents the weight between the k^{th} neuron of the i^{th} layer and the l^{th} neuron of the j^{th} layer, σ represents the activation function. We calculate the gradient of the cost function with respect to the weight of the network as follows:

$$\frac{\partial E}{\partial W_{i,j}^{k,l}} = \frac{\partial E}{\partial h_j^l}\frac{\partial h_j^l}{\partial g_j^l}\frac{\partial g_j^l}{\partial W_{i,j}^{k,l}} \tag{2.2}$$

We can develop as:

$$\frac{\partial E}{\partial W_{i,j}^{k,l}} = \frac{\partial E}{\partial h_j^l} \times \sigma' \times h_i^k \tag{2.3}$$

The input of the activation function of the $j+1^{th}$ layer is the weighted sum of the outputs of the neurons in the previous layer. The gradient of the cost function with respect to the outputs of the neurons can be written as:

$$\frac{\partial E}{\partial h_j^l} = \sum_m \frac{\partial E}{\partial h_{j+1}^m}\frac{\partial h_{j+1}^m}{\partial g_{j+1}^m}\frac{\partial g_{j+1}^m}{\partial h_j^l} \tag{2.4}$$

Then,

$$\frac{\partial E}{\partial h_j^l} = \sum_m \frac{\partial E}{\partial h_{j+1}^m}\sigma'(g_{j+1}^m)W_{j,j+1}^{l,m} \tag{2.5}$$

Eq. 2.5 has a recursive form. The gradient of cost function with respect to the outputs of the j^{th} layer depends on that of the $j+1^{th}$ layer. The computation propagates backwards. Thus, we can calculate the gradient of the loss with respect to all the weight using Eq. 2.3. Then we can update the weights by gradient descent with a learning rate α:

$$\Delta W_{i,j} = -\alpha\frac{\partial E}{\partial W_{i,j}} \tag{2.6}$$

There is a method which could speed up the convergence. It consist in integrating the update calculated in the previous iterations using a coefficient η called momentum.

$$\Delta W_{i,j}^{t} = -\alpha \frac{\partial E}{\partial W_{i,j}^{t}} + \eta W_{i,j}^{t-1} \tag{2.7}$$

2.2.3 Motivations and problems of deep neural networks

By definition, a deep network consist of multiple layers, but a feed-forward network with a single hidden layer is already a universal approximator. So what is the motivation to form a deep neural network? There could be 2 main reasons. Firstly, the deep network is more efficient. Some mathematic results [35] show that certain function, which is representable by a neural network of depth d, need an exponential number of parameters with a network of depth d-1. However, these results are rather theoretical, and it is not clear to what extend this holds in practical applications

The second motivation is to learn a hierarchy of features with increasing level of abstraction. One neuron in the first hidden layer is active when a certain feature is present in the input. Then the activation of a neuron in the next layer means that a group of these features are present. For example, in a neural network trained for object recognition, the first hidden layer is expected to detect elementary visual features like edges, the next layer is expected to detect maybe a part of the object as a texture, and more and more advanced concepts are extracted in succeeding layers. This is in analogy to the visual cortex in the human brain that ressembles a deep architecture with several processing stages of increasing level of complexity and abstraction.

Although, deep networks extract rich and high level features and reduce the need for feature engineering, one of the most time-consuming parts of machine learning practice, increasing the depth does not necessarily improve their performance. Deep learning algorithms also have mainly two disadvantages. Firstly, they easily suffer from the over-fitting problem, that means the model does not generalize well to real-world cases although it fits the training data well. Also the deeper the network, the more difficult to train and the more training data we need for convergence.

The second disadvantage is the vanishing gradient problem [7]. As the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient infinitively small. The error gradients vanish quickly in an exponential fashion with respect to the depth of the network. As a result, as the network goes deeper, its performance gets saturated or even starts degrading rapidly. It turns impossible to train a deep network. To alleviate these two problems, a number of solutions have been proposed. We will introduce some of them, especially for vision tasks, in the following sections.

2.2.4 Convolutional layer

For the computer vision task, the input data are images of sizes usually ranging from several hundreds to several tens of thousands of pixels. If a neural network processed this input matrix with only fully connected neurons, the number of parameters to train would be very large, leading to a high risk of overfitting. The convolutional layer was invented to deal with this problem and became the first successfully trained deep neural network.

The convolutional layer uses two basic ideas: local receptive fields and shared weights to reducing the complexity of the neural network. A CNN receives a matrix as the input, but connects a hidden node to only a small region of nodes in the input layer, since the spatial correlation is local. This region is called the local receptive field for the hidden node. Moreover all the mappings between a local receptive field and a hidden node share the same weights, since the features are not specific to some regions in an image. These two properties reduce dramatically the numbers of parameters of the network.

By applying these two principals, a fully-connected neuron layer is transformed into a convolutional layer as follows:

$$y = \sigma(W * X + b) \tag{2.8}$$

where X is the input image W is the weight matrix, also called a 'filter' or 'kernel'. b is the bias term. The function σ is an activation function and the operator * represents the discrete convolution operation. For a two-dimensional image X as input, the convolution operator is defined as:

$$(W * X)(i, j) = \sum_m \sum_n X(m, n) W(i - m, j - n) \tag{2.9}$$

Intuitively, the output of the convolutional layer is formed by sliding the weight matrix over the image and computing the dot product (see Fig. 2.4). This resulting matrix is called "activation map" or "feature map". In image processing, convolution operations can be employed for edge detection, image sharpening and blurring just by using different the numeric values of the filter matrix. This means that different filters can detect different features from an image and capture the local dependencies in the original image. In convolutional layers, the convolutional kernel or filter, i.e. the coefficients of the weight matrix W, are learnt automatically by the backprop algorithm, and one layer usually contains several such convolution kernels and resulting feature maps.

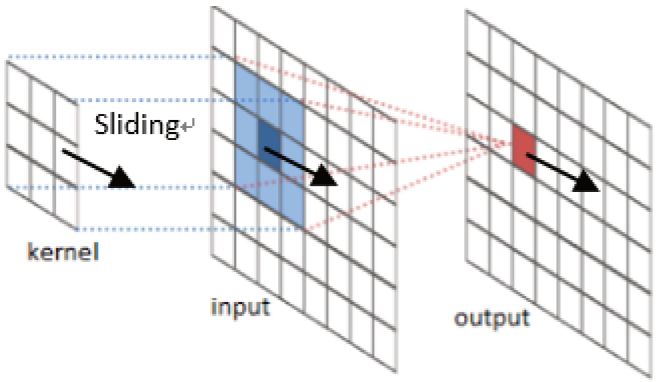


Fig. 2.4 Illustraction of the operation of convolutional layer

2.2.5 CNN architectures

In this section, we will introduce various import CNN architectures from the literature and explain how the basic architecture was improved to achieve state-of-the-art performance in image classification.

LeNet [53] is the most classic CNN which has been developed by Yann LeCun in the early 1990s for handwritten character recognition. The CNN contains 3 types of layers. First, a convolutional layer performs convolution operation on the input image. As the convolution coefficents are learnt automatically by backpropagation, which means that there is no "manual" extraction of features. Then, a pooling layer aggregates the information in a local region and delivers the summarized statistic to the next hidden layer. It can be the maximum or average value, for examples. This allows to reduce the number of parameters and to be invariant to the small translations in the Image. Another convolution and pooling layer repeat this operation to extract higher-level features. At the end, the completely connected layers perform the classification as a feed-forward neural network (like an MLP).

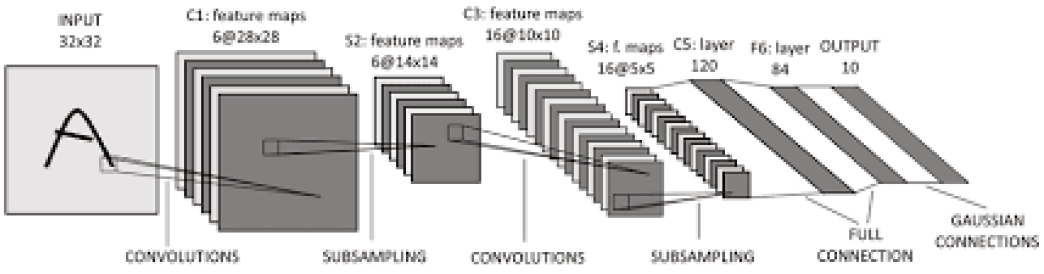


Fig. 2.5 Diagram of LeNet-5

AlexNet [49] is the winning model of the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2012. ImageNet is a dataset of over 15 million labeled high-

resolution images belonging to roughly 22,000 categories. ILSVRC uses a subset of ImageNet with roughly 1000 images in each of 1000 categories. AlexNet improved the ImageNet classification accuracy significantly in comparison to traditional approaches. It is composed of 5 convolutional layers followed by 3 fully connected layers. In Alexnet, the authors proposed solutions of vanishing gradient problem and over-fitting problem.

A new activation function named ReLU (Rectified Linear Unit) is used in the network for the non-linear transformation. The advantage of the ReLu is that it needs light computation and, more importantly, it alleviates the vanishing gradient problem. Since one reason for this problem is that the derivative of sigmoid becomes very small in the saturating region. But the derivative of ReLU is equal to 1 when x is greater than zero, but otherwise it is 0 (see Fig 2.2). So the advantages of ReLU is that when its derivative is back-propagated there will be less degradation of the error signal.

The over-fitting problem is also reduced by the dropout technique, which randomly drops some units after every fully-connected layer from the network during training. Dropout technique has a probability p and it is applied at every neuron of the feature map separately. It randomly switches off the activation with the probability p. One motivation of dropout is to make units learn meaningful features independent from others and to avoid co-adaptations among them. Another view of dropout is related to model ensembles. In fact different activated subsets of neurons represent different architectures and all these architectures are trained in parallel with weights given to each subset. The weighted sum of these random architectures thus actually corresponds to an ensemble of different neural networks.

VGG [101] net replaced large kernel-sized filters in Alexnet (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. Multiple convolutional layers with smaller kernel size can perceive the field of the same size as a larger size kernel with fewer parameters. A better performance can be achieved since multiple non-linear layers increases the depth of the network which enables it to learn more complex features.

ResNet [36] or deep residual network has been proposed by He *et al*. in 2015. The authors noticed a phenomenon with training a deep network: When deeper networks starts converging, a degradation problem has been shown. With the network depth increasing, accuracy gets saturated and then degrades rapidly. The authors of Resnet point out that a deeper network should at least get the same accuracy as a shallow network, since a deeper model's early layers can be replaced with shallow network and the remaining layers can just act as an identity function.

To overcome this degradation problem, the authors proposed to learn a residual function instead of learning a direct mapping function H(x) (A few stacked non-linear layers with

x being the input of the layers). The residual function is defined as F(x)=H(x)-x, where x represents the identity function. So we can reform the equation as H(x)=F(x)+x. The author's hypothesis is that it is easier to optimize the residual mapping function F(x) than to optimize the original, unreferenced mapping H(x). Intuitively, it is easier to learn a function like F(x)=0 rather than F(x)=x using stack of non-linear convolutional layers as function.

To implement this idea in a CNN, the authors add some so called skip-connections, shown in Fig 2.6, to added the input of one layer to the output after one or more layers. This essentially drives the new layer to learn something different from what the input has already encoded. It also alleviate the vanishing gradients problem by allowing the gradient to flow without any changes from the top layers to the bottom by means of identity connections. It leads to the fact that very, very deep networks can be trained. The winning residual network of the ImageNet Challenge 2015 has 152 layers.

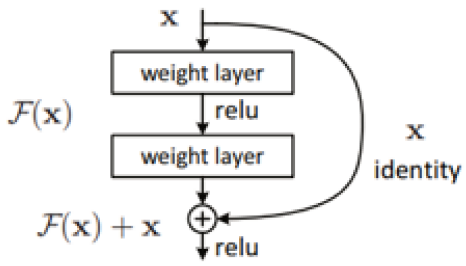


Fig. 2.6 A residual learning block used in deep residual neural networks

With these proposed robust CNN models, the classification accuracy on Imagenet has been continuously improved. However in practice, deep convolutional neural networks have a huge number of parameters, often in the range of millions. Training such q network on a small dataset greatly affects the model's ability to generalize, often resulting in overfitting. Therefore, one practical solution is fine-tuning an existing neural network model. Existing networks that are trained on a large dataset like the ImageNet can be used as initialization for the network. And we replace the last layer and continue to train on the available smaller dataset on another task. Usually a small learning rate should be used. If the provided dataset is not drastically different in context to the original dataset (e.g. ImageNet), the pre-trained model will already have learned features that are relevant to the new task and adjust the learned features to the new task.

2.2.6 Siamese and triplet neural networks

Siamese Neural Networks are neural architectures that receive a pair of examples at the input and produce an output vector. They learn a non-linear similarity metric by repeatedly being

presented pairs of positive and negative examples, i.e. pairs of examples belonging to the same class or not. The principal idea is to train the neural network to map the input vectors into a non-linear subspace such that a simple distance, e.g. the Euclidean distance, in this subspace approximates the "semantic" distance in the input space. That means, two images of the same category are supposed to yield a small distance in this subspace and two images of a different category a large distance.

Siamese Neural Networks have first been presented in [8] using Time Delay Neural Networks (TDNN) and applied to the problem of signature verification, i.e. to verify the authenticity of signatures. This idea was then adopted by Chopra *et al.* [19] who used Siamese CNNs and employed them in the context of face verification. More precisely, the system receives two face images and has to decide if they belong to the same person or not.

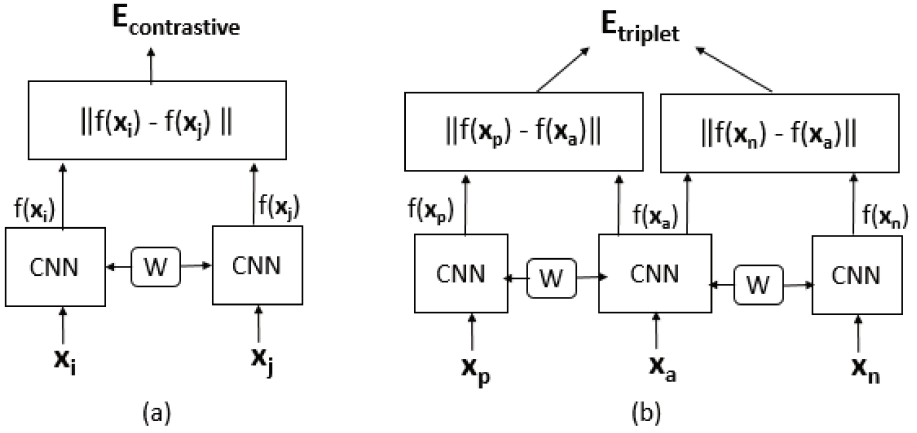


Fig. 2.7 Diagram of (a) Siamese neural network and (b) Triplet neural network.

The most common loss function used by the Siamese architecture is the contrastive loss. The contrastive loss function was based on the Euclidean distance. The objective is to minimize the distance between a similar pair and to separate any two dissimilar data with a distance margin. So the contrastive loss is defined as :

$$E_{contrastive} = \frac{1}{N}\sum^{N} y \cdot \|f(x_i) - f(x_j)\|_2^2 + (1-y) \cdot \max(m - \|f(x_i) - f(x_j)\|, 0)^2, \quad (2.10)$$

where x_i, x_j are an input pair. m is a constant margin, y is the label of the input pair. $y = 1$ for positive pairs and $y = 0$ for negative pairs, f is the projection of the neural network.

The parameters are shared between the sub-networks in a Siamese architecture, that means the weights of the two sub-networks are the same and updated in the same way. Weight sharing can guarantee that the learned distance metric is symmetric, i.e. $d(a,b) = d(b,a)$

and two similar images can not be mapped to very different locations in feature space by two different sub-neural networks. The training is performed with standard backpropagation algorithm. However, the gradient is additive across the sub-networks due to the weight sharing:

$$\Delta W = -\alpha \left(\frac{\partial E}{\partial f(x_i)} \frac{\partial f(x_i)}{\partial W} + \frac{\partial E}{\partial f(x_j)} \frac{\partial f(x_j)}{\partial W} \right) \tag{2.11}$$

where W is the weight of the neural network and α is the learning rate.

The triplet neural architecture is a variant of the Siamese architecture. Instead of having two sub-networks with pairs of images as input, the triplet neural network is composed of three sub-networks (see Fig 2.7). The network is presented with a triplet of images composed of an anchor example, a positive image from the same person as the reference and a negative image from a different person. The weights of the network for the three input images are shared like Siamese network. The triplet loss function is based on a relative distance rather than an absolute distance in contrastive loss. The loss function is defined as:

$$E_{triplet} = \frac{1}{N} \sum^{N} max(\|f(x_a) - f(x_p)\| - \|f(x_a) - f(x_n)\| + m, 0), \tag{2.12}$$

with m being a constant margin, x_a,x_p,x_n are respectively the anchor, positive and negative inputs. The network gets updated when the negative image is nearer than the positive image to the reference image. During training, for a given triplet, the loss function pushes the negative example away from the reference in the output feature space and pulls the positive example closer to it.

Compared to the standard neural network, the Siamese or Triplet architecture has several advantages:

- The Siamese or triplet neural network is able to learn on data pairs or triplets instead of fully labeled data. In other words, the Siamese/Triplet neural network is applicable for weakly supervised cases where we have no access to the labels of training instances: only some side information of pairwise relationship is available.

- The Siamese/Triplet neural network has a greater generalization ability. For the case where there is no common class between training and test set, the learned model can be easily applied to unseen classes in the training set.

- When there is a very large number of classes like tens of thousands or there are very few images per class, it is difficult to train a good classifier and the Siamese/triplet neural network might be a better option.

- It is more practical to update continuously the Siamese/triplet models. When there are new classes added in the training set, we can update the Siamese/triplet model directly with new class images, which is impossible for the model trained with classification loss.

2.3 Person Re-identification

Approaches for person re-identification are generally composed of an appearance descriptor to represent the person and a matching function to compare those appearance descriptors. Over the years, contributions have been made to improve both the representation as well as the matching algorithm in order to increase robustness to the variations in pose, lighting, and background inherent to the problem. Recently, deep learning approaches have been applied to person re-identification and achieved state-of-the-art results. Deep learning approaches for person re-identification learn visual feature representations and a similarity metric jointly. In literature, person re-identification is mostly performed using the person appearance in single color images. However, many approaches use other cues to perform the person re-identification task like temporal information, depth images, gait, camera topology etc. We will introduce all these different types of approaches in this section. At the end, we will present some common used datasets as well.

2.3.1 Feature extraction approaches

Like object recognition tasks, the appearance of pedestrians from static images can be characterized from three aspects: color, shape, and texture. Color histograms are widely used to characterize color distributions. In order to be robust to lighting variations and the changes in photometric settings of cameras, some photometric transformation or normalization methods are proposed:

- In order to model the changes in appearance of objects between two cameras, a brightness transfer function between each pair of cameras is learned from training data in [90]. This transfer function is used as a cue in establishing appearance correspondence.
- Bak *et al*. [4] applied a histogram equalization technique which is based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging conditions (lighting or device). The approach is based on the idea that among all possible histograms, a uniformly distributed histogram has maximum entropy. They apply the histogram equalization to each of the color channels (RGB) to maximize the entropy in each of those channels and obtain the invariant image.

- Liao *et al*. [64] applied the Retinex algorithm to preprocess person images. Retinex considers human brightness and color perception. It aims at producing a color image that is consistent to the human observation of the scene. The restored image usually contains vivid color information, especially enhanced details in shadowed regions.

Only color features alone are not discriminant enough to distinguish people from a large gallery set since the clothing color of some people could be very similar. Therefore, color features are often combined with shape and texture features. Shape context is widely used to characterize local shape structures. Its computation is based on edge or contour detection. However, shape features are subject to the variations in viewpoints and poses, shape features can be less important to re-identify, but are often used for body segmentation as a pre-processing step. Texture features showed to be more effective for person re-identification. Many texture descriptors have been proposed in the literature on object recognition, such as Gabor filter banks [27], Scale-invariant feature transform (SIFT) [73], Local Binary Pattern (LBP) [88], and region covariance [109]. Almost all the proposed descriptors in the literature for person re-identification are the combination of a subset of these texture features and color histograms.

In order to get a feature descriptor which is discriminant, and at same time, robust to the different variations, various extraction strategies have been proposed in the literature. Here we divide the approaches into three classes: patch-based descriptors and body part-based descriptors and stripe-based descriptors, as shown in Fig. 2.8.

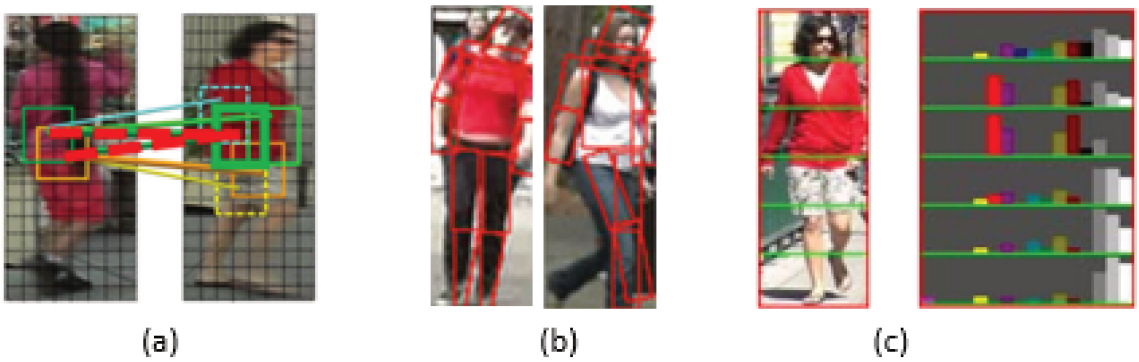


Fig. 2.8 Illustrations of representative approaches of three feature extraction strategies. (a) Patch-based descriptors extracted from a dense grid. Patch matching is performed for person re-identification [99]. (b) Body part-based descriptors extracted from segmented body parts to form a global feature vector [18]. (c) Stripe-based descriptors extracted from horizontal bands to deal witch viewpoint variance [127].

Patch-based descriptors

The holistic representation of these features shows a high robustness but a low discriminative power, because of losing local detail information. A typical solution is to apply the color histograms and texture filters on a dense grid. These local features extracted on small patches can be more discriminative for person re-identification.

Person images are densely segmented into a grid in [135]. 10×10 patches are densely sampled with a step size of 5 pixels. Then, a LAB color histogram is extracted from each patch. To robustly capture color information, LAB color histograms are also computed on downsampled scales. For the purpose of combination with other features, all the histograms are L2 normalized. To handle viewpoint and illumination change, SIFT descriptors are used as a complementary feature to color histograms. As with color histograms, a dense grid of patches are sampled on each human image. Dense color histograms and dense SIFT features are then concatenated as the final multi-dimensional descriptor vector for each patch. For the patch matching, to reduce the computation cost, an adjacent region constraint is imposed on searching matched patches.

Similarly, Liu *et al*. [72] extract the HSV histogram, gradient histogram and the LBP histogram for each local patch. Then they applied local coordinate coding which is a high dimensional nonlinear learning method with data distributed on manifolds. Local coordinate coding approximates a given input point as a weighted linear combination of a few elements called anchor points.

The Bag-of-Words (BoW) model is used in [137]. The 11-dim color names descriptor [127] is extracted for each local patch, and aggregated them into a global vector through a BoW model. The codebook is trained on another pedestrian detection dataset. After generating the codebook on training data, the feature response of each patches are then quantified into visual words. Each pedestrian image is represented as a visual word histogram, which is weighted by TF-IDF (Term Frequency–Inverse Document Frequency) scheme.

Shen *et al*. [99] extract Dense SIFT and Dense Color Histogram as [135] from each patch. They introduced a correspondence structure to encode cross-view correspondence pattern between cameras, and perform a patch matching process by combining a global constraint with the correspondence structure to exclude spatial misalignments between images.

Body part-based approaches

However, these patch-based descriptors cannot encode spatial information. It is also possible to directly compare the features extracted on a fixed dense grid. But it is sensitive to misalignment, pose variation, and viewpoint variation. In order to resolve this problem and

increase the discriminative power, several approaches exploit the prior knowledge of the person geometry or body structure and try to partition the body intelligently to obtain a pose invariant representation:

Wang *et al*. [119] proposed a model of shape and appearance context. The model densely computes a local HOG descriptor in the Log-RGB color space. Then the person body is segmented into regions by two steps. The first step segments images according to their HOG appearance, the second identifies parts by a modified shape context which uses a shape dictionary learnt a priori. Then region labeled images are generated. The context of appearance and shape is modeled by co-occurrence matrix which describes probability distributions and their correlations over the image regions. Descriptor matching is done using the L1-norm distance. This approach just segments person image into uniform regions then use the local features in regions. However, this works only if the viewpoints are similar. Thus, variations of viewpoint or occlusion will cause big differences in the segmentation.

Some approaches segment images into meaningful parts like torso, legs, which are semantic and more robust to the viewpoint variation. One well-know method is Symmetry-Driven Accumulation of Local Features (SDALF) proposed by Farenzena *et al*. [26] (see Fig. 2.9), which partitions the human body into salient and meaningful parts by exploiting asymmetry and symmetry principles. Firstly, the foreground is extracted from the person image, then the person body is partitioned into three regions (head, torso and legs) looking for the asymmetrical axis and symmetrical axis by calculating two operators. The chromatic bilateral operator (C) calculates the sum of the difference between the points located symmetrically with respect to the horizontal axis, and the spatial covering operator (S) calculates the difference of foreground areas for two regions above and below the axis. So the asymmetry axe separating the head and the body is the one which maximizes S and the axis separating the torso and the legs is the line which maximizes C and minimizes S. The vertical symmetrical axis for the torso and leg parts is the vertical line minimizing C and S.

After the head part, the body and leg parts are extracted, then a combination of three features are extracted for the torso and leg regions. The head part is not taken into account. The three features are the Weighted HSV Color Histogram (WCH), Maximally Stable Color Regions (MSCR), and Recurrent Highly Structured Patches (RHSP). WCH takes into consideration the distance to the vertical axes. Pixels located far from the symmetry axis belong to the background with higher probability and will have less weight. MSCR are extracted by grouping pixels of similar color into small stable clusters. Then, the regions are described by their area, centroid, second moment matrix, and average color. RHSP uses entropy to select textural patches that have strong edges and that are highly recurrent. The

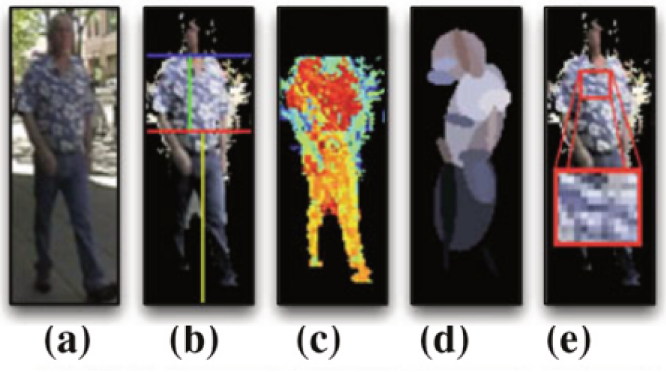


Fig. 2.9 Illustration of the SDALF descriptor [26]. (a) Original image ,(b) segments for meaningful body parts (c) Weighted Color Histogram (WCH), (d) Maximally Stable Color Regions (MSCR), (e) Recurrent Highly Structured Patches (RHSP).

final phase is the feature matching between two person instances, where the three specific feature distances are weighted and unified to a joint matching distance.

Bak *et al*. [4] proposed an approach based on spatial covariance regions by using body part detectors. A human body part detector, based on HOG is trained and applied to detect 5 body parts: the top, the torso, legs, the left arm and the right arm. To build a human signature, a three-level pyramid represents the full body region, the body parts, and sub regions inside the body parts. A covariance matrix capturing pixel location, color, and gradient information is computed for each of these regions.

To better exploit the prior knowledge of body structure, Cheng *et al*. [18] used Pictorial Structures to localize the body parts and match their descriptors. They made a modification of pictorial structures to better localize body parts using multiple images of a person to guide the MAP (Maximum A Posteriori) estimates the body configuration. The body configuration is composed of head, chest, thighs, and legs on pedestrian images and each part is used to extract HSV histograms and MSCRs. Re-identification is performed by matching signatures coming from articulated appearances.

Stripe-based approaches

Since person images from video surveillance cameras are of low resolution, generally, it is difficult to get an accurate body part segmentation. The body part detection errors can influence re-identification results, especially in difficult scenes like complete profile images and partial occlusion. One solution is to perform a rough segmentation. Since we know the pedestrian images are seen from an arbitrary horizontal viewpoint, we can disregard the horizontal dimension as it is less relevant, which leaves us with a set of stripes which span the entire horizontal dimension as feature regions.

Gray *et al*. [33] first proposed to divide the pedestrian image into 6 equally-sized horizontal stripes. The approximately correspond to the image regions of the head, upper and lower torso, upper and lower legs and fee. In each stripe, 8 color channels (RGB, HS, YCbCr) and 19 texture channels (Gabor and Schmid filter banks) are represented. Similarly, Mignon *et al*. [85] build the feature vector from RGB, YUV and HSV channels and the LBP texture histograms in horizontal stripes.

Yang *et al*. [127] introduced the salient color names based color descriptor (SCNCD) for pedestrian color descriptions. SCNCD utilizes salient color names (e.g. "black", "red", "yellow") to guarantee that a higher probability will be assigned to the color name which is nearer to the color. Based on SCNCD, color distributions over color names in different color spaces are then obtained and fused to generate a feature representation. SCNCD is extracted in six horizontal stripes of equal size.

The problem of feature extraction in stripes is that it can be influenced by the background clutter and it may also lose spatial details within a stripe, thus affecting its discriminative power. To solve these problems, some approaches proposed to combine the stripe and patched based descriptor to encode salient local patch in horizontal stripes.

Covariance matrices were adopted in [76]. The underlying features are Gabor filter response magnitude images extracted from different spatial scales. Neighboring scale responses are grouped to form a single band and magnitude images are computed using the max operator within each band. The appearance model is not the covariance matrices but differences between matrices between consecutive bands.

Liao *et al*. [64] proposed LOcal Maximal Occurrence (LOMO) features. Two scales of Scale-Invariant Local Ternary Patterns (SILTP) and an $8 \times 8 \times 8$-bin joint HSV histograms are extracted in sliding sub-windows. The sub-window size is 10×10, with an overlapping step of 5 pixels describing local patches. Following the same procedure, features are extracted at 3 different image scales. For all sub-windows on the same image line, only the maximal value of the local occurrence of each pattern among these sub-windows is retained. In that way, the resulting feature vector achieves a large invariance to view point changes and, at the same time, captures local region characteristics of a person.

Matsukawa *et al*. [81] densely extracted local patches inside a region and regard the region as a set of local patches. They firstly model the region as a set of multiple Gaussian distributions, each of which represents the appearance of one local patch. The pixel features are extracted : location in vertical direction, gradient orientation and color channel values. The characteristics of the set of patch Gaussians are again described by another Gaussian distribution. The parameters of the region Gaussian are then used as feature vector to represent the region.

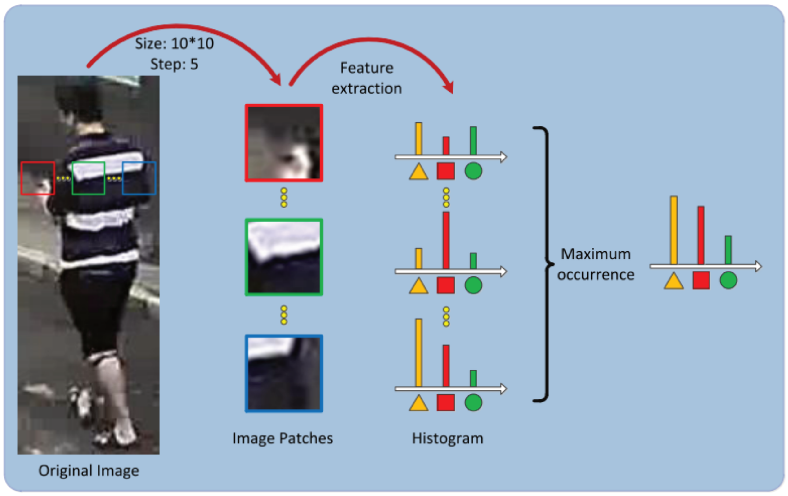


Fig. 2.10 Illustration of LOMO features [64]

2.3.2 Matching approaches

Based on the extracted features, we distinguish two types of matching methods. The first consists of learning a matching function in a supervised manner, and the other learns a distance metric in feature space.

Matching function learning

Given feature based representations of a pair of images, an intuitive approach is to compute the geodesic distance between the descriptors, for instance, using the Bhattacharyya distance between the histogram-based descriptors or the L2-norm between descriptors in a Euclidean space. However, some features may be more relevant for appearance matching than others. To this end, several approaches have been proposed to learn a matching function in a supervised manner from a dataset of image pairs. For instance:

In [98], the high-dimensional feature is transformed into a low-dimensional discriminant latent space using a statistical tool called Partial Least Squares (PLS) in a one-against-all scheme, which considers class information during reduction. For the one-against-all scheme, the discriminative appearance of a person is learned using information about the appearances of other persons. PLS gives higher weights to features located in regions containing discriminative characteristics.

Lin *et al*. [66] proposed an approach based on learnt pairwise dissimilarity profiles. The re-identification problem is considered as a multi-classification problem where each person is a different class. Nearest neighbor classification is performed with a combination of a direct distance and an indirect distance. The direct distance is based on the Kullback–Leibler

divergence between the probability density function of the two images in a 4D space (3 colors + the height coordinate). The indirect distance is based on a pairwise dissimilarity profile which models the properties that are very discriminative between two persons and are learnt from all pairs in the training set.

These two approaches consist in building a model for each person in the training set. However, the drawback is that it heavily depends on the given data. For very crowded scene, it becomes difficult and it has to be re-computed if new samples are added.

Some other approaches hold an idea similar to metric learning, they select more discriminant features using boosting. Gray *et al*. [33] use boosting to find the best ensemble of localized features for matching. Instead of designing the model by hand, boosting is used to construct a model that provides maximum discriminability for a set of training data. The learned model is an ensemble of localized features, each consisting of a feature channel, a region, and a likelihood ratio test for comparing corresponding features. Once the model has been learned, it provides a similarity function for comparing pairs of pedestrian images.

Prosser *et al*. [91] proposed ensemble RankSVM to solve person re-identification as a ranking problem. The goal is to learn a subspace where the potential true match is given as the top rank. Existing methods such as RankSVM do not scale very well on large datasets due to high computational costs and memory requirements with a large number of possible pairs.To address this issue, ensemble learning is applied. The authors train weak SVMs on small subsets of the data and then combine them to form a strong ranker using a boosting principle. In this way, memory costs can be significantly reduced without loss in performance.

Metric learning

Compared to standard generic distance measures, e.g. Euclidean or Bhattacharyya distance, a metric that is learnt specifically for person images is more discriminative for the given task of re-identification and more robust to large variations of person images across views.

Most distance metrics learning approaches learn a Mahalanobis-like distance: $D_2(x,y) = (x-y)^T M(x-y)$ where M is a positive semi-definite (PSD) matrix of witch the elements are to be learnt. The main advantage is that the optimization of the Mahalanobis matrix can be seen as a constrained convex programming problem which can be solved with existing efficient algorithms. However, guaranteeing that M is PSD can be computationally expensive. Hence, several works factorize M as $M = w^T w$, ensuring the PSD constraint and implicitly defining a (potentially low-dimensional) projection into an Euclidean space which reflects the distance constraints. We distinguish two types of methods explained in the following.

Distance constraint based optimization The first class of methods generally defines an objective function based on distance constraints. The global idea of constraints is to keep all

the vectors of the same class closer while pushing vectors of different classes further apart. M is solved by a constrained convex optimization method.

Mignon *et al*. [85] presented Pairwise Constrained Component Analysis (PCCA) which projects the feature vector to a low-dimensional space in which the training constraints are respected by minimizing an objective function penalizing distances greater than a threshold for positive pairs and lower than the same threshold for negative pairs. They also developped a kernelized version Kernel PCCA to deal with non-linearity in the data.

Different from [85], which aims to minimize the distance between examples from the same class in an absolute sense, the objective function of the Probabilistic Relative Distance Comparison (PRDC) model proposed by [139] aims to maximize the probability of a pair of true match having a smaller distance than that of a pair of related wrong match. This is achieved using an iterative optimization algorithm.

Unlike previous approaches learning the metric from pairs of images, in [23], the metric is learned with a distance constraint defined on neighborhood of examples. The method is called Large Margin Nearest Neighbor classification with Rejection (LMNN-R). LMNN learns a matrix that minimizes the distance between each training point and its K nearest similarly labeled neighbors, while maximizing the distance between all differently labeled points, which are closer than the aforementioned neighbors' distances plus a constant margin (see Fig. 2.11). The authors proposed a modified cost function which forces the closest impostors of a training point to be at least a certain distance away, determined by this average distance of all K nearest neighbor pairs in the training set. The net effect of this modification is that a universal threshold on pairwise distances can be used for determining rejection, while still approximately preserving the local structure of the large margin metric learning. The matrix M is finally solved by the Iterative Subgradient optimization method.

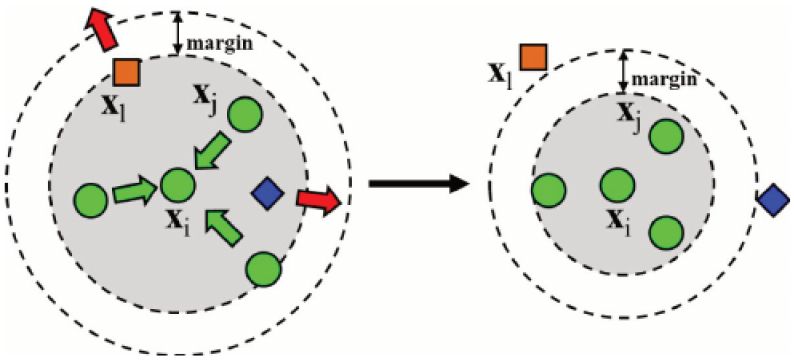


Fig. 2.11 Illustration of Margin Nearest Neighbor classification [23]

All the previous methods solve the matrix M relying on a tedious iterative optimization procedure. Koestinger *et al*. [48] presented their "Keep It Simple and Straightforward Method" (KISSME) to solve the matrx M in a closed form. KISSME learns a distance metric from equivalence constraints and put a probabilistic view on learning a Mahalanobis metric. The distance between examples can be represented as a likelihood ratio test, i.e. the logarithm of the ratio of the probability that the pair is dissimilar and the probability that the pair is similar. By assuming a Gaussian structure of the difference space, we can relax the problem. It turns out that the learnt Mahalanobis matrix is simply the covariance matrix of the pairwise difference for the similar pairs and that for the dissimilar pairs. Solving the problem in a closed form makes this method scalable to large datasets, as it just involves computation of two small sized covariance matrices.

Discriminative subspace learning The methods of the second class are generally variants of the Linear Discriminative Analysis (LDA). The approaches are based on the difference of the feature vectors of two classes. The positive class consists in pairs of images of the same person acquired by different cameras, and the negative class consists in pairs of images of different person acquired by different cameras. They learn directly the projection **w** to a discriminative low-dimensional subspace where the between-class variance is maximized and the within-class variance is minimized. In LDA, the objective function is formulated as:

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{2.13}$$

where S_b and S_w are the between-class and within-class scatter matrices, respectively.

Pedagadi *et al*. [89] proposed LFDA which combines Linear Discriminant Analysis and Locality-Preserving Projection for dimensionality reduction. Both the within-class scatter matrix S_W and the between class scatter matrix S_B are weighted by an affinity matrix of the data. The affinity value between two samples separated by a small Euclidean distance will be higher than two samples separated by a larger distance. A regularized form of the supervised dimensionality reduction was also proposed for the LFDA technique used in second stage of the framework.

The Cross Quadratic Discriminitive Analysis (XQDA) metric [64] combines Quadratic Discriminative Analysis (QDA) and KISSME. The original feature dimensions is generally large. Thus, to apply the KISSME, a PCA is generally applied to perform dimension reduction. But this reduction does not consider the distance metric. So instead of PCA, QDA is applied to find the subspace where between-class variation is maximized and within-class variation is minimized. This subspace projection is solved in a similar way to LDA.

Zhang *et al*. [131] further employed the Null Foley-Sammon Transform (NFST) to learn a discriminative null space which satisfies a zero within-class scatter S_w and a positive

between-class scatter S_b. NFST aims to learn a discriminative subspace where the training data points of each of the C classes are collapsed to a single point, resulting in C points in the space. In order to make this subspace discriminative, these C points should not further collapse to a single point.

2.3.3 Deep learning approaches

As introduced in section 2.2, the deep learning is a collection of machine learning methods which model high-level abstraction in data through multiple layers of nonlinear transformations. After the big success of the well-known Alexnet [49] at ILSVRC 2012, more and more methods based on CNNs are applied to person re-identification. Different from the more traditional methods, feature extraction is implicitly learned by CNNs instead of manually designed. However, we can still find the idea of feature extraction based on stripes, on patches or on body parts.

Stripe-based architecture

Due to pose change across different camera views, features appearing at one location may not necessarily appear in the same location for its paired image. Since all the images are resized to a fixed scale, it is reasonable to assume a horizontal row-wise correspondence.

Yi *et al*. [128] first applied a CNN on re-identification. Given two person images, they are first separated into three over-lapped horizontal stripes respectively and the image pairs are matched by three Siamese Convolutional Neural Networks (SCNN). With parameters being shared between the two sub-networks, the network is more appropriate for the general re-identification task. But the authors proposed that a SCNN that is specific to a pair of cameras can also be trained by not sharing the parameters.

The different stripes are not equally discriminative, and it is reasonable to give more importance to some stripes than to others. Varior *et al*. [112] proposed to integrate a gating layer in a SCNN to compare the extracted local patterns for an image pair at the midium-level and promote the local similarities in stripes along the higher layers so that the network propagates more relevant features to the higher layers of the network. Therefore, the matching gate first summarizes the features along each horizontal stripe for a pair of images and compares it by taking the Euclidean distance along each dimension of the obtained feature map. Once the distances between each individual dimension are obtained, a Gaussian activation function is used to output a similarity score ranging from 0 (dissimilar) to 1 (similar). These values are used to gate the stripe features and finally, the gated features are added to the input features to boost them thus giving more emphasis to the

local similarities across view-points. The matching gate is formulated as a differentiable parametric function to facilitate the end-to-end learning. The drawback of the method is that the feature representation is specific to image pairs. We cannot extract a representation for a single image.

Spatial dependency between stripes are exploited in [113]. Unlike previous approaches using convolutional layers, the authors extracted LOMO [64] and SCNCD [127] features in horizontal stripes. The Long Short Term Memory (LSTM) network, a specific type of recurrent neural network, has been applied to model the spatial correlations between the equence of stripesto enhance the discriminative ability of the deep features. LSTM receives the features extracted in horizontal stripes one-by-one (see Fig. 2.12) and progressively captures and aggregates the relevant contextual information.

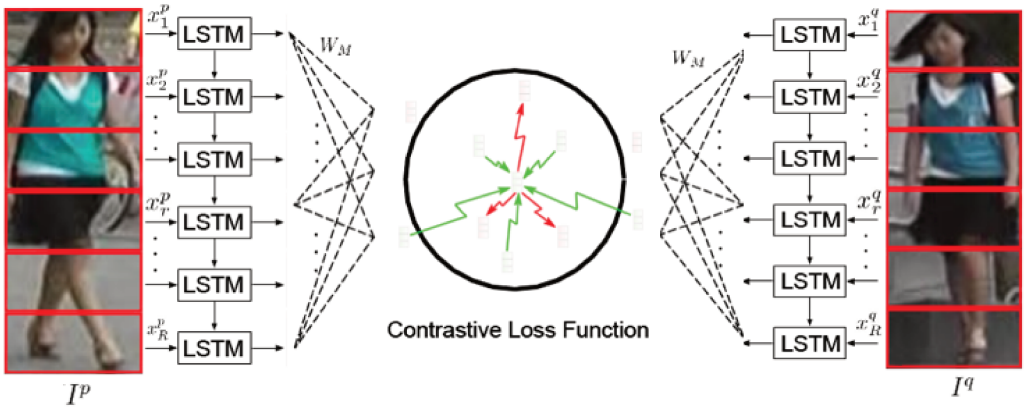


Fig. 2.12 A diagram of the siamese LSTM architecture in [113].

Cheng *et al*. [17] proposed to combine the global and stripe feature extraction. A multi-channel CNN model that learns both the global full-body features in one branch and learns the local parts in other branches. These features are integrated together to produce the final feature representation of the input person.

Patch-based architecture

Some methods directly integrate some kind of patch matching into a CNN architecture to handle misalignment and geometric transformations. Matching all patches in two images is computationally expensive. Thus for the two following representative approaches, one perform the patch matching within horizontal stripes, and the other matches patches in a nearby neighborhood.

Li *et al*. [61] proposed an architecture called Filter pairing neural network. The pair of images observed in two different camera views form the input of the CNN. The photometric

transform between twos cameras are modeled by the convolutional filters of the first layer. Then, the feature maps are divided into patches. For each stripe in one feature map (see Fig. 2.13), their method constructs a matrix in which each element is the product of two patches in this stripe. These displacement matrices encode the spatial patterns of patch matching under the different features. Then, one layer of max-pooling and one layer of convolution are added to obtain the displacement matrices of body parts on a larger scale. Finally, a fully-connected layer combines all the possible part displacements and represents a global geometric transform.

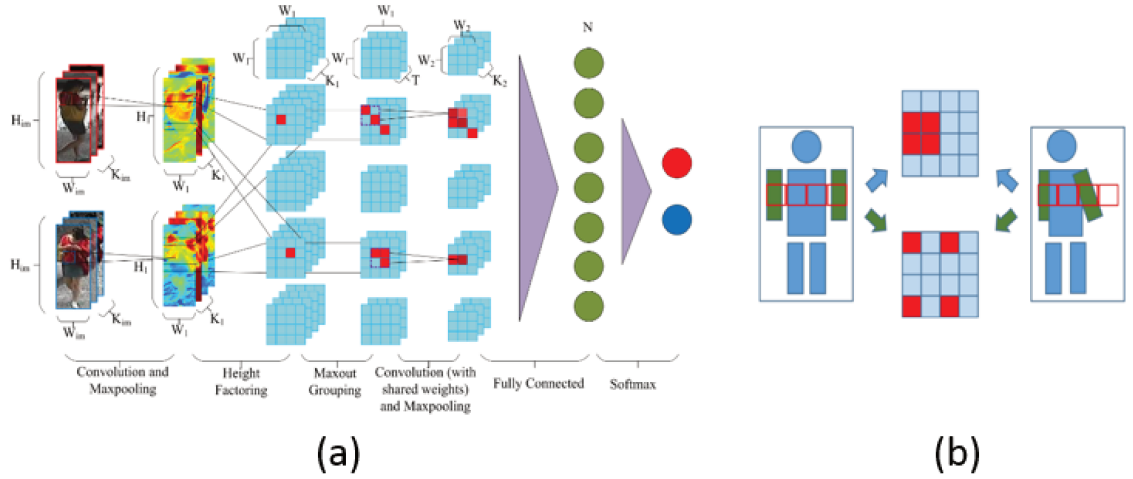


Fig. 2.13 (a) Filter pairing neural network [61]; (b) one stripe generates two displacement matrices. One matrix for blue feature the other for green features.

The architecture is computationally complex. To simplify the network, a cross-input neighborhood difference layer was introduced in [1]. Two layers of convolutions and max-pooling are first performed. The cross-input neighborhood difference layer then compares convolutional image features in each patch of one input image to the same features computed on nearby patches in the other input image. Later a subsequent layer summarizes each patch's neighborhood differences to model the cross-view relationships of the features. Wu *et al*. [122] employed a similar but deep architecture, called PersonNet, achieving a A better accuracy.

Body part-based architecture

In the datasets in which the bounding boxes are annotated by automatically pedestrian detection method, the misalignment of pedestrian in images is a big challenge for person re-identification. In this case, the assumption of division into stripes does not hold any more.

To deal with this, some methods integrate a body part-based feature extraction within a deep neural network architecture.

Zhao *et al*. [133] for example proposed an architecture composed of three components: body region proposal network, feature extraction network and feature fusion network. The body region proposal network is train on a dataset with human landmark annotation. This network is used to locate human body joints from one input image. The feature extraction network takes the person image together with the region proposals as input and computes one global feature vector of the full image and sub-region feature vectors corresponding to the proposed body sub-regions. With the feature fusion network, a final feature vector can be computed by merging the full image feature vector and the sub-region feature vectors together. The final feature vector can be used to distinguish different persons.

Without requiring more body structure annotated data, Li *et al*. [56] proposed to localize latent pedestrian parts through Spatial Transform Networks (STN) [41], which is originally proposed to learn image transformation. To adapt it to the pedestrian part localization task, prior knowledge based constraints are given on the learned transformation parameters: e.g. the predicted parts should be near the prior center points, the predicted parts should have a reasonable extent and the cropped parts should be inside the pedestrian image. Finally, the features of the full body and body parts are learned by separate networks and then are fused in a unified framework to perform person re-identification.

Zhao *et al*. [134] proposed to jointly model the human body regions that are discriminative for person matching with neither prior knowledge nor labeled data and compute a compact representation. An image feature map is first extracted by a deep Fully Convolutional Neural network (FCN). Then the network is connected to several branches. Each branch receives the image feature map from the FCN as the input, detects a discriminative region by multiplying a learned weighting mask (see Fig 2.14), and extracts the features over the detected region as the output. Then all the part features are concatenated to yield the global human representation. The triplet loss is used to train the network in an end-to-end manner.

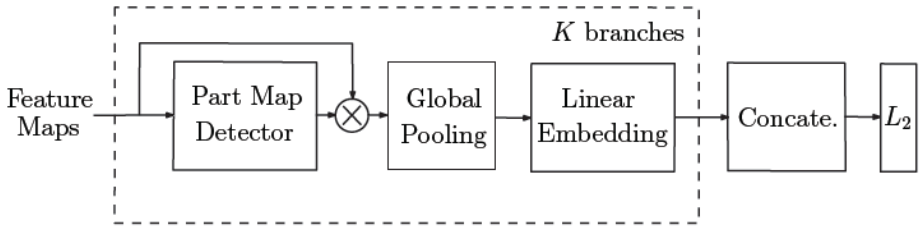


Fig. 2.14 Part-based feature extraction branch of the neural network architecture proposed in Zhao *et al*. [134].

Loss functions

The loss functions learning a non-linear projection into a feature space in which the similarity of pedestrian is well represented. Several loss functions for person re-identification in the litterature. Yi *et al*. [128] used deviance loss in a Siamese network, as follows:

$$J_{deviance} = ln(e^{-2sl} + 1), \tag{2.14}$$

where $-1 \leqslant s \leqslant 1$ is the similarity score and $l = 1$ or -1 is the label. And in [1, 61], the re-identification task is considered as an image pair classification problem deciding whether an image pair is from the same person or not. One disadvantage of pairwise approaches is the data imbalance since there are much more possible negative pairs than positive pairs in the training set. This may lead to over-fitting when using a pairwise loss function. A class weighting can be applied to solve the problem, but it adds a free parameter that may be different for different datasets. The triplet loss, which uses a reference example, a positive example and a negative example, overcomes this problem. Ding *et al*. [24] first applied the triplet loss to train a CNN for person re-identification.

Some methods combine different losses to improve the performance. For example, Cheng *et al*. [17] proposed an improved variant of the triplet loss function combining it with the contrastive loss. Zheng *et al*. [140] combined image pair classification loss and contrastive loss. Chen *et al*. [14] applied a quadruplet loss which samples four images from three identities and minimizes the difference between a positive pair from one identity and a negative pair from two different identities and they combine this quadruplet loss with the triplet loss.

New trends in deep learning for re-identification

New architectures of CNN have continuously been proposed. These new ideas give some new trends to apply CNN to the person re-identification task. One trend could be the generative adversarial network (GAN). Many variations present in person re-identification problems. But with the limited size of datasets, the examples can not represent all possible variations. Some approaches considered to use the generative models. These are models that can learn to create data that is similar to data that we provide. GAN has been proposed in [30] and quickly applied to different tasks. GANs learn two sub-networks: a generator and a discriminator. The discriminator reveals whether a sample is generated or real, while the generator produces samples to deceive the discriminator.

Zheng *et al*. [141] first applied GAN to re-identification problem. The idea is direct, using generated images to augment training data. They adopt the deep convolutional generative

adversarial network (DCGAN) for sample generation, and a baseline CNN for representation learning. And experiments show that adding the GAN-generated data improves the learned CNN embeddings. Liu *et al*. [71] instead of generate whichever pedestrian images proposed to generate images of an identity with various poses with different skeletons. A guider module is proposed to impose the identity of the generated images. Wei *et al*. [121] proposed to solve the domain gap problem by GAN. The problem commonly exists between datasets caused by the variations of camera setting and capture conditions. This leads that the model learned in available training data cannot be effectively leveraged for new testing domains. To relieve the expensive costs of annotating new training samples, the authors performed Image-to-Image Translation by GAN to bridge the domain gap, that means transfering the images in domain to another.

Another trend is the attention mechanism. The deep learning models without attention mechanism relies on all images or all frames in sequence of images. However, on one hand, the displacement of pedestrian in bounding box or the pose variations, may some regions should be paid more attention, on the other hand due to occultation or viewpoints variation, for a sequence images, some frames are more important. To this end, Li *et al*. [63] proposed to jointly learn multi-level soft pixel attention and hard regional attention along with simultaneous optimisation of feature representations. Li *et al*. [58] combine spatial and temporal attentions into spatiotemporal attention models to address the challenges in video-based person re-identification. For spatial attention, a penalization term is used to regularize multiple redundant attentions. Temporal attention is used to assign weights to different salient regions on a per-frame basis to take full advantage of discriminative image regions.

2.3.4 Other cue-based approaches

Almost all previous discussed approaches are based on color images of pedestrian. There are some other cues to perform the person re-identification task or improve the performance by using additional information to increase the discriminative power of the appearance-based approaches. These cues includes biometrics, gait, depth information, camera network topology etc.

Temporal information

From a surveillance camera, an image sequence of a person can be extracted by tracking algorithms. Some methods exploit the temporal information in the sequence to perform

a video based person re-identification. We distinguish three types of use of the temporal information in person re-identification.

In the first case, some early work used the sequence of images to perform image pre-processing. For example, In [26, 6], the video clip is used for background subtraction in order to better extract feature vectors. In [29], the spatio-temporal information is used for body part segmentation. Secondly, some methods just extend single image-based models by aggregating frame-level features. For example, frame features are aggregated by max or average pooling which yield better accuracy in [136]. In [26, 6], the minimum Euclidean distance between two sets of frame-based features is calculated as the distance of the image sequences.

Finally, some methods exploit the temporal information as a cue to distinguish identities. This information is also called gait, which characterizes and discriminates the way people walk. The advantages of gait recognition are that gait information is robust in very low resolution videos and no assumption being made on subject cooperation. Each person seems to have a distinctive way of walking. The differences in gait can be observed from many walking parameters, like gait cycle information, person's centroid vertical oscillations range, the maximum height of the foot when it leaves the floor, etc. For example, Han *et al*. [80] used a gait energy image which is the aligned and averaged binary image over one gait cycle. Gait can be helpful to distinguish people, but in general using only temporal information is not discriminative enough in the case of large scale gallery set. So most approaches use spatio-temporal features. Wang *et al*. [118] uses space-time feature to describe the gait. In the lower part of person images, they calculate Flow Energy Profile (FEP) which is the sum of the norm of optical flow vectors. The candidate sequence consists of 10 frames before and after the local minima of the FEP. HOG3D features extracted on the lower part of each image frame of these sequences. Then multi-instance ranking is performed on the difference of the pair of feature vectors of the candidate sequence. Some deep recurrent neural networks are also applied to exploit temporal information and build a sequence level representation, for example, recurrent neural network used in [84] and LSTM used in [126].

Biometrics

Biometric approaches which use biological characteristics of people are the most accurate way to re-identify people. There are several biometrics which are exploited to be applied to identification such as the face, fingerprints, and the iris. The face represents the most common way for human beings to recognize identity. Face recognition has been a popular subject research since several decades and very good accuracy has been achieved. Some approaches apply face recognition into person re-identification in surveillance video. Vaquero *et al*. [111]

was the first to introduce facial characteristics for person re-identification, like eyewear type, and facial hair type for 3 face parts. The classifiers of facial attributes are based on Haar-like features. These attributes are saved in database to facilitate the zero-shot people search a posteriori. Iris recognition is a method of identifying people based on unique patterns within the ring-shaped region surrounding the pupil of the eye. The iris usually has a brown, blue, gray, or greenish color, with complex patterns that are visible upon close inspection. Finally, identification methods using fingerprints use the impressions made by the minute ridge formations or patterns found on the fingertips. Many research works assume that no two people have the same iris pattern and fingerprints. Although, biometrics represent a viable approach to the person re-identification problem, they are only applicable under very constrained conditions. They require specific sensors, high resolution images, collaboration actions from people or a frontal body orientation.

Attributes

Since biometric cues are not always available, some approaches consider higher-level appearance attributes, so-called soft-biometrics, such as hair-style, shoe-type or clothing-style. An attribute-based person representation is similar to a description provided verbally by a human. This semantic representation can be complementary to a classic feature representation. More detail discussions will be brought up in the Chapter 3.

Depth or 3D information

Depth images are often captured by time-of-flight camera. A depth image can have several advantages over color images. These include for example a much higher robustness to different viewpoints and lighting conditions. On the other hand, the known disadvantages are its sensitivity to solar infrared light and the limited functioning range. Depth image-based re-identification approaches often use the shape and skeleton of body (see Fig 2.15). With depth information, a 3D point cloud and skeleton of a person can be easily extracted. Moreover, with depth value of each pixel, pedestrians can be more easily segmented from background in the corresponding image. Munaro *et al*. [87] uses only skeletons to extract physical features of the body. The extraction of skeleton information is mainly based on the computation of a few limb lengths and ratios (see Fig 2.15). In [86], besides using skeleton to extract physical information, applying point clouds converted from depth images for 3D body shape matching is also considered. In [34], a deep model is applied to classify the person point cloud sequences, in which feature extraction and classification are jointly modeled.

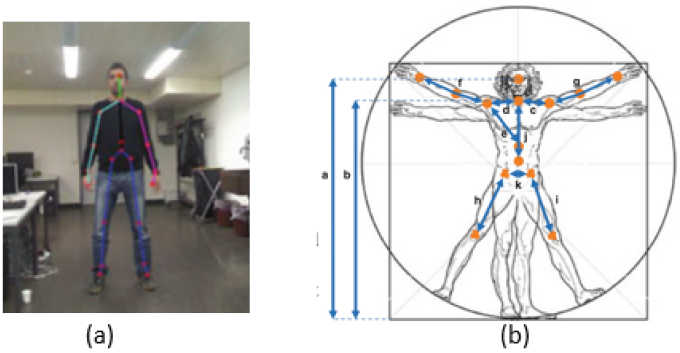


Fig. 2.15 Illustration of (a) skeleton extraction and (b) skeleton based physical features

Camera network topology

In the large-scale camera network, it is not efficient to perform re-identification only by using appearance-based methods, as they do not take the structure of the camera network into account. Therefore, they have to examine every possible person candidate in the video streams to re-identify a person. Instead of examining every person, we can restrict and reduce the search space by inferring the spatio-temporal relation between cameras. A person cannot be at two different locations at the same time and cannot travel a given distance in a time which corresponds to incoherent velocity, etc. Cai *et al*. [11] exploit these spatio-temporal constraints for positive/negative training sample collection. and person re-dientificaiton. In the case, the camera network topologie is unknow. In [42], motion trends of people and cars are used to establish correspondences across non-overlapping cameras for tracking. Makris *et al*. [79] proposed a topology inference method based on a simple event correlation model between cameras.

Group context

The fact that many people are wearing clothing with similar color and style, increase the ambiguity and uncertainty in the matching process. One possibility is to seek more holistic contextual constraints in addition to localized visual appearance of isolated individuals. In public scenes people often walk in groups, either with people they know or strangers. The availability of richer visual content in a group of people over space and time could provide vital contextual constraints for more accurate matching of individuals within the group. We will discuss this in detail in Chapter 5.

Re-ranking

Re-ranking methods consist in improving the re-identification results by using obtained ranking list.The intuition is that there should be a certain symmetry between the query and the highest ranked results, i.e. by using the results as a new query, the original query, in turn, should be highly ranked. The main advantage of many re-ranking methods is that they can be implemented without requiring additional training samples, and that they can be applied to any initial ranking result. A number of works utilize the k-nearest neighbors to explore similarity relationships to address the re-ranking problem. Chum *et al*. [20] proposed the average query expansion method, where a new query vector is obtained by averaging the vectors in the top-k returned results, and is used to re-query the database. Zhong *et al*. [142] ,instead of k-nearest neighbors, exploit the k-reciprocal neighbors to improve person re-identification.

2.3.5 Datasets

In the literature, numerous of datasets for person re-identification have been released over the last years. In this section we present the most used state-of-the-art datasets for person re-identification, categorizing them into tree classes and highlighting their challenging aspects. A summary of the common datasets is shown in the Table 2.1. Some example images from these datasets are shown in the appendix.

Datasets	# identities	# images	#views	Bounding box annotation
VIPeR	632	1264	2	manual
GRID	250+775 distractors	1275	2	manual
PRID2011(S)	400+534 distractors	1334	2	manual
CUHK01	971	3884	2	manual
CUHK03	1467	13164	10	manual/detector
Market1501	1501	32217	6	detector
DukeMTMTC-Reid	1812	36441	8	manual
PRID2011(V)	400+534 distractors	24541	2	manual
iLIDS	300	42495	2	manual
Mars	12611	1191003	6	detector+tracker

Table 2.1 Person re-identification dataset overview

Single-shot datasets

Datasets are qualified “single-shot” if, in gallery set in the test set of the dataset, each identity has one image for query and one true match image. The most commonly used single-shot datasets are VIPeR [33], GRID [74], PRID2011(S) [39], CUHK01 [60]. They are relatively

small, containing hundreds of identies. VIPeR, CUHK01 was captured in university campus scenes. PRID2011(S) was captured in street scenes. GRID was collected in a subway station. In VIPeR, CUHK01 and PRID2011(S), viewpoint changes are the main challenge since the images are captured under very different angles. PRID2011(s) and VIPeR have illumination changes between two cameras. GRID and PRID2011(S) have a number of distractor images (additional images that do not belong to any of the probes) in the gallery set.

Multi-shot datasets

In multi-shot datasets, the test sets are composed such mean that each identity has one or several images for the query and several true match images in gallery set. The most common multi-shot datasets are CUHK03 [61], Market1501 [137] and DukeMTMC-Reid [141]. All these three datasets are collected in campus scenes. They include each more than 1400 identities and more than 6 camera views. Besides the various viewpoint and pose, the large scale variation is the main challenge in these datasets. CUHK has two versions: one with manually labeled and one with automatically detected bounding boxes. Market-1501 is composed of the bounding boxes from a pedestrian detector and images from DukeMTMC-Reid are manually annotated.The datasets that are constituted by automatic detection of pedestrians usually produce misaligned bounding boxes and images, which poses a big challenge for re-identification.

Video datasets

The difference between multi-shot datasets and video datasets are that video datasets are composed of continuous image sequences and the multi-shot datasets contains several single images from different views but not continuous. The video datasets allow to exploit the temporal information. Common video datasets for person re-identification are ILIDS [118], PRID2011(V) [39] and MARS [136]. Images in ILIDS are captured in an airport. PRID2011(V) and MARS are respectively the video extension of PRID2011(S) and of Market-1501.

Evaluation measure

Cumulated Matching Characteristics (CMC) and mAP (mean Average Precision) are two most used evaluation measures for person re-identification. CMC evaluates the top n nearest images in the gallery set with respect to one probe image. If a correct match of a query image is at the k^{th} position $(k <= n)$, then this query is considered as success of rank n. mAP is the

mean value of average precision of all queries. The average precision is defined as the area under the Precision-Recall curve.

2.4 Conclusion

In this chapter, we first introduced deep learning, especially CNN which can learn a high-level abstraction feature extraction by the backpropagation algorithm. Compared to classical methods with "handcrafted" features, learned feature extractions can better adapt to the task. In the case of person re-identification, CNNs can learn more discriminative features to distinguish identities than manually designed features. Meanwhile, to learn such a feature extraction, a large amount of annotated data is necessary. Thus, on the relatively small dataset like VIPeR, GRID, classical methods combining "handcrafted" feature and Manahanobis distance learning can still give superior results and is computationally cheaper. But on large datasets like Market1501 or DukeMTMC-Reid, the CNN-based methods can outperform classical methods by a large margin.

In spite of the different conceptions of these two kinds of methods, we can still find many similar ideas on dealing with the challenges related to large appearance variations. Most of both CNN architectures and classical feature-based methods are based on stripe, patch or body part segmentations. We summarize the advantages and disadvantages of these different types of methods in Table 2.2.

Methods	Advantage	Disadvantage
Patch based	more discriminative local features	spatial information is not encoded, patch matching is computationally expensive
Stripe based	invariant to viewpoint and pose changes	sensitive to bounding box misalignment and background clutter
Part based	semantic and invariant to pose changes and misalignment	Body part segmentation error can affect results, eventually body part data or annotations are needed

Table 2.2 Comparison of different types of person re-identification methods.

Similarly, some loss function in CNN and metric learning method are based on the same metric constraint,Contrastive loss and PCCA metric [85], for example, both penalize the long distance between positive pairs and short distances between negative pairs. Triplet loss and PRDC metric [139] both penalize the case where a false match is closer than true match with respect to a reference or query instance.

As to methods based on other cues than the appearance, combining these methods with an appearance-based method can add more discriminative power for person re-identification, but usually some extra constraints or conditions need to be satisfied.

Most existing person re-identification approaches assume that, for a given query image, there is a true match in the gallery set. However, in real-world applications, this is not necessarily the case. Thus a threshold on the similarity score is required. This threshold can be very difficult to define or learn. The second limitation is that in real application for surveillance network, there are not that much annotated data. Therefore, the future research direction could be extending person re-identification to person verification. Moreover, unsupervised or semi-supervised learning should be paid more attention to. Training a model on an annotated dataset, then performing domain adaptation or transfer learning on available public datasets can be another way to be exploited for this issue.

Following the trend of the increase of the amount of surveillance data and given their large robustness and excellent performance in visual recognition task, in this thesis, we chose to focus on deep learning methods for the person re-identification task. On the one hand, we propose to use higher-level semantic information like attributes, body orientation and group context to overcome some difficulties of existing methods for re-identification. On the other hand, we improve deep learning methods by introducing some novel CNN architectures and loss functions.

Chapter 3

Pedestrian Attribute-assisted Person Re-identification

3.1 Introduction

Recently, visual attributes recieved much attention and have been successfully used for object recognition [25], action recognition [70] , face recognition [50] etc. Pedestrian attributes are defined as semantic mid-level descriptions of persons, such as gender, accessories, clothing and so on. The advantage of attributes is that they are more robust to visual changes related to the viewing angle, body pose or lighting for instance, and that they can be used for "zero-shot" identification (querying by an attribute-based description instead of an image). Since biometric features like faces are often not visible or of too low resolution to be helpful in surveillance, pedestrian attributes could be considered as soft-biometrics and provide helpful information for many surveillance applications like person detection [107], person retrieval [111], or abnormal event detection. For example, a description like "a male in a black shirt with a backpack" can be effectively used in person retrieval applications.

The main challenges for pedestrian attribute recognition are the large visual variation and large spatial shifts due to the descriptions being at a high semantic level. For instance, the same type of clothes (e.g. shorts) can have very divers appearances. The large spatial shifts with respect to the detected pedestrian bounding boxes are caused by different body poses and camera views, and a finer body part detection or segmentation is challenging in surveillance-type videos. Furthermore, in realistic settings, illumination changes and occlusion make the problem even more challenging.

In this chapter, we present a CNN-based pedestrian attribute assisted person re-identification framework. In the first step, the attribute learning is performed. In order to deal with the

large spatial shift of attributes, we propose to use a specific CNN architecture with 1D convolution layers operating on several horizontal parts of the input feature maps to learn different feature representations and model the displacements of different body parts. For an even larger spatial invariance, our approach additionally extracts LOMO features, which have been specifically designed for viewpoint-invariant pedestrian re-identification. These low-level handcrafted features are fused with the high-level learned CNN features at a late training and processing stage to get a more robust feature representation modelling the diverse appearances of attributes. Our experiments show that the proposed method improves the state of the art on pedestrian attribute recognition on three public benchmarks.

In the second step, the learned attribute embedding is used for person re-identification. The framework fuses two neural networks. One is our attribute recognition network pre-trained with attribute labels, the other is a CNN pre-trained with person identity labels. Then we integrate these two neural networks into a triplet architecture to learn the optimal fusion parameters to perform re-identification. To this end, an improved triplet loss with hard example selection is used. We experimentally show that the fusion leads to a better re-identification performance, and our approach achieves state-of-the-art results.

3.2 Related work to pedestrian attribute recognition

3.2.1 Attribute recognition

In the pioneering work of Vaquero *et al*. [111], the person image is parsed into regions, and each region is associated with a classifier based on Haar-like features and dominant colors. The attribute information is then used to index surveillance video streams. Layne *et al*. [52] annotated 15 attributes on the VIPeR dataset and proposed an approach to extract a 2784-dimensional low-level color and texture feature vector for each image and to train an SVM for each attribute. Zhu *et al*. [145], in their work, introduced the pedestrian attribute database APiS. Their method determines the upper and lower body regions according to the average image and extracts color and texture features (HSV, MB-LBP, HOG) in these two regions. Then, an Adaboost classifier is trained on these features to recognize attributes. The drawback of these approaches is that all attributes are treated independently. That is, the relation between different attributes is not taken into account.

Some later works try to overcome this limitation. Zhu *et al*. [143] proposed an interpolation model based on their Adaboost approach [145] learning an attribute interaction regressor. The final prediction is a weighted combination of the independent score and the interaction score. Deng *et al*. [21] constructed the pedestrian attribute dataset “Peta” and their approach

uses a Markov Random Field (MRF) to model the relation between attributes. The attributes are recognized by exploiting the context of neighbouring images on the MRF-based graph.

Some CNN models have been proposed for pedestrian recognition. For example, Li *et al*. [55] fine-tuned the CaffeNet (similar to AlexNet) trained on ImageNet to perform simple and multiple attribute recognition. Similarly, Sudowe *et al*. [105] proposed the Attribute Convolutional Net (ACN) which adds custom layers to Alexnet to jointly learn attributes. Zhu *et al*. [146] proposed to divide the pedestrian images into 15 overlapping parts where each part connects to several CNN pipelines with several convolution and pooling layers. They further pre-define connections between the parts and the attributes in the fully-connected layers to deal with the shift problem. Later they proposed an improved version [144], where a fully-connected layer is connected to all attributes in stead of using manually defined connections.

Recently, some approaches combining deep features and "hand-crafted" features have been proposed in different tasks like saliency detection [57], face recognition [75] and person re-identification [123]. These approaches implement a deep neural network framework embedded with low-level features. In our work, we exploit this "handcrafted" and deep feature combination in the attribute recognition context. Our method effectively fuses shift-invariant lower-level features with learned higher-level features to build a combined representation that is more robust to the large intra-class variation which is inherent in pedestrian images.

3.2.2 Attribute assisted person re-identification

Some works have used pedestrian attributes to assist with the re-identification task. Based on an attribute recognition SVM approach, Layne *et al*. [52] first proposed to use attributes as a mid-level representation for improving person re-identification. The final distance between two pedestrian images is computed as a weighted sum of low-level feature distance and attribute distance. Li *et al*. [54] embeds middle-level clothing attributes via a latent SVM framework for more robust person re-identification. The approach introduced by Khamis *et al*. [46] learns a discriminative projection to a joint appearance-attribute subspace in order to leverage the interaction between attributes and appearance for matching.

Other related methods use attributes with a CNN model to perform person re-identification, but the ways of integrating these attributes are different, for example using a weighted combination, simple concatenation, multi-task learning or attribute pre-training. Zhu *et al*. [146] recognize the attributes with deep neural networks then calculate a pedestrian distance by weighting the attribute distance and a low-level feature-based person appearance distance. McLaughlin *et al*. [83] propose to perform person re-identification and attribute recognition

in a multi-task learning. The proposed loss function is a weighted sum of the attribute and identification classification loss as well as a Siamese loss. They show that this multi-task joint learning improves the re-identification performance. Matsukawa *et al*. [82] propose to fine-tune the well-known Alexnet with attribute combination labels to increase the discriminative power. Further they concatenated the CNN embedding directly with LOMO features and used the metric learning method XQDA [64] to learn a feature space. Su *et al*. [103] proposed a three-stage procedure that pre-trains a CNN with attribute labels of an independent dataset, then fine-tunes the network with identity labels and finally fine-tunes the network with the learned attribute feature embedding on the combined dataset. The main difference of these approaches to ours is the way of making use of attributes to assist in the re-identification task. In summary, two CNN embeddings are learned based on attribute and identity annotation. Then, an improved triplet loss is used to learn the fusion. We will experimentally show the performance improvement brought by this fusion achieving state-of-art results on a public person re-identification benchmark.

3.3 Deep and low-level feature based Attribute Learning for Person Re-identification

In this section, we first describe our attribute recognition method and then introduce the attribute and identity-based re-identification framework.

3.3.1 CNN based pedestrian attribute recognition

Overall procedure

The architecture of the proposed attribute recognition approach is shown in Fig. 3.1. The framework consists of two branches. One branch is a CNN extracting higher-level discriminative features by several succeeding convolution and pooling operations that become specific to different body parts at a given stage (P3) in order to account for the possible displacements of pedestrians due to pose variations. Another branch extracts the viewpoint-invariant Local Maximal Occurrence (LOMO) features, a robust visual feature representation that has been specifically designed for viewpoint-invariant pedestrian attribute recognition and achieving state-of-the-art results [64] (cf. Section 3.3.1). The extracted LOMO features are then projected into a linear subspace using Principal Component Analysis (PCA). The aim of this step is two-fold: first, to reduce the dimension of the LOMO feature vector removing potential redundancies, and second, to balance the contribution of CNN features and LOMO

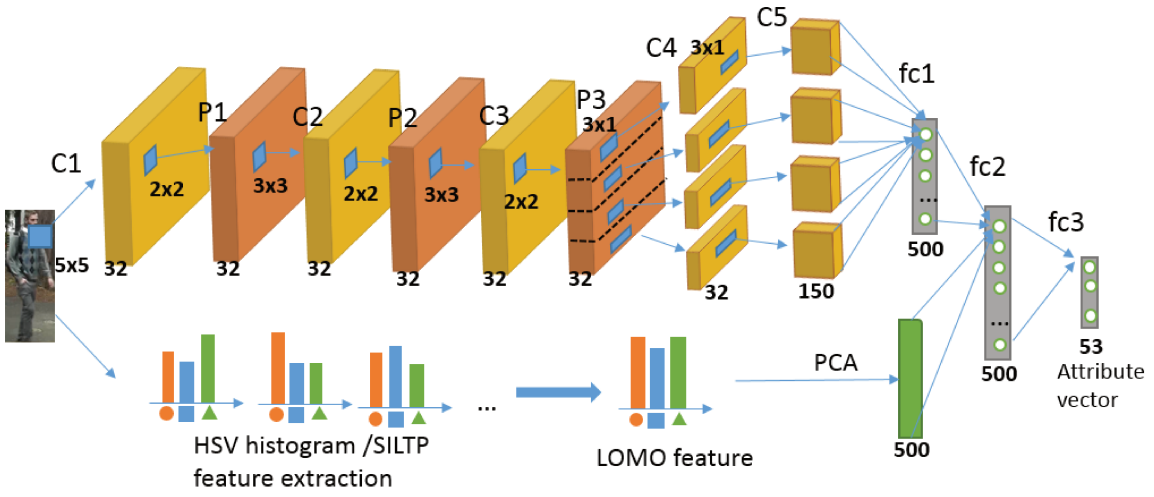


Fig. 3.1 Overview of attribute recognition approach

features in the succeeding fusion that combines information represented in the two feature vectors.

To carry out this fusion, the output vectors of the two branches are concatenated and given to a neural network of two fully-connected layers (fc2+fc3) effectively performing the final attribute classification. We will explain these steps in more detail in the following.

Part-based CNN

We first propose to extract deep feature hierarchies by a CNN model providing a higher level of abstraction and a larger discrimination power since the features are directly learned from data. As illustrated in Fig. 3.1, we use an Alexnet-like CNN architecture, *i.e.* 5 convolutional layers with maxpooling. We designed a relative small CNN rather than using directly the large models like Alexnet, VGG, since we want to keep the model "light" and add little extra-cost to the person re-identification task by using attributes. Since pedestrian bounding boxes are of very small size and small convolution kernels have less network parameters, we choose to use relatively small convolution/pooling kernels. Besides, the well-known VGG network [101] has shown that small convolution kernels can also be effective.

Therefore the size of the first convolution (C1) is set to 5×5 to have a larger perceptive field. And the two following (C2, C3) are of size 3×3. The kernel size of max-pooling (P1-P3) is 2×2, and the number of channels of convolution and pooling layers is 32 respectively. The resulting feature maps in the last pooling layer are divided vertically into 4 equal parts roughly corresponding to the regions of head, upper body, upper legs and lower legs. For each part, similar to [113], we use two layers (C4, C5) with 1D horizontal convolutions of

size 3×1 without zero-padding reducing the feature maps to single column vectors. These 1D convolutions allow to extract high-level discriminative patterns for different horizontal stripes of the input image. In the last convolution layer, the number of channels is increased to 150, and these feature maps are given to a fully-connected layer (fc1) to generate an output vector of dimension 500. All the convolution layers in our model are followed by batch normalization and ReLU activation functions.

LOMO extraction

Recently, pedestrian re-identification methods using LOMO features proposed by [64] have achieved state-of-the-art performance. We apply these low-level features on the related task of attribute recognition in order to extract relevant cues from pedestrian images and to complement the CNN features by providing a higher viewpoint invariance. In the LOMO feature extraction method, the Retinex algorithm is integrated to produce a colour image that is consistent with human perception. To construct the features, Scale-Invariant Local Ternary Patterns (SILTP) [65] and HSV histograms are extracted in the sliding windows at 3 different image scales. For all sliding windows on the same image line, only the maximal value of the local occurrence of each pattern among these sub-windows is retained. In that way, the resulting feature vector achieves a large invariance to view point changes and, at the same time, captures local region characteristics of a person. More details can be found in chapter 2. In our approach, as illustrated at the bottom of Fig. 3.1, we perform a dimensionality reduction projecting the extracted LOMO features of size 26 960 on a linear subspace of dimension 500, in order to facilitate the later fusion. The projection matrix is computed using PCA on the LOMO feature vectors computed on the training dataset.

Training

To train the parameters of the proposed CNN, the weights are initialised at random and updated using stochastic gradient descent minimising the global loss function (Eq. 3.1) on the given training set. Since most attributes are not mutually exclusive, i.e pedestrians can have several properties at the same time, the attribute recognition is a multi-label classification problem. Thus, the multi-label version of the sigmoid cross entropy is used as the overall loss function:

$$E = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}[w_l y_{il}\log(\sigma(x_{il})) + (1-y_{il})\log(1-\sigma(x_{il})]\,, \tag{3.1}$$

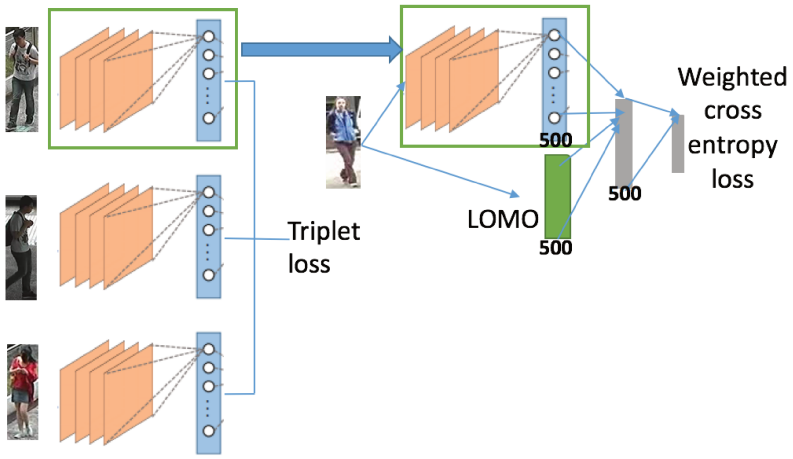


Fig. 3.2 Illustration of the transfer learning from a re-identification task to attribute recognition. *Left:* the (shared) weights of the triplet CNN are pre-trained in a weakly supervised manner for pedestrian re-identification using the triplet loss function. *Right:* the CNN weights are integrated in our attribute recognition framework and the whole neural network is fine-tuned using the weighted cross-entropy loss.

$$\text{with} \quad \sigma(x) = \frac{1}{1+exp(-x)} \,,$$

where L is the number of labels (attributes), N is the number of training examples, and y_{il}, x_{il} are respectively the l^{th} label and classifier output for the i^{th} image. Usually, in the training set, the two classes are highly unbalanced. That is, for most attributes, the positive label appears generally less frequently than the negative one. To handle this issue, we added a weight w to the loss function: $w = -log_2(p_l)$, where p_l is the positive proportion of attribute l in the dataset.

As we will show in our experiments, for smaller training dataset (like VIPeR). It is beneficial to pre-train the CNN with a (possibly larger) pedestrian re-identification dataset in a triplet architecture on the re-identification task. Since pedestrian attribute recognition and re-identification are two similar tasks, the visual features learned from re-identification can be useful for recognising attributes. Thus, after this pre-training, we transfer the re-identification knowledge to attribute recognition by fine-tuning the pre-trained convolution layers on the actual small attribute datasets. Figure 3.2 illustrates this transfer learning approach. The CNN branch of the framework is pre-trained with person re-identification data in a triplet architecture. Then we use the learned weights as the initialization of the network.

From the re-identification data, the network learns informative features that distinguish individuals, and the semantic attributes that we want to recognise can be considered as such identity features at a higher level. Therefore, this pre-learned knowledge can be effectively

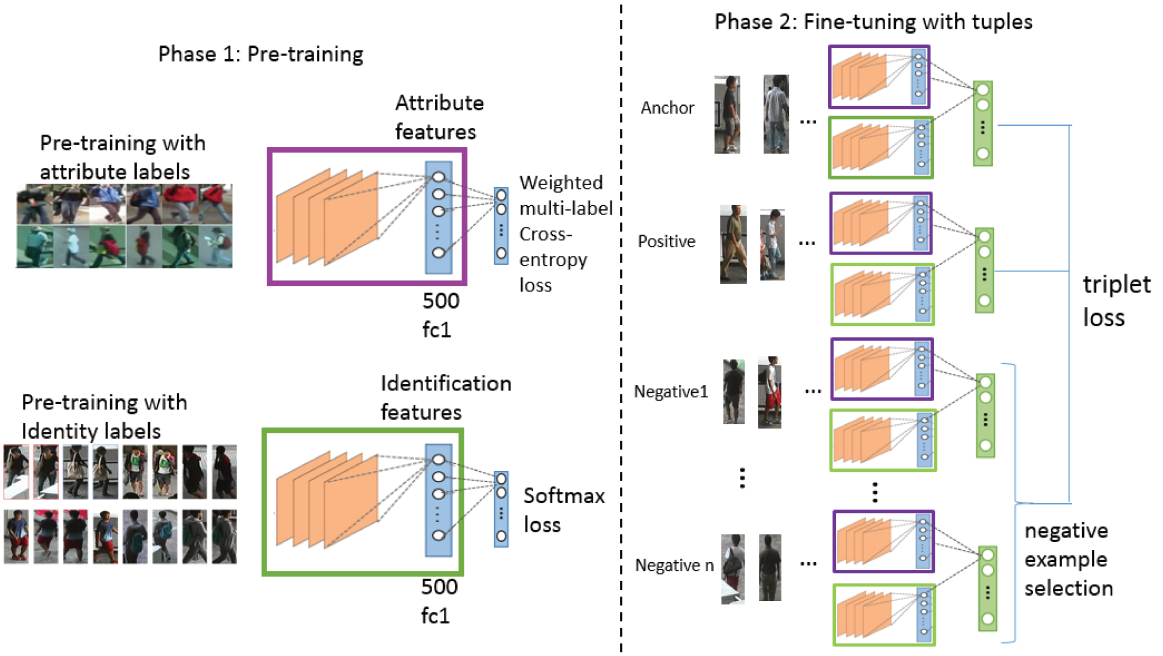


Fig. 3.3 Overall architecure of our re-identification method

transferred to this problem. In the next section, we will also show that, inversely, attribute information can support the re-identification task.

3.3.2 Person re-identification

We propose a new CNN-based approach for pedestrian re-identification, that effectively combines automatically learned visual features with semantic attributes. To this end, we make use of our attribute recognition neural network presented in Section 3.3.1. Attributes are important cues for a human to identify persons by appearance. The attribute learning consists in representing data instances by projecting them onto a basis set defined by domain-specific axes which are semantically meaningful to humans. Compared to features that are directly learned form appearance, semantic attributes are more consistent for the same person and are more robust to the different variations. Since the attribute information adds additional constraints to person identity consistency, i.e. appearance consistency and attribute consistency, the combination with the attribute CNN embedding which encodes these attribute constraints improves the person re-identification performance, as we will show experimentally in Section 3.4.2.

However, only using the attributes is not discriminant enough to perform re-identification. It is very probable that pedestrians have the same or similar attributes. Then, the idea is to combine the attribute embedding with identification embedding. Since the identification task is very close to the person re-identification, the identification embedding can also be helpful to discriminant persons. But the identification learned on training set has a limited generalization ability to different person on test set. So a triplet based fine-tuning is necessary to adapt to the re-identification task. The advantage of using the identification task as a pre-training step is that it is more efficient to train the CNN by using the cross-entropy loss than using directly the triplet loss. Since the supervised identification constraint of the loss is stronger than a relative distance constraint. Fine-tuning the CNN on the re-identification task based on pre-trained identification and attribute embeddings can be more effective than re-identification training from scratch.

The overall framework is shown in Fig. 3.3. The framework is composed of two neural networks that are pre-trained. The first is a CNN that is trained in a supervised way to classify identities on a separate training set. Then we remove the output classification layers of the network and keep the other parts of the network which are related to feature selection. The second part is our attribute recognition network that is trained as described in section 3.3.1. After training, we also remove the output layers and keep all the other layers up to the first fully-connected layer (fc1).

The output vectors from the hidden layers of the two CNNs are concatenated and they represent high-level features related to attributes and pedestrian identities respectively. As we will show experimentally, the information extracted by the two CNNs is complementary, thus using both leads to an overall performance improvement. In order to combine the extracted features effectively, we propose to integrate both output vectors in a new neural network that automatically learns these fusion parameters on the re-identification task in a triplet architecture. this leads to a fully neural architecture that can be trained and fine-tuned as a whole to maximise the re-identification performance. We will explain these steps in more detail in the following.

Supervised identification CNN

For the identification, we employ a similar network architecture as the attribute recognition network, consisting of 5 convolutional layers and 4 max-pooling layers. The details are presented in Table 3.1. The first convolutional layer has a kernel size of 5x5 and the following 3 convolutional layers have a kernel size of 3×3. The last one has a kernel size of 3×1 without zero-padding increasing the number of channels but reducing their size to a single column. At the end, there are two fully-connected layers. All max pooling layers have a

layer	type	filter size	padding	output size
C1	Convolution	5×5	yes	128×48×32
P1	Max-Pooling	2×2	-	64×24×32
C2	Convolution	3×3	yes	64×24×32
P2	Max-Pooling	2×2	-	32×12×32
C3	Convolution	3×3	yes	32×12×64
P3	Max-Pooling	2×2	-	16×6×64
C4	Convolution	3×3	yes	16×6×128
P5	Max-Pooling	2×2	-	8×3×128
C5	Convolution	3×1	no	8×1×400
fc1	Fully-connected	-	-	500
fc2	Fully-connected	-	-	N

Table 3.1 Identification network parameters.

kernel size of 2×2. The output of the network is a vector of N dimensions with N being the number of identities in the training set. Batch normalization and ReLU activation function are applied after the convolutional layers and fully connected layers. As mentioned earlier, this CNN is pre-trained in a supervised way, using images and identity labels from a separate training dataset. To this end, we minimise the following softmax cross-entropy loss on the given classification task:

$$P(y_j = 1|x) = \frac{e^{W_j^T x + b_j}}{\sum_{k=1}^{N} e^{W_k^T x + b_k}} \tag{3.2}$$

$$E_{identification} = -\sum_{k=1}^{N} y_k log(P(y_k = 1|x)) \tag{3.3}$$

Where N is the number of identities. y is the one-hot-coded identity label. x is the input of the last fully connected layer. W and b are weight and bias term of the last fully connected layer. $P(y_j = 1|x)$ is the probability predicted that the input x corresponds to identity j. The intuition of this is that this learned feature representation can be used for learning similarities between arbitrary pedestrian images and thus be transferred to the task of re-identification.

Fusion by Triplet architecture

The pre-trained attribute CNN and identification CNN are combined and trained in a triplet architecture. Here, we propose to use an improved triplet loss with hard example selection to learn the optimal fusion of the two types of features. The fc1 layer of the attribute network and the fc1 layer of the identification network are normalized and concatenated, and another fully-connected layer which allows to merge attribute and identification features is added.

Unlike with classic triplet loss, a (n+2)-tuple of images instead of a triplet is projected into the feature space. The tuple includes one anchor image, one positive image and n negative images. Training enforces that the projection of the positive example is placed closer to the anchor than the projection of the closest negative example among n negative examples. This constraint is defined as following:

$$min(\|f(a)-f(n^j)\|_2^2)-\|f(a)-f(p)\|_2^2> m \tag{3.4}$$

Similarly to the distance learning approach “Top-push" proposed by [129] , hard example mining in [1, 96] or moderate positive example mining in [100], the idea is finding the appropriate example to update the model. The negative example that is closest to the anchor is considered the hardest example and having the highest potential for improvement. The network is thus updated efficiently by pushing the hardest example further away from the anchor. The intuition is that if the positive example is ranked in front of the hardest negative example then the positive example is ranked first, which is our goal. In classic triplet loss, a big part of triplets does not violate the triplet constraint (c.f Eq 3.4). These triplets are useless for learning. The selection among n negative examples reduces the number of unused training data and can make the training more efficient.

To further enhance the loss function, as an extention of [17], we add a term including the distance between the anchor example and the positive example. The loss function is defined as follows:

$$\begin{aligned} E_{min-triplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(a_i)-f(p_i)\|_2^2 \\ -min(\|f(a_i^j)-f(n_i^j)\|_2^2)+m,0)+\alpha\|f(a_i)-f(p_i)\|_2] \end{aligned} \tag{3.5}$$

The first part of the loss is a comparison of two distances which defines a relative relationship in the feature space. The second part corresponds to an absolute distance in feature space. Combining these two constraints leads to a more efficient learning of the resulting manifold that better represents the semantic similarities.

Using this loss function, we train the additional fully-connected layer for the fusion, and, at the same time we fine-tune the other parts of the network, i.e. the weights are updated at a lower rate. Since pedestrian attributes are difficult to annotate, especially for large re-identification dataset. Unlike other approaches[83, 54, 46, 52], the advantage of our method is that the attributes do not need to be annotated on the re-identification dataset. We can make use of a separate attributes dataset with annotated attributes and transfer this information to a re-identification dataset by fine-tuning.

Fig. 3.4 Some example images from pedestrian attribute datasets.

3.4 Experiments

In this section, the proposed attribute recognition methods are evaluated on the VIPeR [32] dataset with the annotation from [51], PETA dataset[54] and the APiS dataset [145] (see Fig. 4). Finally, we test our proposed CNN architecture for person re-identification on the CUHK03 dataset [61].

3.4.1 Attribute recognition experiments

Datasets

We evaluated our approach on three public benchmarks: PETA, APiS and VIPeR (see Fig. 3.4).

- The **PETA dataset** [21] is a large pedestrian attribute dataset which contains 19 000 images from several heterogeneous datasets. 61 binary attributes and 4 multi-class attributes are annotated. In our attribute recognition evaluation, we follow the experimental protocol of [21, 55]: dividing the dataset randomly in three parts: 9 500 for training, 1 900 for validation and 7 600 for testing. Since different approaches[21, 55, 144] have

been evaluated on different subsets of attributes, in our experiment we use the union of all these subsets, i.e. 53 attributes. The used attribute names are shown in Appendix.

- The **APiS dataset** [145] contains 3 661 images collected from surveillance and natural scenarios. 11 binary attributes are annotated such as male/female, shirt, backpack, long/short hair. All the images are already resized to 128x48 pixels by bilinear interpolation. We followed the experimental setting of [145]. A 5-fold cross-validation is performed, and the final result is the average of the five tests.

- The **VIPeR dataset** [32] contains 632 pedestrians in an outdoor environment, each having 2 images from 2 different view points. 21 attributes are annotated by [51]. Each dataset is divided into two equal-size non-overlapping parts for training and testing (images from the same person are not separated). We repeat the process 10 times and report the average result.

During training, we perform data augmentation by randomly flipping and shifting the images slightly.All the inputs are resized to a resolution 128×48. Since images from some subsets of Peta, VIPeR and APIS datasets have already been resized to this size and the other images from Peta are almost in this range. We suppose that increasing the image size by interrelation does not bring any new information to improve the performance.

Parameters setting

All weights of the neural network are initialised from a Gaussian distribution with 0 mean and 0.01 standard deviation, and the biases are set to 0. The learning rate is set to 0.01. We used dropout [102] for the fully-connected layers with a rate of 0.6.

For tests on APiS and VIPeR, a dimension of 500 is used for the layers fc1 and fc2 as well as for the PCA projected LOMO feature, and the batch size is set to 50. Since the PETA dataset has more data, for tests on PETA, the fc1 layer, fc2 layer and PCA projected LOMO features are set to 1000 dimension and the batch size is set to 100.

The neural network is learned with random initialisation for tests on PETA and APiS. Since for VIPeR we have only 632 training image, the network is pre-trained with triplet loss on CUHK03 dataset[61] which contains 13164 images of 1 360 pedestrians. Then, the CNN part is fine-tuned on VIPeR with a lower learning rate (0.0005).

Evaluation measure

The test protocol of PETA [21] proposes to use the attribute classification accuracy as evaluation measure. The APiS protocol [145] uses the average recall at a False Positive

Rate (FPR) of 0.1 and the Area Under Curve (AUC) of the average Receiver Operating Characteristics (ROC) curve as performance measures. As mentioned in [144], accuracy is not sufficient to evaluate the classification performance on unbalanced attributes. In our experiments, we thus use all these three measures to evaluate our approach.

Comparison with different variants

We first evaluated the effectiveness of different components: the 1D horizontal convolution layers, body part division, the feature fusion and person re-identification pre-training. The comparisons among different variants of the method using Peta and VIPeR datasets are respectively shown in Tables 3.2 and 3.3.

Our baseline is a CNN with 3 consecutive convolution and max-pooling layers (C1-P3) and a multilayer perceptron using LOMO features of different PCA output dimensions as input. Then we implemented different variants of the proposed method: the baseline with 2 layers of 3×3 convolution or 2 layers of 3×1 convolution, CNNs with and without body part division, and CNNs with and without LOMO feature fusion. The 1D convolutions have less parameters and sligntly improves the results compared to 2D convolutions. The body part division improves more than 1% point on the recall score. The fusion of deep features and LOMO features improves about 1% point on accuracy and 2% points on recall score. Comparing Tables 3.2 and 3.3, we can note that LOMO features are more performant on the VIPeR dataset and the deep features are more performant on the PETA datasets, since the images in VIPeR undergo extreme viewpoint changes, which LOMO is specifically designed to deal with. The Peta dataset actually comprises several datasets, where various clothing appearances becomes the major aspect of variation. The rich learned feature representation of the CNN is more robust in this case. Finally, the fusion increases the overall recall and accuracy on both datasets. We can conclude that these two kinds of features are complementary and the fusion make the framework more robust.

In Table 3.3, we can see that the pre-training on person re-identification data increases the accuracy by 1.4% points , the recall by 4.4% points and the AUC score by 3.4% points. This shows that the capacity to discriminate people learned from person re-identification can assist the attribute learning. In Section 3.4.2, we further show that, inversely, the attributes can assist and improve the person re-identification task.

	Accuracy	Recall@FPR=0.1	AUC
LOMO (dim 500)	88.7	72.5	89.8
LOMO (dim 1000)	89.8	73.7	90.3
baseline	89.7	76.2	92.0
baseline + 2D conv	90.0	76.9	92.2
baseline + 1D conv	90.5	77.3	92.1
baseline + part-based 1D conv	90.8	78.7	92.3
baseline + 1D conv + LOMO (dim 1000)	91.5	79.4	91.7
baseline + part-based 1D conv + LOMO (dim 1000)	**91.7**	**81.3**	**93.0**

Table 3.2 Attribute recognition result comparison of different variants of our approach on PETA.

Variants	Accuracy	Recall@FPR=0.2	AUC
non pretrained CNN	81.0	61.5	75.9
pretrained CNN	82.4	65.9	79.3
LOMO	83.1	68.2	**81.0**
pretrained CNN + LOMO	**83.9**	**69.6**	80.9

Table 3.3 Attribute recognition result comparison of different variants of our approach on VIPeR.

Comparison with the state-of-the-art methods

The comparison with the state of the art on PETA is shown in Table 3.4. In the literature, there are two evaluation settings for the PETA dataset with 35 and 45 attributes respectively. Table 3.4 shows the results on the 27 attributes that they have in common in order to compare all methods. We also display the average results for 35 and 45 attributes. A more exhaustive result comparison of all 53 attributes is shown in Appendix. Our method outperforms the state-of-the-art approach mlcnn on the 27 attributes by 3.4%, 14.3%, 6% points for average accuracy, recall and AUC respectively and by 3.5%, 15%, 6.1% points on the 45 attributes. It also outperforms the DeepMar method by 9% points on accuracy.

Moreover, our approach achieves a better score on almost all individual attributes. This superior performance comes from, the fusion of the viewpoint-invariant LOMO features and the rich deep feature representation, on the one hand, and from a better generalisation of our architecture and training compared to the mlcnn approach, on the other hand. We performed a simple test by removing the "dropout" mechanism from the training. Dropout is a form of regularisation and usually improves the generalisation capacity. On Peta, without dropout, the recall dropped from 79.9% to around 71%, which still above the state of the art.

attribute	Accuracy Rate (%)				Recall@FPR=0.1		AUC	
	MRFr2[21]	DeepMar[55]	mlcnn[144]	ours	mlcnn[144]	ours	mlcnn[144]	ours
personalLess30	**86.8**	85.8	81.1	86.0	63.8	**80.8**	88.5	**93.8**
personalLess45	83.1	81.8	79.9	**84.7**	59.4	**74.9**	84.6	**91.9**
personalLess60	80.1	86.3	92.8	**95.4**	70.2	**83.0**	87.7	**92.8**
personalLarger60	93.8	94.8	97.6	**98.9**	90.7	**94.6**	94.9	**96.8**
carryingBackpack	70.5	82.6	84.3	**85.5**	58.4	**70.2**	85.2	**91.9**
carryingOther	73.0	77.3	80.9	**85.7**	46.9	**65.1**	77.7	**88.4**
lowerBodyCasual	78.2	84.9	90.5	**92.1**	56.2	**76.1**	87.5	**93.1**
upperBodyCasual	78.1	84.4	89.3	**91.2**	62.1	**74.2**	87.2	**92.5**
lowerBodyFormal	79.0	85.2	90.9	**93.3**	72.5	**82.8**	87.8	**92.7**
upperBodyFormal	78.7	85.1	91.1	**93.4**	70.5	**83.4**	87.6	**92.9**
accessoryHat	90.4	91.8	96.1	**97.5**	86.1	**89.9**	92.6	**95**
upperBodyJacket	72.2	79.2	92.3	**94.7**	53.4	**77.4**	81.0	**92.1**
lowerBodyJeans	81.0	85.7	83.1	**87.6**	67.6	**83.2**	87.7	**94.5**
footwearLeatherShoes	87.2	87.3	85.3	**90.2**	72.3	**87.8**	89.8	**95.7**
hairLong	80.1	88.9	88.1	**91.3**	76.5	**88.3**	90.6	**95.6**
personalMale	86.5	**89.9**	84.3	88.9	74.8	**87.0**	91.7	**95.8**
carryingMessengerBag	78.3	82.0	79.6	**84.5**	58.3	**70.7**	82.0	**89.8**
accessoryMuffler	93.7	96.1	97.2	**98.8**	88.4	**93.6**	94.5	**96.2**
accessoryNothing	82.7	85.8	86.1	**89.0**	52.6	**71.5**	86.1	**92.1**
carryingNothing	76.5	83.1	80.1	**84.5**	55.2	**71.8**	83.1	**91.3**
carryingPlasticBags	81.3	87.0	93.5	**96.6**	67.3	**83.6**	86.0	**92.2**
footwearShoes	78.4	80.0	75.8	**80.8**	52.8	**68.3**	81.6	**89.4**
upperBodyShortSleeve	75.8	87.5	88.1	**90.7**	69.2	**86.2**	89.2	**94.5**
footwearSneaker	75.0	78.7	81.8	**85.7**	52.0	**73.0**	83.2	**92.0**
lowerBodyTrousers	82.2	84.3	76.3	**83.4**	56.2	**75.2**	84.2	**92.0**
upperBodyTshirt	71.4	83.0	90.6	**93.3**	63.5	**82.7**	88.7	**92.8**
upperBodyOther	**87.3**	86.1	82.0	86.2	73.2	**80.8**	88.5	**93.5**
27 attributes average	80.8	85.4	86.6	**90.0**	65.6	**79.9**	87.0	**93.0**
35 attr in [21, 55] average	75.6	82.6		**91.7**		78.9		**92.0**
45 attr in [144] average			87.2	**90.7**	67.3	**82.3**	87.7	**93.8**
53 attributes average				91.7		81.3		93.0

Table 3.4 Attribute recognition results on PETA (in %).

attribute	Accuracy	Recall@FPR=0.1			AUC			
	ours	fusion[145]	interact[143]	ours	fusion[145]	interact[143]	DeepMar[55]	ours
long jeans	93.5	89.9	89.2	**93.8**	96.1	96.2	96.5	**97.4**
long pants	94.2	78.7	80.6	**93.3**	92.5	93.9	**97.1**	**97.1**
M-S pants	93.7	76.7	85.1	**90.0**	92.4	92.8	95.5	**96.0**
shirt	88.4	68.2	**74.5**	65.5	83.9	83.9	**88.0**	87.3
skirt	95.6	58.3	61.3	**80.5**	90.0	**91.2**	91.0	90.5
T-shirt	79.6	56.2	56.5	**66.3**	85.4	85.5	**90.6**	88.7
gender	81.6	55.2	56.5	**65.1**	85.5	86.1	**90.0**	88.1
long hair	92.3	55.2	58.3	**68.9**	85.2	86.1	86.2	**88.1**
back bag	93.1	54.6	54.8	**61.2**	83.6	83.6	**86.6**	85.2
hand carrying	87.7	52.1	52.1	**60.6**	81.8	81.8	**84.3**	83.9
S-S bag	82.8	38.5	42.9	**54.0**	77.3	78.3	**83.7**	82.9
average	89.3	62.1	64.7	**72.7**	86.7	87.2	**90.0**	89.5

Table 3.5 Attribute recognition results on APiS (in %).

The results on the APiS dataset are shown in Table 3.5. Our method outperforms the Adaboost approach with fusion features and interaction models by a margin of 6% and 2.3% points respectively in recall at FPR=0.1 and AUC. The Adaboost fusion and interaction methods use simple low-level features like color histograms, LBP features. The improvement of our approach is mainly due to the richer feature presentation of the CNN and the horizontal local maximum extraction mechanism of LOMO. For the AUC, DeepMar achieves a slightly better result (0.5% points) which could be explained by its pre-training on the large ImageNet dataset.

Finally, the results on the VIPeR dataset are shown in Table 3.6. Our approach achieves a 9.8% point improvement in accuracy and 4.1% points on recall at FPR=0.2 compared to the CNN-based state-of-the-art approach mlcnn-p. For most of the attributes, our method obtains a better score.

In summary, our approach outperforms the state-of-the-art (including CNN-based methods) on two datasets and is on par with the best method on the third one. This demonstrates the robustness of the combined feature representation w.r.t to the high intra-class variation and the discriminative power of the proposed part-based CNN architecture. In addition, some attribute recognition examples results are shown in Fig. 3.5.

attribute	Accuracy			Recall@FPR=0.2			AUC
	svm[52]	mlcnn-p[146]	ours	svm[52]	mlcnn-p[146]	ours	ours
redshirt	85.5	91.9	**94.4**	88.4	88.9	**95.9**	95.2
blueshirt	73.0	69.1	**91.5**	60.8	70.8	**75.5**	83.1
lightshirt	83.7	83.0	**84.4**	87.8	85.3	**88.2**	91.7
darkshirt	**84.2**	82.3	83.3	**87.5**	85.8	86.1	90.9
greenshirt	71.4	75.9	**96.2**	54.3	69.4	**84.6**	88.7
nocoat	70.6	71.3	**74.2**	59.3	57.2	**65.4**	80.4
notlightdarkjean	70.3	90.7	**96.7**	57.2	78.6	**80.0**	86.0
darkbottoms	75.7	78.4	**78.9**	70.2	**76.2**	74.9	85.7
lightbottoms	74.7	76.4	**76.5**	69.5	**73.3**	72.3	83.6
hassatchel	47.8	57.8	**70.9**	22.0	31.7	**39.1**	64.8
barelegs	75.6	84.1	**92.2**	68.7	85.4	**92.2**	92.8
shorts	70.4	81.7	**92.3**	59.8	82.9	**87.3**	88.6
jeans	76.4	77.5	**80.6**	72.7	74.7	**81.7**	87.6
male	66.5	69.6	**74.7**	48.2	57.2	**67.9**	82.1
skirt	63.6	78.1	**94.3**	40.7	60.7	**61.3**	72.8
patterned	46.9	57.9	**90**	26.3	41.0	**49.9**	68.1
midhair	64.1	**76.1**	75.2	43.0	**63.5**	54.1	73.1
darkhair	63.9	**73.1**	67.5	39.6	**58.4**	49.7	71.9
hashandbagcarrierbag	45.3	42.0	**90.9**	17.4	18.5	**27.5**	55.1
hasbackpack	67.5	64.9	**72.7**	47.9	49.9	**57.4**	76.3
average	68.9	74.1	**83.9**	56.1	65.5	**69.6**	80.9

Table 3.6 Attribute recognition results on VIPeR (in %).

3.4.2 Re-identification experiments

Dataset

The CUHK03 dataset [61] includes 13164 images of 1 360 pedestrians and is one of the largest publicly available person re-identification dataset. Each person is taken from two different views. There are two settings labelled with human-annotated bounding boxes and the more challenging detected with automatically generated bounding boxes. In this experiment, we use the latter as this is closer to real-world scenarios. There are 100 identities for test and the rest for training and validation, with 20 training/test splits (provided by [61]).

We use one camera view as the probe set, and the other as the gallery set. For the gallery, we randomly sample one image for each identity. For the probe set, we use all the images, computing the CMC curve for each of them, and then average over them. This evaluation process is repeated for 20 times and the mean value is reported as the final result.

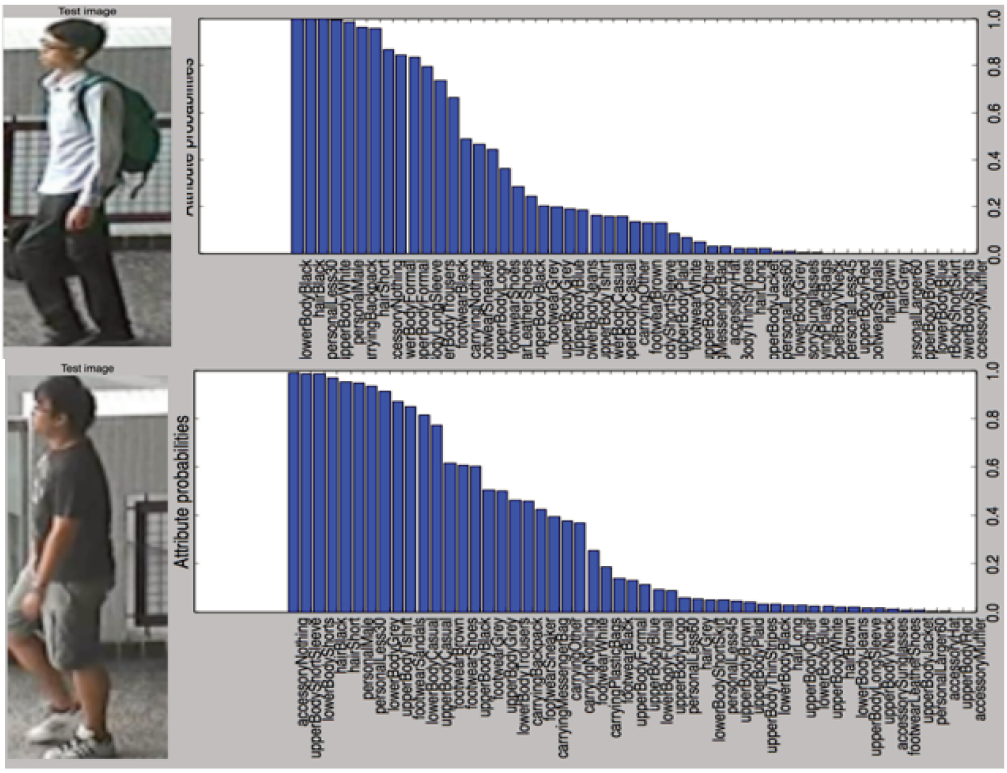


Fig. 3.5 Some attribute recognition result examples.

Training setting

The training is performed in two stages. In the first step, we pre-train the two subnets of the framework. Since the CUHK03 dataset does not have attributes annotations, the attribute network is pre-trained on the PETA dataset as described in section 3.4.1. The identification network is pre-trained with the CUHK03 dataset with 1160 identities in the training set. In the second stage, we remove the output layers of the two subnetworks, and then add a fully-connected layer for the fusion. We train the new fusion layer and fine-tune the rest of the network with a lower learning rate.

For the pre-training of the identity network, the learning rate is set to 0.005 and the batch size is set to 100. For the fusion phase, the initial learning rate is set 0.005, and we fine-tune the other part with initial learning rate of 0.0005. The learning rate is then reduced by a factor of 0.7 each 2000 iterations. Much smaller learning rate is applied for the previous layers since the two subnets are pre-trained and there is no need for a large step of update. The weights are initialised from zero-mean Gaussian distribution with a standard deviation of 0.01. We used dropout[102] for the fully-connected layers with a rate of 0.7. We generated 50 tuples in each iteration. In each tuple, we randomly select one reference image and one positive image from the same person but from a different camera view, and 5 negatives

Method	rank=1	rank =5	rank =10
FPNN [61]	19.9	49.3	64.7
Convnet [1]	45.0	75.3	83.4
LOMO+XQDA [64]	46.3	78.9	88.6
SS-SVM [132]	51.2	80.8	89.6
SI-CI [115]	52.2	84.3	92.3
DNS[131]	57.3	80.1	88.3
S-lSTM [113]	57.3	80.1	88.3
S-CNN SQ [112]	61.8	80.9	88.3
CAN[69]	63.1	82.9	88.2
ours Identity only	59.7	86.1	93.3
ours fusion Id&Attr	**65.0**	**90.3**	**95.1**

Table 3.7 Re-identification result on CUHK03 (“detected”).

images from different persons. The learning parameters are chosen by the performance on the validation set. In the loss function, the coefficient for the absolute part α is set to 0.02 as in [17] and the margin is set to the default value 1.

In both training phases, we perform data augmentation by randomly flipping the images and by cropping the center regions with random perturbation. All the inputs are resized to a resolution of 128×48.

Results

We compare our proposed re-identification approach on the CUHK03(detected) dataset with state-of-art methods Our method achieves the best results on rank 1, rank 5 and rank 10 accuracies. The hard example selection helps the training to converge slightly faster. Fusing the identity with the attributes can improve the results by 5.3, 4.2 points on rank 1 and rank 5. This shows that the identity and attribute information are complementary for the re-identification task and the robustness of the attribute features.

3.5 Conclusion

In this chapter, we have proposed a pedestrian attribute-assisted person re-identification framework. In our approach, the attribute learning is performed by merging low-level and high-level features learned by a body part-based CNN. By fusing these two kinds of features, the resulting model is robust to spatial variations due to pose or view point changes and incorporates rich feature representations that are able to model the divers appearance of pedestrian attributes. Our attribute recognition model outperforms the state of the art on

three public benchmarks. Further, to improve person re-identification, our model uses an improved triplet loss to fuse pedestrian identities and an attribute embedding. We have shown that making use of attributes enhances the re-identification performance. Our final re-identification method achieves the state-of-the-art result on the challenging CUHK03 benchmark.

Chapter 4

Person Re-identification with a Body Orientation-Specific Convolutional Neural Network

4.1 Introduction

The main difficulty of person re-identification is that the pedestrian appearance can be very different with different body orientations under different viewpoints, *i.e*. images of the same person can look quite different and images of different persons can look very similar.

Most existing approaches consider that pedestrian images come from a single domain. The viewpoint-invariant feature representations are either designed "manually" or learned automatically by a deep neural network. Though, re-identification can be considered as a multi-domain problem, *i.e*. pedestrians with the same body orientation have similar silhouettes and those with different body orientations have dissimilar appearance. Some metric learning approaches, for example, learn to transfer the feature space from one camera to another. But this requires a model for all the combination of cameras. Some other metric learning methods learn to transfer the different view-specific feature spaces to a common subspace where features are discriminative. This addresses the lighting and background variations, but it cannot be generalised to new camera views, and pedestrian images still have variations from different body orientations even if they come from the same camera.

To tackle this issue, we use a multi-task deep Convolutional Neural Network (CNN) to perform body orientation regression in a gating branch, and in another branch separate orientation-specific layers are learned as local experts. The combined orientation-specific

CNN feature representations are used for the person re-identification task. Our main contributions are:

- a mixture-of-expert deep CNN to model the multi-domain pedestrian images for person re-identification. We show that learning and combining different feature embeddings of different orientations improves the re-identification performance,
- a novel multi-task CNN framework with combined person orientation estimation and re-identification, where the estimated body orientation is used to steer the orientation specific mixture of experts for re-identification,
- an experimental evaluation showing that our approach outperforms most state-of-the-art methods on the CUHK01 and Market-1501 datasets.

4.2 Related Work to Orientation based Person Re-identification

Most existing methods for person re-identification focus on developing a robust representation to handle the variations of view. Some methods take into account the view as extra information. For example, Ma *et al*. [78] divide the data according to the additional camera position information and learn a specific distance metric for each camera pair. Lisanti *et al*. [68] proposed to apply Kernel Canonical Correlation Analysis which finds a common subspace between the feature space from disjoint cameras. Yi *et al*. [128] proposed to apply a Siamese CNN to person re-identification. Similar to [68], the weights of two subnetworks are not shared to learn a camera view projection to a common feature space. In these approaches, camera information is used but the body orientation which is only partly due to different camera views is not modelled. That is, in the same camera view, pedestrians can exhibit different orientations and thus largely different appearances in the resulting images.

In order to solve this issue, Bak *et al*. [5] perform an orientation-driven feature weighting and the body orientation is calculated according to the walking trajectory. some other approaches [114, 95] deal with the orientation variations of pedestrian images by using Mixture of Experts. The expert neural networks map the input to the output, while a gating network produces a probability distribution over all experts' final predictions. Verma *et al*. [114] applied an orientation-based mixture of experts to the pedestrian detection problem. Sarfraz *et al*. [95] proposed to learn the orientation sensitive units in a deep neural network to perform attribute recognition. Garcia *et al*. [28] used orientations estimated by a Kalman filter and then trained two SVM classifiers for pedestrian images matching with respectively similar orientations and dissimilar orientations. And the approach of Li *et al*. [59] learns a

mixture of experts, where samples were softly distributed into different experts via a gating function according to the viewpoint similarity.

Sharing the idea of mixture of experts, we propose to build a multi-domain representation in different orientations with deep convolutional neural networks. Intuitively, an orientation-specific model should have a better generalization ability than a camera view-specific model, since we cannot incorporate all possible surveillance camera views. Further, instead of using discrete orientations for the gating activation function, in our method, we use a regressor to estimate an accurate and continuous body orientation. This allows to continuously weight different expert models for re-identification and also avoids combining contradictory orientations.

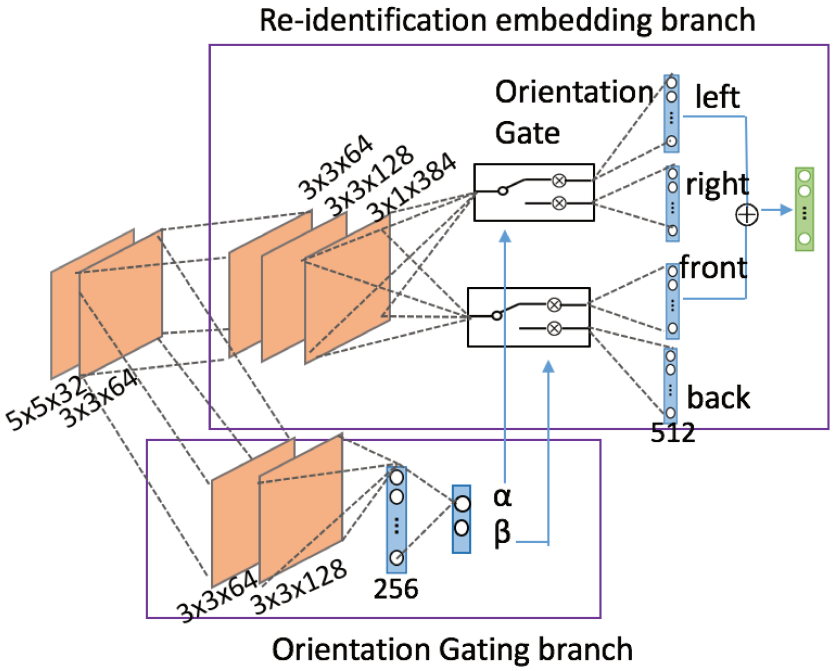


Fig. 4.1 Overview of the OSCNN architecture.

4.3 Proposed method

The overall procedure of our re-identification approach OSCNN is shown in Fig. 4.1. The network contains an orientation gating branch and a re-identification branch consisting of 4 feature embeddings regarding the 4 main orientations: left, right, frontal and back. The final output feature representation is a linear combination of the four expert outputs and is steered by an orientation gate unit which is a function of the estimated orientation.

4.3.1 OSCNN architecture

The proposed neural network architecture consists of two convolution layers shared between an orientation gating branch and a re-identification feature embedding branch. As CNN used in Chapter 3, we propose an Alexnet-like architecture and we apply the same kernel size and the number of layer setting as previously. In the re-identification branch, there are 3 further convolution layers followed by 4 separate, parallel fully-connected layers (left,right,front,back) of 512 dimensions, each one corresponding to a local expert. Thus, our network learns different projections from different orientation domains to a common feature space, as shown Fig. 4.2.

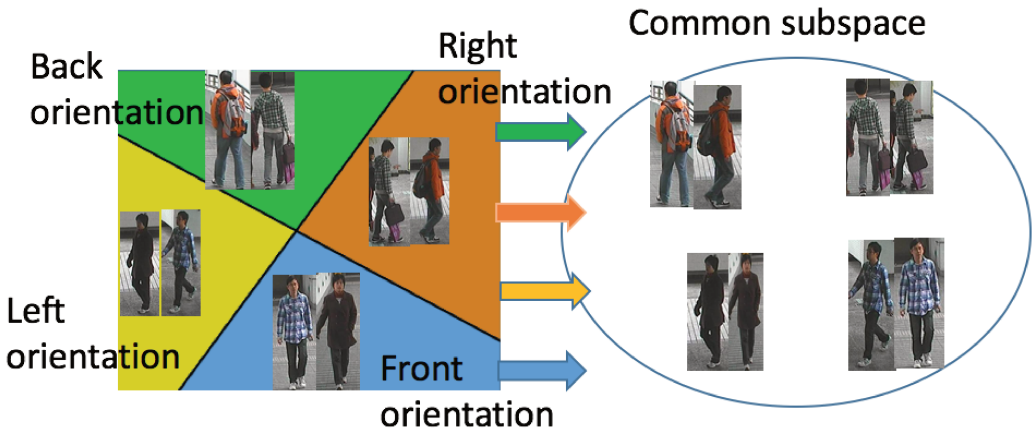


Fig. 4.2 Pedestrian images from different orientations can be considered as different domains. Our method learns different orientation-specific projections into a common feature space.

In the orientation regression branch, 2 convolution layers and 2 fully connected layers are connected to the common convolutional layers. The estimated orientation output by the orientation gating branch is represented by a two-dimensional Cartesian vector $[\alpha,\beta]$ constructed by projecting the orientation angle on the left-right axis (x) and on the front-back (y) axis and then normalizing it to a unit vector. Based on this vector, the orientation gate selects and weights either the left or the right component and either the front or the back component of the re-identification branch. Let $f_{\{left,right,front,back\}}$ be the output feature vectors of the 4 different orientation branches. Since any orientation can be expressed as combination of two of these four main directions (left or right and front or back), the final re-identification output vector is the sum of the left-right component and the front-back component:

$$f_{ouput} = max(\alpha,0)f_{left} + max(-\alpha,0)f_{right} + max(\beta,0)f_{front} + max(-\beta,0)f_{back} \quad (4.1)$$

Different from the classic mixture of experts approach, our orientation gate is set before the local experts, and we perform a regression instead of a classification. The advantage of our orientation gate is that it avoids combining contradictory orientations like front and back. Computationally, only two among four orientations are used and combined according to the sign of α and β. This further allows saving computation.

4.3.2 Training

There are two stages to train the model as shown in Fig. 4.3. In the first stage, the orientation regressor and a general re-identification feature embedding are both trained in parallel with two separate objective functions. In the second stage, the network is specialized to different orientations. These two steps are detailed in the following.

Multi-task network training

We start training the network with pedestrian identity labels and orientation labels respectively.

- **Identification**: for identification learning, we temporarily add an N-dimensional fully-connected layer to the re-identification branch, N being the number of the identities in the training set. The estimated probability of the i^{th} identity is calculated with the softmax function: $p(i) = \frac{exp(z_i)}{\sum_{j=1}^{N} exp(z_j)}$, where $z = [z_1, z_2, ..., z_N]$ is the output of this last fully connected layer. Then, we train the CNN by minimizing the cross-entropy loss:

$$L_{id} = -\sum_{i=1}^{N} log(p(i)) l_{id}(i), \tag{4.2}$$

  where l_{id} is the ground truth one-hot coded identity vector for a given example.

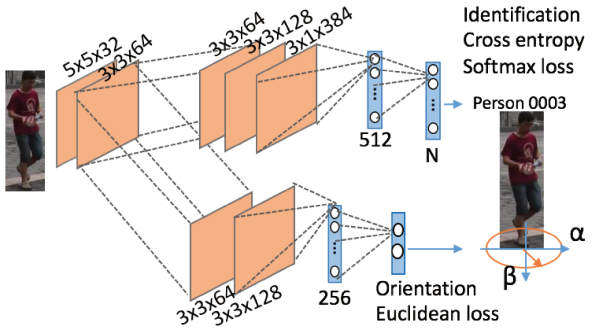
- **Orientation regression**: for the body orientation, we use the Euclidean loss to train the orientation regression of α and β. For a given training example, we have:

$$L_{orien} = \frac{(\alpha - \widehat{\alpha})^2 + (\beta - \widehat{\beta})^2}{2} \tag{4.3}$$

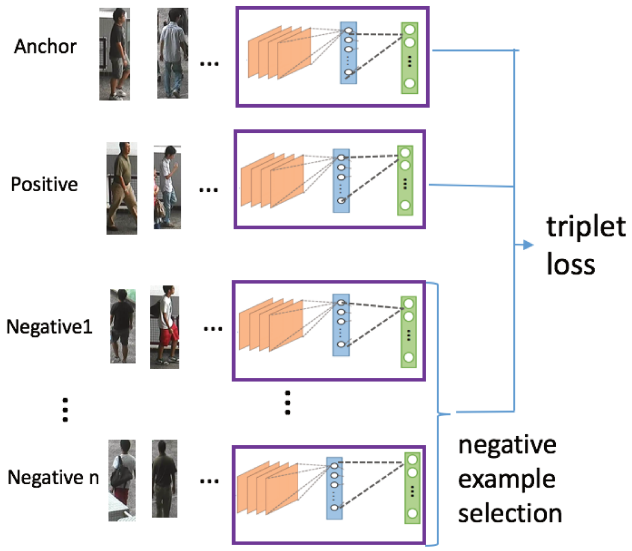
  where $\widehat{\alpha}, \widehat{\beta}$ are predicted orientation labels of the example. Due to the difficulty in estimating the precise body angle, even for humans, orientation is annotated with 8 discrete labels. For training we convert the orientation class to the vector $[\alpha, \beta]$. To

get a more robust orientation learning, we add a uniform random noise of 10 degrees to the orientation labels.

For datasets that have both identity and orientation labels, we train the network with a combined loss $L_{multi-task} = L_{id} + \lambda L_{orien}$. Then, orientation and identification are learned jointly. Otherwise, the two branches are trained separately.



(a)



(b)

Fig. 4.3 The two training steps of our method. (a) In the first step, we train the model with identity and orientation labels . (b) Then, we fine-tune the model to train the orientations-specific layers with hard triplets.

Orientation-specific fine-tuning with triplets

In the second training stage, we fine-tune the network parameters using similarity metric learning in order to specialize the 4 different local experts. For the re-identification branch,

we remove the last fully-connected layer and duplicate four times the the first fully-connected layer. Two orientation gates are integrated to select and weight different orientation projections. Since the different choices and weightings are performed according to the orientation of the person in the input image, the four orientation-specific layers are updated in different ways, whereas the other layers keep their pre-trained weights.

For the similarity metric learning, we use the improved triplet loss with hard examples, as in Chapter 3 (more details see Sec.3.3.2). A (n+2)-tuple of images instead of a triplet is projected into the feature space. The tuple includes one anchor image a, one positive image of the same person p and k negative images of different persons n^j. The loss function for N training examples is defined as follows:

$$E_{triplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(a_i)-f(p_i)\|_2^2 - \min_{j=1..k}(\|f(a_i^j)-f(n_i^j)\|_2^2)+m,0) \\ +\gamma\|f(a_i)-f(p_i)\|_2] \tag{4.4}$$

The network is updated efficiently by pushing the hardest example further away from the anchor. In classic triplet loss, a part of the triplets does not violate the triplet constraint and thus is useless for learning. The selection among k negative examples reduces the number of unused training data and can make the training more efficient. To further enhance the loss function, as [17], we add a term including the distance between the anchor example and the positive example. The parameter γ is used to weight this distance.

4.3.3 Implementation details

The first convolutional layer has a kernel size of 5×5 and the following have a kernel size of 3×3. All following max-pooling layers have a kernel size of 2×2 except the last one in the re-identification branch which has a kernel size of 3×1 without zero-padding increasing the number of channels and reducing the number of parameters by reducing their size to a single column. Batch normalization and a Leaky ReLU activation function with a slope of 0.2 are applied after the max-pooling layers and fully connected layers.Leaky ReLU is one attempt to fix the "dying ReLU" problem. In [125], the author showed improvement with respect to Relu activation function.The first fully-connected layers of the re-identification branch and the orientation gating branch output a vector of respectively 512 and 256 dimensions. Dropout is applied to the fully-connected layers to reduce the risk of overfitting. The optimization is performed by Stochastic Gradient Descent with a learning rate of 0.005, a momentum of 0.9 and a batch size of 50. The constant k is set to 5 and γ is set to 0.002 in Eq. 4.4 as in Chapter 3.

4.4 Experiments

4.4.1 Datasets

The **Market-1501 Dataset** [137] is one of the largest publicly available datasets for human re-identification with 32668 annotated bounding boxes of 1501 subjects.

The **Market-1203 Dataset** [77] is a subset of Market-1501 containing 8570 images from 1203 identities under two camera views. 8 body orientations are annotated. We use 601 identities for training and 602 identities for the test. The test on Market-1203 is performed in the way as Market-1501, that means, we take one image for each identity and each camera view as query (if there's only one image, no image will be taken) and the rest as gallery images. The gallery images from the same identity and the same camera view as the query will be considered as "junk images" which have zero impact on search accuracy. The rank 1 accuracy (R1) and the mean average precision (mAP) are used for performance evaluation.

The **CUHK01 Dataset** [60] contains 971 subjects, each of which has 4 images under 2 camera views. We mannualy annotated each image with 8 body orientations. According to the protocol in [1], the data set is divided into a training set of 871 subjects and a test set of 100 and the extra data from the CUHK03 dataset [61] is also used in training. The Cumulative Match Curve (CMC) is employed as evaluation measure.

For all datasets, to reduce over-fitting, we perform data augmentation by randomly flipping the images and by cropping central regions with random perturbation. For the tests on Market-1501 and on CUHK01 with extra data from CUHK03, since only a part of the images has orientation annotations, the re-identification branch and the orientation gating branch are trained separately. For the test on Market-1203 and the one using only the CUHK01 dataset, The joint multi-task training is performed with the combined loss from Section 4.3.2 and with $\lambda = 0.01$.

4.4.2 Experimental results

Orientation regression evaluation

We first evaluate the performance of orientation regression. We tested the model after the first training stage on the Market-1203 dataset. The confusion matrix is shown in Fig. 4.4. We can see that most predictions are the correct orientation or the adjacent orientation. We calculated also the accuracy measure proposed in [77], *i.e.* the result is considered correct if the predicted and true orientation classes are equal or adjacent. Since person appearances obtained in adjacent orientations are very similar, the exact orientation is less important.

Thus, this accuracy evaluation criterion is more suitable for the person re-identification problem. On the Market-1203 test set, we can get an accuracy of 97.7%.

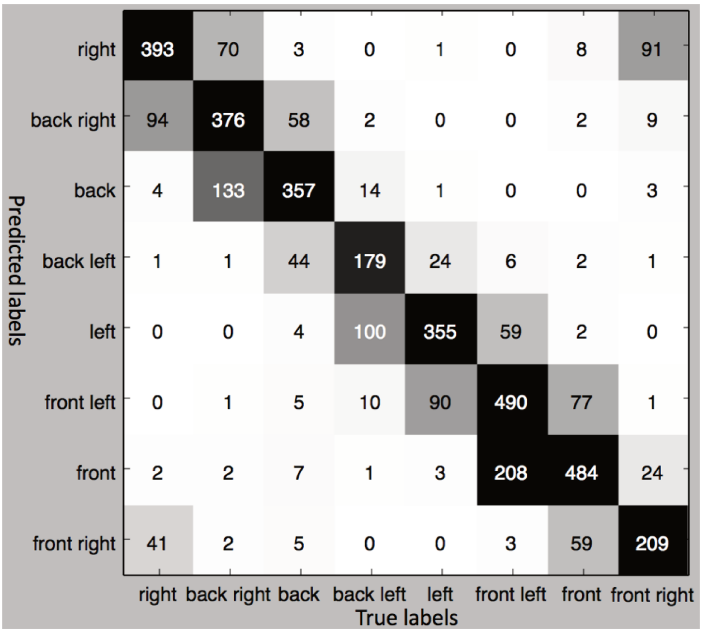


Fig. 4.4 Orientation confusion matrix on Market-1203.

Orientation gate evaluation

To evaluate the effectiveness of our OSCNN, we set up a baseline method. The baseline performs identity learning with softmax loss, then fine-tuning on hard triplets without the orientation gate. The results on Market-1203, Market-1501 and CUHK01 are respectively shown in Tables 4.1, 4.2 and 4.3. Compared to the baseline, integrating the orientation-based local experts in the CNN framework achieves a 1.8% point improvement for rank1 on CUHK01,1.6% and 1.3% points for R1 and mAP on Market-1501 and 1.8% and 1.8% points for R1 and mAP on Market-1203 . This demonstrates the effectiveness of the orientation gate and the specific projections into a common feature subspace.

The improvement is clear but not huge. There are some possible reasons. First, for the mixture-of-expert model, a specific local expert is trained with less training examples (images with specific orientation) than a global model. Secondly, 97.7% of the predictions are the true and their adjacent classes. This may be not enough, since the annotation error maybe up to 45 degrees. Thirdly, the classic CNN is already a very robust model and it has some ability of modeling multi-domain images.

Methods	R1	mAP
Baseline	62.0	64.6
OSCNN	**63.8**	**66.4**

Table 4.1 Experimental evaluation of OSCNN on the Market-1203 dataset.

Analysis of the multi-task learning parameter

To investigate the effect of the parameter λ on the performance accuracy, we conducted experiments using cross validation on the CUHK01 dataset, and the R1 score in function of $log_{10}(\lambda)$ are shown in Fig. 4.5. We can see that the best performances is obtained around $\lambda = 0.01$.

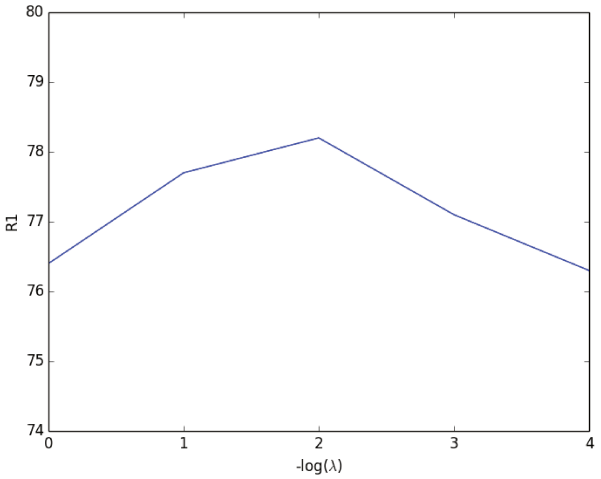


Fig. 4.5 Analysis of the multi-task learning parameter λ.

Comparison with state-of-the-art methods

We compared our OSCNN to the state-of-the-art approaches on Market-1501 and CUHK01. Following the test protocol in [1, 122, 15], we added also the CUHK03 images to the training for the test on the CUHK01 and we compared to the methods only using these two datasets for training. As Table. 4.3 shows, our method is superior to most results of the state of the art. Even without much extra CUHK03 training data, our method shows a competitive performance.

On the Marke-1501 dataset, our OSCNN achieves the same level results as some state-of-the-art methods. Although the result is under the best score of the state of the art, the advantage of our approach is that the model does not need a pre-training step with a much

larger pre-training dataset composed of ImageNet as [141, 56, 142, 106] . And our model has less complexity (1.15×10^8 FLOPs of our model compared to 1.45×10^9 FLOPs of JLML and to 3.8×10^9 FLOPs of SVDNet). Recently some state-of-the-art approaches employ re-ranking [142, 130] which uses information from nearest neighbors in the gallery and significantly improves the performance. As Table. 4.2 shows, our approach can largely benefit from this technique and achieves a state-of-the-art result on Market-1501.

Methods	R1	mAP
Baseline	77.3	53.9
OSCNN	78.9	55.2
OSCNN+re-rank [142]	83.9	**73.5**
LOMO+XQDA [64]	43.8	22.2
PersonNet [122]	37.2	18.6
Gated SCNN [17]	65.9	39.6
Divide fues re-rank [130]	82.3	72.4
LSRO [141]	78.1	56.2
DeepContext [56]	80.3	57.5
K-reciprocal re-rank [142]	77.1	63.6
SVDnet [106]	82.3	62.1
JLML [62]	**85.1**	65.5

Table 4.2 Experimental evaluation of OSCNN on the Market-1501 dataset.

Methods	R1	R5	R10	R20
Baseline(CUHK01)	76.6	93.8	97.0	98.8
OSCNN(CUHK01)	78.2	94.1	97.3	99.1
OSCNN(CUHK01+03)	**83.5**	96.4	**99.0**	**99.5**
LOMO+XQDA [64]	63.2	83.9	90.1	94.2
ImporvedDL [1]	65.0	88.7	93.1	97.2
PersonNet [122]	71.1	90.1	95	98.1
Deep Embedding [100]	69.4	-	-	-
Norm X-Corr [104]	81.2	-	97.3	98.6
Multi-task [15]	78.5	**96.5**	97.5	-

Table 4.3 Experimental evaluation of OSCNN on the CUHK01 dataset.

4.5 Conclusion

In this chapter, we presented a person re-identification approach based on an orientation specific CNN architecture and learning framework. Four orientation-based local experts are trained to project pedestrian images of specific orientations into a common feature subspace. An orientation gating branch learns to predict the body orientation and an orientation gate unit uses the estimated orientation to select and weight the local experts to compute the final feature embedding. We experimentally showed that the orientation gating improves the performance of person re-identification, and our approach outperforms most of the previous state-of-the-art re-identification methods on two public benchmarks.

Chapter 5

Person Re-identification with Listwise Similarity Learning

5.1 Introduction

The deep learning models that have been proposed in the previous chapters have the advantage that they incorporate feature representations and a distance metric in an integrated framework, and they are learned jointly. For training this type of neural networks, different loss functions have been proposed in the litrature such as contrastive loss, triplet loss or quadruplet loss. Unlike these existing functions, in this chapter, we introduce a novel listwise loss function which we call Rank-Triplet loss. It is based on the predicted and ground truth ranking of a list of instances with respect to a query image.

Furthermore, existing deep learning methods are solely based on the minimization of a loss function defined on a certain similarity metric between different examples. However, the final evaluation measures are computed on the overall ranking accuracy. Inspired by the learning-to-rank method LambdaRank [10], our optimization approach directly incorporates these evaluation measures in the loss function. During training, each image in the training batch is used as probe image in turn and the rest as gallery. For each query, the mean average precision and rank 1 score are calculated. And triplets are formed by the probe image and a pair of mis-ranked true and false correspondence.

The loss of one triplet is weighted by the improvement of these evaluation measures by swapping the rank positions of the true and false correspondences. This evaluation measure-based weighting makes better use of difficult triplets which can bring a larger rank improvement and are more effective for the learning, and at the same time, keep the learning

stable by using all misranked pairs. Only using the hardest examples can in practice lead to bad local minima early in training.

The main contributions of this chapter can be summarized as follows:

- A novel listwise loss function that combines triplet loss and LambdaRank loss. This loss considers the re-identification ranking problem in a conceptually more natural way than previous work by directly taking into account the ranking evaluation scores.
- A thourough experimental evaluation showing that the proposed loss outperforms other common loss functions and achieves state-of-the-art results on the challenging Market-1501, DukeMTMC-Reid, CUHK03 datasets. Moreover, to further show the effectiveness of our loss function, we apply it to an image retrieval task with Holidays dataset and it shows a competitive result compared to the state-of-the-art.

5.2 Related work to Rank-triplet loss

5.2.1 Variants of triplet loss

The triplet loss was first applied by Ding *et al*. [24] to train a CNN for person re-identification. Cheng *et al*. [17] proposed an improved variant of the triplet loss function by combining the contrastive loss and a CNN network processing parts and the entire body. Chen *et al*. [14] applied a quadruplet loss which samples four images from three identities and minimizes the difference between a positive pair from one identity and a negative pair from two different identities and they combine this quadruplet loss with the triplet loss.

Using triplets reduces the imbalance problem in siamese networks, but another drawback of both siamese and triplet loss learning is that the trivial pairs or triplets become inactive at a later training stage. To tackle this problem, some methods exploit hard example mining to enhance convergence and overall performance. Ahmed *et al*. [1], for example, used the difference of feature maps to measure the similarity and performed hard negative example pair mining. Shi *et al*. [100] proposed to perform moderate positive and negative example mining to ensure a stable training process and avoid perturbing the manifold learning by using hard examples. On the contrary, Hermans *et al*. [38] proposed to use the hardest positive and negative examples in each training batch to perform an effective triplet learning.

5.2.2 Learning-to-rank

Learning-to-rank is a class of techniques that learns a model for optimal ordering of a list of items. It is widely applied in information retrieval and natural language processing. Many

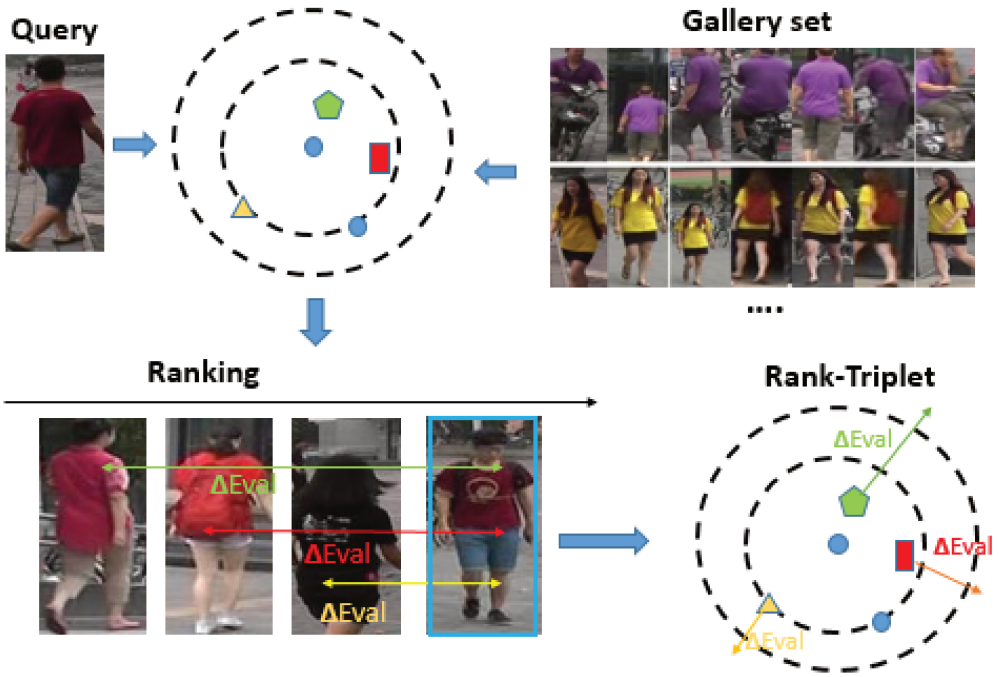


Fig. 5.1 Schematic illustration of Rank-triplet. An image of a person of interest on the left (the query) is used to rank images from a gallery according to how closely they match that person. The correct match, highlighted in a blue box, can be difficult to find given the similar negative images, pose and viewpoint variations and occlusions. During training, we propose to estimate the importance of misranked pairs by the gain of the evaluation measure incurred by swapping the rank positions and to weight the loss according to their importances. In this example, swapping the falsely ranked (positive) image on the right with the leftmost one would lead to the biggest improvement (ΔEval).

learning-to-rank methods have been proposed in the literature, like pairwise approaches RankSVM [37], RankNet [9] and listwise approaches ListNet [13], ListMLE [124] and LambdaRank [10], taking the entire ranked list of objects as the learning instance. Since person re-identification could be considered as a retrieval problem based on ranking, some person re-identification approaches applied these techniques like Prosser *et al*. [91] who reformulated the person re-identification problem as a ranking problem. Their method learns a set of weak RankSVMs, each computed on a small set of data, and then combines them to build a stronger ranker using ensemble learning. Wang *et al*. [117] applied the ListMLE method to the person re-identification problem: their approach maps a list of similarity scores to a probability distribution, then utilizes the negative log likelihood of ground truth permutations as the loss function.

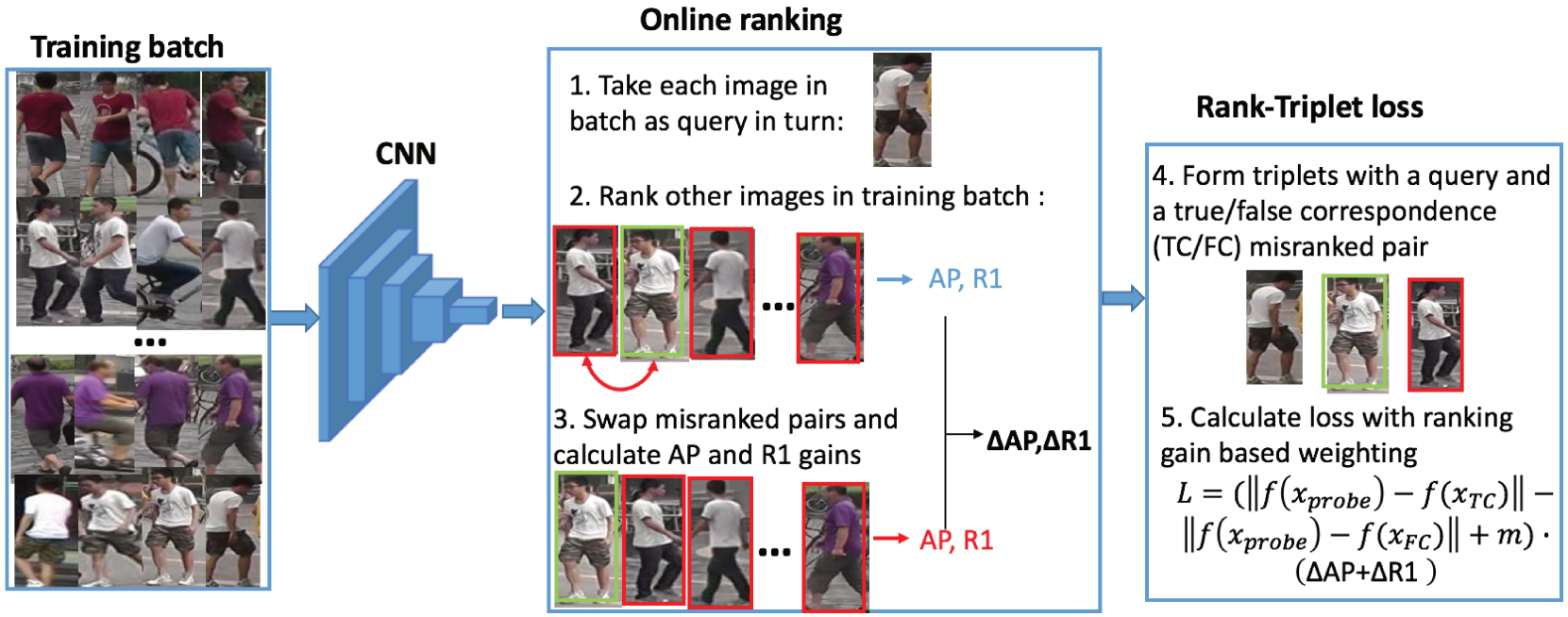


Fig. 5.2 Overview of the training procedure of the proposed Rank-Triplet approach

5.3 Rank-triplet loss function

In the following, we will first describe the learning-to-rank methods Ranknet and LambdaRank and the person re-identification evaluation measures. Then we will explain how to perform our proposed Rank-Triplet loss learning in terms of the evaluation measures. An overview of our approach is shown in Fig. 5.2.

5.3.1 Ranknet and LambdaRank

Ranknet is a neural network based learning-to-rank method. Query dependent features are extracted as the inputs of the network. Given the scores s_i and s_j of two items with respect to the query, $s_{ij} = s_i - s_j$. Let $x_i \triangleright x_j$ denote the event that the item x_i should be ranked higher than x_j. Then the two scores are mapped to a probability that x_i should be ranked higher than x_j via a sigmoid function:

$$P_{ij} \equiv P(x_i \triangleright x_j) = \frac{1}{1+e^{-s_{ij}}} \tag{5.1}$$

To learn the model, the cross entropy cost function is minimized .It penalizes the deviation of the model output probabilities from the desired probabilities: let $P_{ij} = \{-1, 1\}$ be the known probability that training x_i should be ranked higher than training x_j. Then the loss function is

$$L_{ranknet} = -\overline{P}_{ij} log P_{ij} - (1-\overline{P}_{ij}) log(1-P_{ij}) \tag{5.2}$$

RankNet uses a neural network model that is trained using this pair-based cross entropy cost. Thus, it is minimizing the number of pairwise errors and does not consider other information retrieval measures. To directly optimize these measures is difficult as they are not differentiable, thus leading to a non-convex problem that cannot be solved with gradient descent-based algorithms commonly used for neural network models. To tackle this problem, Burges *et al*. [10] proposed LambdaRank which, at each training iteration, simply scales the gradient of the loss function by the difference of the document retrieval evaluation measure Normalized Discounted Cumulative Gain (NDCG) incurred by swapping the rank positions of two items, as shown in Eq. 5.3. They show that this approach improves the overall ranking performance.

$$\lambda = \frac{\partial L_{ranknet}}{s_{ij}} \cdot \Delta NDCG \tag{5.3}$$

The triplet learning has shown good performance on verification problems like image classification [116], face recognition [96] and person re-identification [24, 17, 38]. However, in triplet learning for person re-identification, we face a similar problem to RankNet. The classical triplet loss is defined on the partial order relations among identities, However, the ranking measures are calculated on the global order. That means that the triplet loss iteratively enforces pair-wise order relationships w.r.t. reference examples, but it is difficult to generalize this approach for optimizing the global order. In this regard, a listwise ranking is a better approximation of this global order relation, and we adapt it to the person re-identification problem, as explained in Section 5.3.3.

5.3.2 Person re-identification evaluation measure

Cumulated Matching Characteristics (CMC) and mean average precision (mAP) are widely used performance measures for person re-identification. CMC evaluates the top n nearest images in the gallery set w.r.t. one probe image. If a correct match of a query image is at the k^{th} position (k$\leqslant$n), then this query is considered a success of rank n. In most cases, we look at the success of rank 1 (R1), *i.e.* the person has been correctly re-identified. The CMC curve shows the probability that a query identity appears in different-sized ordered candidate lists. As for mAP, for each query, we calculate the area under the Precision-Recall curve, which is known as average precision (AP):

$$AP = \int_0^1 P(R)\,\mathrm{d}R, \tag{5.4}$$

where P(R) is the precision for a given recall R. Then, the mean value of AP of all queries, *i.e.* $mAP = \frac{1}{n}\sum_n^{i=1} AP_i$, where n is the number of queries, considers both precision and recall of an algorithm, thus providing a more suitable evaluation for the setting in which there are several true correspondences in the gallery set.

Since P(.) and R(.) are discrete functions, the area under the precision-recall curve is approximated as [137]:

$$AP = \sum_{k=1}^{N} \frac{p(k)+p(k-1)}{2}[r(k)-r(k-1)], \tag{5.5}$$

where k is the rank in the sequence of retrieved items. p(k) and r(k) are respectively the precision and recall at the rank k position. We define also p(0)=1 and r(0)=0. N is the number of images in the gallery set.

Since in our method the AP is calculated at each iteration during training, we propose to simplify this computation. In ranking problems, the recall is the fraction of items that are relevant to the query and successfully retrieved, the variation r(k)-r(k-1) is different from zero only when a relevant item is retrieved through the sequence of retrieved items. We only need to take into account the true correspondence ranking position and the variation of recall equals always $\frac{1}{M}$, where M is the number of the true correspondences of a query. Thus AP can be calculated as:

$$AP = \frac{1}{2M}[1 + p(\pi_1) + \sum_{i=2}^{M} p(\pi_i) + p(\pi_{i-1}))], \tag{5.6}$$

where π_i is the rank index of the i^{th} true correspondence. Precision is defined as the proportion of retrieved non-relevant items out of all non-relevant items available. Thus the precision at ranking position π_i is: $p(\pi_i) = \frac{i}{\pi_i}$. We can further simplify the equation:

$$AP = \frac{1}{M} \sum_{i=1}^{M} \left[\frac{i}{\pi_i} \right] - \frac{1}{2\pi_M} + \frac{1}{2M} \tag{5.7}$$

5.3.3 Rank-Triplet loss

The triplet loss uses triplets of examples to train the network with an anchor image a, a positive image p from the same person as a and a negative image n from a different person. Training imposes that the projection of the positive example is placed closer to the anchor than the projection of the negative example. This constraint is defined as following:

$$\|f(a_i) - f(p_i)\|_2^2 < \|f(a_i) - f(n_i)\|_2^2 \tag{5.8}$$

The weights of the network for the three input images are shared, and to train the network, the constraint of Eq. 5.8 is formulated as the minimization of the following triplet loss function:

$$E_{triplet} = -\frac{1}{N} \sum_{i=1}^{N} [max(\|f(a_i) - f(p_i)\|_2^2 - \|f(a_i) - f(n_i)\|_2^2 + m, 0)], \tag{5.9}$$

where N is the number of triplets, f is the projection of the network, and m is a margin. With the triplet loss function, the network learns a semantic distance metric by "pushing" the negative image pairs apart and "pulling" the positive images closer in the feature space.

In order to update the weights of the network, it is crucial to select triplets that violate the triplet constraint 5.8. However, in practice, the majority of the triplets does not violate

the constraint at a later learning stage. Hard triplet mining is an effective way to tackle this problem, but some too hard triplets may distort the manifold [16]. We propose to select the triplets according to the ranking order, *i.e*. only mis-ranked matches will be selected. Not using only the hardest examples stabilizes the training, and weighting the triplets according to their contribution makes the learning more effective.

The overall training algorithm is presented in Algorithm 1 and shown in Fig 5.2. In order to optimize directly the AP and R1 scores, we estimate the gain for AP and R1 of the triplets from the ranking within a training batch. A training batch is formed by M images of N identities. For each example in the batch, we perform a ranking among the rest of images in the batch. For the sake of a robust metric, we add a margin m to the distance of ranking positions between the true correspondences and the probe before ranking. The AP and R1 scores are computed for each query ranking. Then, with respect to one probe, we form all possible mis-ranked pairs (false correspondences ranked before the true correspondence), and we re-calculate the new AP and R1 scores by swapping positions of the pair in the ranking and thus obtain the gains ΔAP and $\Delta R1$, respectively. The loss of each triplet is weighted by the sum of these gains. The final Rank-triplet loss is calculated as follows:

$$E_{rank-triplet} = \frac{1}{MN}\sum_{i=1}^{MN}\frac{1}{K_i}\sum_{j\in TC_i}\sum_{\substack{k\in FC_i \\ r_k^i<r_j^i}}[\|f(x_i)-f(x_j)\|_2^2 - \|f(x_i)-f(x_k)\|_2^2+m]\cdot(\Delta AP_{jk}^i+\Delta R1_{jk}^i), \quad (5.10)$$

where x_i is the i^{th} training example in a training batch, K_i is the number of misranked pairs w.r.t. the i^{th} example as query, and r_j^i is the rank of the j^{th} example w.r.t. the i^{th} image as query. TC_i/FC_i is the true/false correspondence set of the i^{th} example. ΔAP_{jk}^i is the gain of AP by swapping the j^{th} and k^{th} examples w.r.t. the i^{th} example as query and analogously for R1.

With our evaluation based weighting, we make a trade-off between the moderate hard examples and hardest examples, i.e. more weight is given to the hardest examples to make the learning efficient, and, at the same time, the less hard example are used to stabilize the training.

5.4 Experiments and results

In this section, we report the experimental results carried out on the person re-identification datasets Market-1501 [137], DukeMTMC-Reid [141] and CUHK03 [61] to compare our

Algorithm 1: Similarity learning with Rank-Triplet loss

Input: Training image set, identity label set, learning rate λ
Output: The network weights W

Initialize W **for** $t = 1...T$ **do**
  Randomly sample M identities
  Randomly sample K images for each identity
  Form the training batch with images and identity labels $X = \{x_i\}_{i=1}^{KM}, ID = \{Id_i\}_{i=1}^{KM}$
  Forward pass to obtain image embeddings $Y = \{y_i\}$:
  $Y = f_w(X)$
  L = 0
  for $i = 1...KM$ **do**
    $\mathscr{D} \leftarrow dist(y_i, y_{j=1...KM, j\neq i})$
    foreach j *that* $Id_i = Id_j$ **do**
      $\mathscr{D}_j \leftarrow \mathscr{D}_j + margin$
    end
    $\mathscr{R} \leftarrow sort(\mathscr{D})$
    $AP, R1 \leftarrow calculateAPR1(\mathscr{R})$
    foreach j *with* $Id_i = Id_j$ **do**
      foreach k *with* $\mathscr{R}_k < \mathscr{R}_j$ *and* $Id_k \neq Id_j$ **do**
        $\mathscr{R}' \leftarrow swap(\mathscr{R}_k, \mathscr{R}_j)$
        $AP', R1' \leftarrow calculateAPR1(\mathscr{R}')$
        $eval\_gain \leftarrow AP' - AP + R1' - R1$
        $L \leftarrow L + (\mathscr{D}_j - \mathscr{D}_k) \times eval\_gain$
      end
    end
  end
  $Loss \leftarrow \frac{L}{N}$
  $W^t \leftarrow W^{t-1} - \lambda \frac{\partial Loss}{\partial W}$
end
Return W

approach with the state-of-the-art approaches. We also perform a comparison with other loss functions commonly used in the literature. We further perform a more detailed analysis of the proposed method in several aspects, like its convergence behaviour and training time. Finally, to show the genericity of our approach, we applied it to an image retrieval task. We performed experimental evaluations on the Holidays dataset and compared it to the state-of-the-art methods and results.

5.4.1 Person re-identification

Datasets

The Market-1501 dataset [137], DukeMTMC-Reid dataset [141] and CUHK03 dataset [61] are used for the evaluation (for a detailed description of the datasets seeSection 2.3.5). For the CUHK03 dataset, our experiments will be conducted on the detected version which is a more realistic scenario. We followed the new test protocol proposed in [142] which splits the CUHK03 dataset into training set and testing set similar to that of Market-1501, which consist of 767 identities and 700 identities respectively.

All the three datasets follow the same test protocol. The authors randomly select one image from each camera as the query for each identity and use the rest of images to construct the gallery set. In evaluation, true matched images captured from the same camera as the query are not considered. Thus, these images have no influence on the re-identification accuracy.

Implementation Details

We used Alexnet [49] and Resnet-50 [36] as the model architecture and the weights pre-trained on the ImageNet dataset are used as initialization. Any other CNN architecture, pre-trained or not, could have been used. But for the matter of comparison with the state of the art, and to concentrate on the proposed Rank-triplet approach which is independent from the underlying model, we evaluated our method using standard CNN architectures. We replaced the final layer of the models by a fully-connected layer with 256 output dimensions. Each input image is resized to 224 $\times$112 pixels. Data augmentation is performed by randomly flipping the images and cropping central regions with random perturbation. The Adam optimizer [47] is used and the initial learning rate is set to 10^{-4}. Each 80 epochs the learning rate is decreased by a factor of 0.1. The weight decay is set to 0.0005. The training is performed in 200 epochs. And the batch size is set to 128 from 32 identities with 4 images each. The 32 identities are randomly selected without replacement until all the identities have been taken.

	Resnet		Alexnet	
	R1	mAP	R1	mAP
hardbatch	81.0	63.9	-	-
baseline	82.1	66.5	70.9	47.3
Rank-triplet	**83.6**	**67.3**	**72.7**	**49.1**

Table 5.1 Re-identification performance on the Market-1501 dataset in terms of rank 1 (R1) and mean average precision (mAP) (in %) for different loss functions and neural network models.

Experimental results

Comparison of different neural network models. Alexnet and Resnet have been proven to be a very successful model for image classification. We implemented the hardbatch triplet loss, our baseline and rank-triplet for the training of both models. Table 5.1 shows the re-identification results on Market-1501. The Resnet50-based model trained with Rank-triplet shows a better performance than the Alexnet-based model by a margin of 9.9% points for R1 and 18.2% points for mAP. Integrating our evaluation measure gain weighting of Rank-triplet increased the mAP by 1.8% points and 0.8% points and R1 by 1.8% points and 1.5% points with Alexnet and Resnet respectively. The hardbatch approach with Alexnet cannot converge on the Market-1501 dataset. This demonstrates that hard example mining can make the learning more effective, but only using the hardest examples may severely perturb the learning process.

Comparison of different loss functions. We conducted experiments with different common loss functions, and results are shown in Table 5.2. For the supervised classification with identity labels, the softmax cross entropy loss is used. The margin in the Siamese loss and triplet loss is fixed to the default value m =1. For the pairwise Siamese learning the contrastive loss is used, we generate all possible pairs of images within a batch. The loss is calculated with Eq. 2.10.

The triplet loss is calculated according to Eq. 5.9. And the hard batch triplet loss takes only the hardest positive image and negative image, the hardbatch triplet loss is calculated as follows.

$$L_{hard-batch} = \frac{1}{N}\sum_{i=1}^{N} max(\max_{j \in TC_i} \|f(x_i) - f(x_j)\|_2^2 - \min_{k \in FC_i)} \|f(x_i) - f(x_k)\|_2^2 + m, 0). \quad (5.11)$$

where N is the number of triplets, TC_i/FC_i is the true/false correspondence set of the ith example.

The quadruplet loss in [14], based on triplets, pushes away also negative pairs from positive pairs w.r.t different probe images. The loss is formulated as:

$$E_{quadruplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(x_i)-f(x_j)\|_2^2-\|f(x_i)-f(x_k)\|_2^2+m_1,0) + max(\|f(x_i)-f(x_j)\|_2^2-\|f(x_k)-f(x_l)\|_2^2+m_2,0)], \quad (5.12)$$

where x_j is the feature embeddings of an image from the same identity as x_i and x_k,x_l are from different identities. As [14], we set the $m_1 = 1, m_2 = 0.5$.

Finally, we implement the baseline called "rank triplet selection" by calculating the loss without the term of evaluation gain weighting. But the triplet selection is still based on online ranking orders.

Rank-triplet achieved the best performance among these loss functions. The Rank-triplet improves the baseline "rank triplet selection" by a margin of 1.5% points for R1 and 0.8% points for mAP. This shows the effectiveness of the listwise evaluation measure-based weighting. The baseline "rank triplet selection" and the Rank-triplet gives also better results than the hardbatch. This shows that using moderate difficult examples and weighting them helps the metric learning. In fact, hardbatch is a particular weighting with weight=1 given to the hardest examples and 0 given to the rest. Also hardbatch shows better performance than the normal triplet loss, confirming the effectiveness of the hardest example mining in [1, 38]. Using the quadruplet loss also slightly improves the performance with respect to triplets. This could eventually be combined with our loss.

Loss function	R1	mAP
Classification loss	74.3	51.0
Siamese loss	62.9	46.6
Triplet loss	74.3	56.5
Quadruplet loss	74.9	58.1
Hardbatch*	81.0	63.9
rank triplet selection	82.1	66.5
Rank-triplet	**83.6**	**67.3**

Table 5.2 Re-identification results(in %) on Market-1501 with different loss functions.*: The result of our re-implementation of [38]. To notice that we did not use the same training parameters and fc layer settings as [38] and in [38], the test data augmentation is performed.

Comparison with state-of-the-art methods. We compared our method with state-of-the-art methods on the three benchmark datasets. The results are shown in Table 5.3.

Methods	Market-1501		DukeMTMC-Reid		CUHK03-NP	
	R1	mAP	R1	mAP	R1	mAP
Hardbatch triplet loss [38]	81.0	63.9	62.8	42.7	46.4	50.6
Our baseline rank triplet selection	82.1	66.5	72.4	52.0	45.3	48.9
Our Rank-Triplet loss	83.6	67.3	74.3	55,6	47.8	52.4
Rank-Triplet+re-rank [142]	86.2	**79.8**	78.6	**71.4**	**60.4**	**60.8**
LOMO+XQDA [64]	43.8	22.2	30.8	17.0	12.8	11.5
LSRO [141]	78.1	56.2	67.7	47.1	-	-
Divide and fuse [130]	82.3	72.4	-	-	30.0	26.4
K-reciprocal re-rank [142]	77.1	63.6	-	-	34.7	37.4
ACRN[97]	83.6	62.6	72.6	52.0	-	-
SVDNet [106]	82.3	62.1	76.7	56.8	41.5	37.3
JLML [62]	85.1	65.5	-	-	-	-
DPFL [16]	**88.6**	72.6	**79.2**	60.6	40.7	37.0

Table 5.3 Comparison with the the sate-of-the-methods on person re-identification

Our method Rank-triplet achieves better results than most of the other methods on the three benchamarks. Only on Market-1501, DPFL obtains a slightly better result, and SVDNet and DPFL on DukeMTMCReid.

On the CUHK03 benchmark, our methods achieves the best results. DPFL and SVDNet are based on a classification loss, and the CUHK03 dataset contains fewer imaegs per person. That is not enough to train a good classifier. However, triplet loss is not much affected because we could still form a large number of triplets even there's less image per peron. Since the main contribution of these two state-or-the-art methods focus on the network architecture, their methods can potentially be eventually combined with our loss function.

The hardbatch triplet learning on DukeMTMC-Reid had difficulty to converge with an initial learning rate of 10^{-4}. The convergence improved when the learning rate was reduced to 2×10^{-5}. But it still gave an inferior final performance.

Some successful and failed top-10 Rank-triplet results are shown in Fig. 5.3 and 5.4. As can be seen, most of the errors are due to the high clothing similarity among pedestrians and to some partial occlusion. Even for a human, it is difficult to decide if they represent a real match or not.

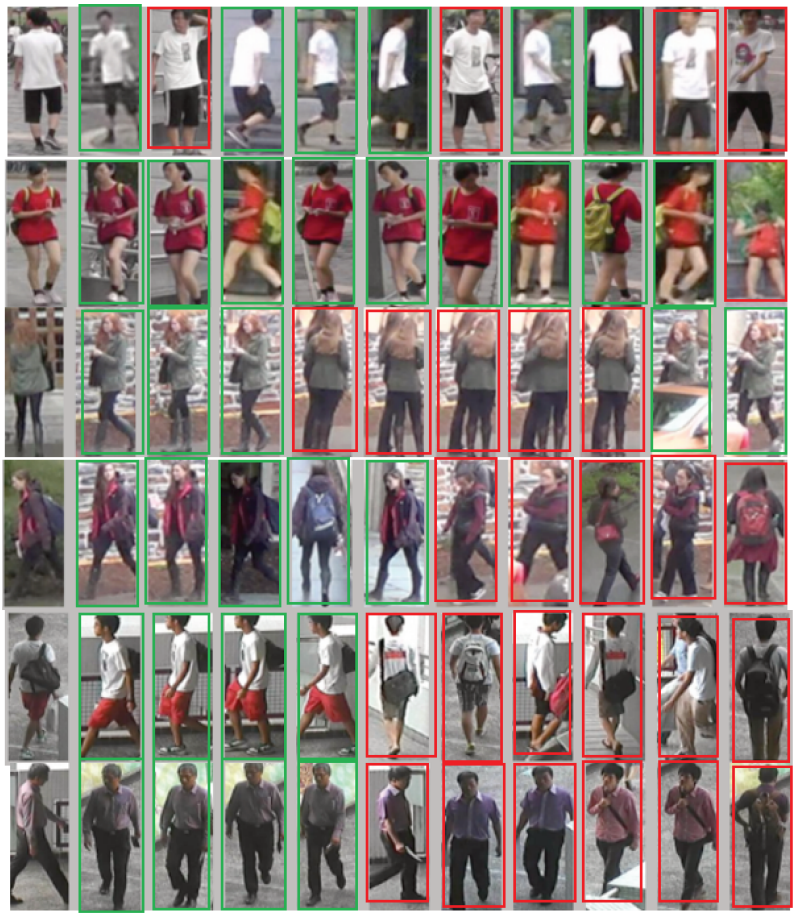


Fig. 5.3 Some successful ranking results. The query image is on the left and the true matches are surrounded by green boxes. The two top rows are from Market-1501, the two middle rows are from DukeMTMC-Reid, and the two bottom rows are from CUHK03

Fig. 5.4 Some failed ranking results. The top row is from Market-1501, the middle row is from DukeMTMC-Reid, and the bottom row is from CUHK03

Analysis of the proposed method

Evolution of R1 and mAP during training. We further analyzed the R1 and mAP values computed for each batch during training. Further, a separate validation batch is formed with 128 random images from 32 persons that were not used for training We evaluate the mAP, R1, number of misranked pairs and loss for each epoch. The curves showing the evolution of these measures during training are shown in Fig. 5.5.

We can observe that the R1 and mAP computed on the training batches converge to 1 and the number of misranked pairs almost converges to 0. The validation mAP and R1 also increase and the misranked pairs decrease during the training. The validation loss naturally increases because the loss is the average among the misranked pairs, and after solving the simple cases, it remains only the mis-ranked pairs giving a high loss with a large evaluation weighting.

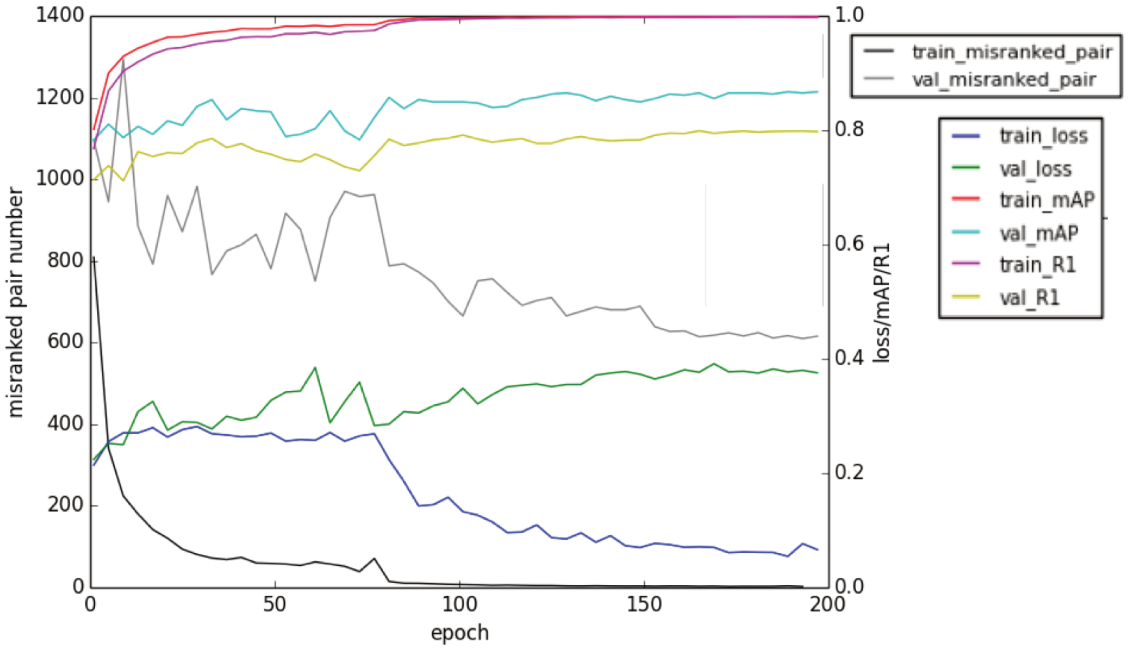


Fig. 5.5 Training and validation mAP, R1 and misranked pair number curve

Training time analysis. In order to analyze the complexity of Rank-triplet, we compared its training time with the one of hardbatch. The results are shown in Table 5.4. All algorithms are implemented in Pytorch. The training was performed with Intel i7-5930K 3.50GHZ CPU and 2 Nvidia GTX Titan Maxwell GPUs. The hardbatch triplet loss has first been implemented with Eq. 5.11. Surprisingly, this implementation takes about 1 hour more for training than our Rank-triplet. A probable explanation is that the Hinge function slowed down the training. That means, even if the triplets do not violate the constraint, the gradient

is still calculated on these triplets. Then we optimized the Hardbatch code by adding a condition flow in order to calculate the loss only for the necessary triplets. The training time for Rank-triplet is considerably reduced. After this optimization, the hardbatch takes about 40 minutes less than the Rank-triplet loss. That difference roughly corresponds to the time of ranking, evaluation measure computation and the use of more triplets at each iteration. However, we consider that this is a reasonable extra cost given the overall performance improvement.

Loss	Training time
Hardbatch hinge triplet loss	6h10min
Hardbatch triplet loss with condition flow	4h25min
Rank-triplet loss	5h04min

Table 5.4 Training time on Market-1501

Batch size analysis. To investigate the effect of the training batch size on the performance accuracy, we conducted on experiment with a smaller batch size on the Market-1501 dataset and the results are shown in Tab. 5.5. The batch size is set to 72, from 18 identities with 4 images each. We can see that the Rank 1 score and mAP are dropped respectively 2.6% and 2.8% points by using a smaller batch size. In fact, the online ranking that we perform during training is a kind of simulation of the ranking on the test set. We expect to improve the test ranking by correcting the online ranked list. The ranking performed with a larger training batch size has a better generalization ability and more triplets can be exploited, which helps the network converge to a more robust solution. However, the computation and the memory use of a large batch size is more expensive.

Training batch size	R1	mAP
72	81.0	64.5
128	**83.6**	**67.3**

Table 5.5 The effect of batch size on results on Market-1501

5.4.2 Image retrieval

To further evaluate our method, we tested our Rank-triplet loss on a more general content-based image retrieval problem, where the task is to retrieve images from a gallery set that belong to the same category as the probe image or are similar to it. As with the person re-identification task, the challenge of image retrieval is translation, rotation and scaling transformation of the objects of interest in the images and also illumination changes. We

use the INRIA Holidays [43] dataset to perform the test. Images are considered from the same category/class, i.e. relevant to a specific query, if they are taken in the same scene or showing the same object under different viewpoints. The dataset contains 500 queries and 991 corresponding relevant images. For the training, we used the landmark dataset as in [3, 31]. However, we were only able to use a subset of the dataset due to broken URLs. In total, we used 28777 images of 560 landmarks for training. For the training and the test, the input image is randomly cropped to 320×320 from 362×362. Since there is a high variance of translation and scale of relevant objects inside images, we replaced the last global average pooling by a global max-pooling as in [108]. All other experimental settings remain the same.

Table 5.6 shows the comparison with the state-of-the-art methods and with the baseline. Our method performs slightly better than the baseline and is superior to most state-of-the-art results. A probable explanation of lower improvement in the image retrieval task is that in the landmark dataset, more types of variations are present in images, it means the formed triplets are generally difficult and the triplet weighting could be less effective. The ROI-triplet method uses also a triplet network and integrates a pre-trained ROI pooling to localise the salient image content. This technique could also be integrated in our model to further improve the instance retrieval performance.

Method	mAP (in%)
Neural codes [3]	75.9
R-MAC [108]	85.2
NetVLAD [2]	83.1
Cross-dimension weighting [45]	84.9
Hard siamese [92]	82.5
ROI-Triplet [31]	**90.7**
Baseline	85.1
Our Rank-triplet	85.8

Table 5.6 Experimental evaluation on the Holidays dataset

5.5 Conclusion

In this chapter, we proposed to use a listwise loss function to perform similarity learning for person re-identification. We introduced a novel listwise loss function based on ranking evaluation measuresand the idea of LamdaRank. An online ranking within training batches is performed to evaluate the importance of different triplets composed of probe, misranked true and false correspondences and to weight the loss with the rank improvement for a given query.

We experimentally showed that taking into account the evaluation measures during training and calculating the loss in a listwise way improves the overall ranking and recognition performance. Further, our proposed loss function outperforms other common functions in the literature and achieved state-of-the-art results on three different person re-identification benchmarks. Finally, we applied the proposed approach to a more general image retrieval problem with photographs of very diverse content. Without any major modifications, our algorithm outperformed most state-of-the-art methods on the Holiday benchmark showing the genericity of our approach.

Chapter 6

Person re-identification Using Group Context

6.1 Introduction

Most existing approaches only use the visual appearance of a *single* person for its re-identification in different images. However, this can lead to strong ambiguities, for example when people wear similar clothes and colour, as shown in Fig. 6.1. The problem becomes increasingly difficult when there is a large number of candidate persons since, in practice, many persons have similar appearance as they share the same visual attributes. To address this problem, context information about the surrounding group of persons can be used. In realistic settings, people often walk in groups rather than alone. Thus, the appearance of these groups can serve as visual context and help to determine whether two images of persons with similar clothing belong to the same individual.

However, matching the surrounding people in a group in different views is also challenging. On the one hand, it undergoes the variations of single person appearances. On the other hand, the number of persons and their relative position within the group can vary over time and across cameras. Further, partial occlusions among individuals are very likely in groups.

In this chapter, we propose to extract group feature representations using a deep convolutional neural network. First, we train the model with single-person re-identification data, and then, transfer it to the group association problem. In order to cope with the relative displacements of persons in a group, we applied a Global Max-Pooling (GMP) operation of CNN activations to achieve translation invariance in the resulting representation. Furthermore, we measure a group context distance with this representation and then combine it

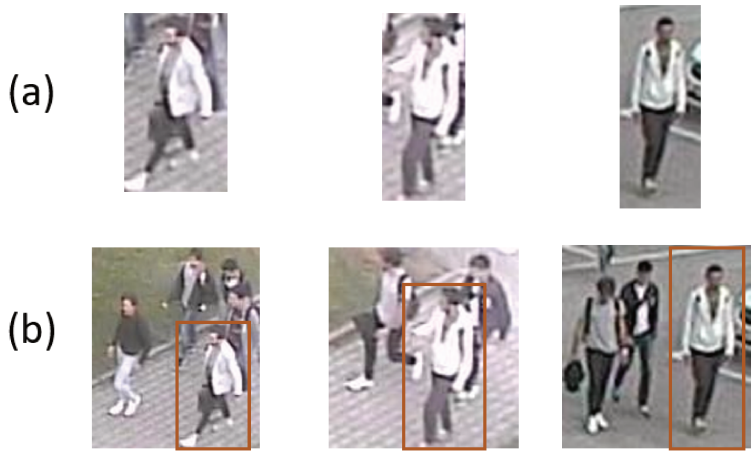


Fig. 6.1 (a) Single person images. (b) Corresponding group images of (a). Even for a human, it may be difficult to tell if the three top images belong to the same person or not. Using the context of the surrounding group, it is easier to see that the middle and right images belong to the same person and the left image belongs to another person.

with the distance measure based on single-person appearance to enhance the re-identification accuracy.

The main contributions of this chapter are the following:

- we learn a deep feature representation with displacement invariance and apply it to the group association problem. Our experiments show that this approach outperforms the state-of-the-art on group association.
- we propose a novel way to combine group context and single-person appearance and experimentally show that the group information can improve the person re-identification performance.

6.2 Related Work to Group association

In the literature, there are several group association (or group re-identification) approaches. Zheng *et al*. [138] extracted visual words which are the clusters of SIFT+RGB features in a group image. Then they built two descriptors that describe the ratio information of visual words between local regions to represent group information. Cai *et al*. [12] used covariance descriptor to encode group context information. And Lisanti *et al*. [67] proposed to learn a dictionary of sparse atoms using patches extracted from single person images. Then the learned dictionary is exploited to obtain a sparsity-driven residual group representation. These approaches can be severely affected by background clutter, and thus a preprocessing

stage is necessary. For example, in [138, 12] a background subtraction is performed before feature extraction. And in [67], three pedestrian detectors based on respectively deformable part models, aggregated channel features and RCNN were used to weight the contribution of each pixel in the histogram computation.

Some other approaches use trajectory features to describe group information. Wei *et al*. [120], for example, presented a group extraction approach by clustering the persons' trajectories observed in a camera view. They introduced person-group features composed of two parts: SADALF features [26], extracted after background subtraction and representing the visual appearance of the accompanying persons of a given individual, and a signature encoding the position of the subject within the group. Similarly, Ukita *et al*. [110] determined for each pair of pedestrians whether they form a group or not, using spatio-temporal features of their trajectories like relative position, speed and direction. Then, the group features composed of the trajectory features (position, speed, direction) of individuals in each group, the number of persons as well as the mean colour histograms of the individual person images. However, when people walk in group, the position and speed are not always uniform. Thus, the trajectory-based features may not be precise and change significantly over time.

Unlike these methods, the advantage of our approach is that there is no need for a pre-processing stage of person detection or background subtraction. Our model is pre-trained on single-person re-identification data to learn the discriminative features that distinguish identities in images. The applied global max-pooling operation captures maximum activations over feature maps,which correspond to salient discriminative patches in the input image. Thus the proposed model is, by design, invariant to displacements of individuals within a group. Moreover, the deep neural network that we employed can provide a richer feature representation to describe groups than the colour and texture features used by existing methods.

6.3 Proposed method

In this section, we first describe our group association method and further introduce how we use the group information to improve person re-identification performance. An overview of the method is shown in Fig. 6.2

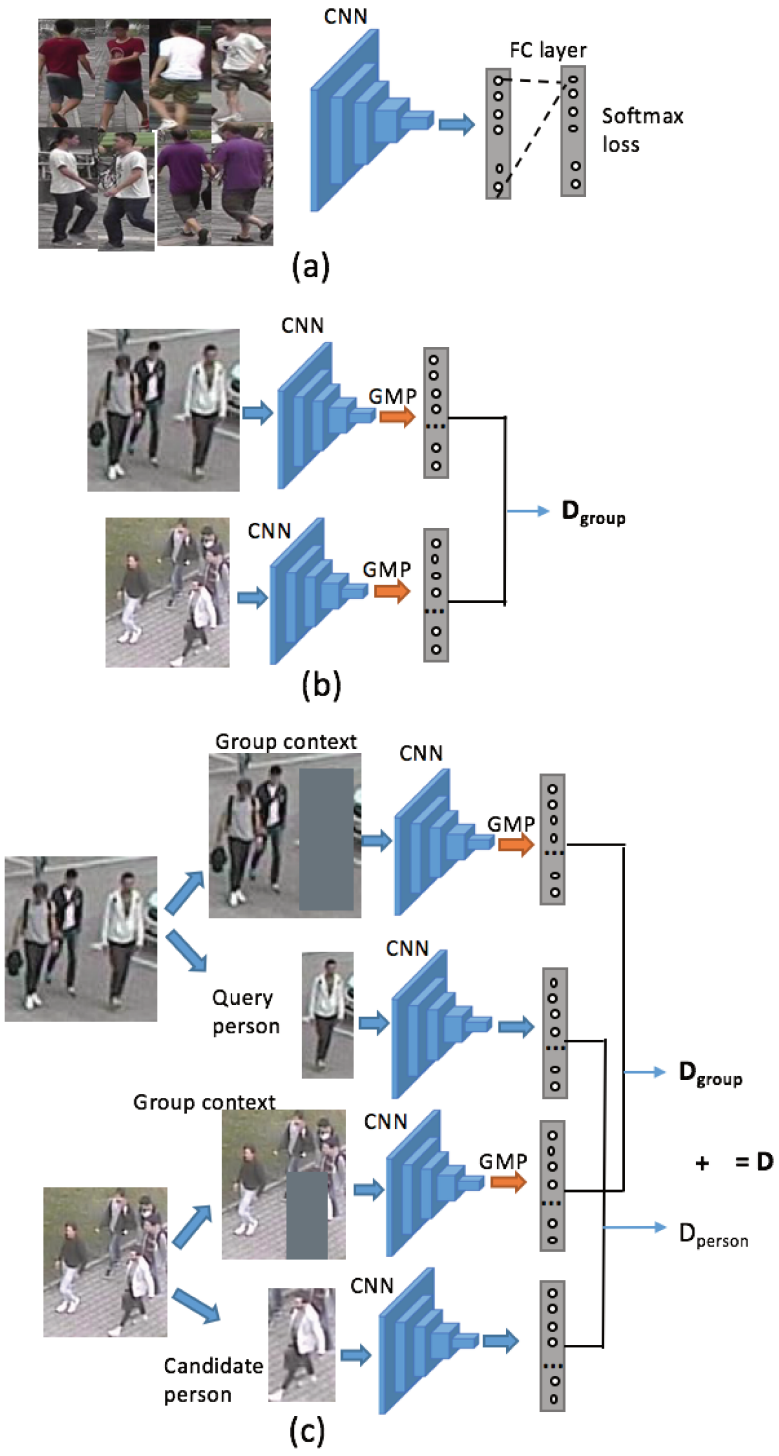


Fig. 6.2 Overview of our group association assisted re-identification method. (a) A CNN is first trained with person images. (b) The CNN with GMP is applied to group images to measure a group distance. (c) For person re-identification, group context distance and single-person distance are computed and summed to obtain the final distance.

6.3.1 Group association

In the first step, we train a neural network predicting the identities of the images, given an input image resized to 64x124. We can either train a CNN from scratch or use a a CNN pre-trained on ImageNet, like ResNet-50 [36]. The final fully-connected (FC) layer is replaced by another FC layer with an output dimension of N, with N being the number of identities in the training set.

Then, the CNN is fine-tuned in a supervised way, using images and identity labels from a separate person re-identification dataset. To this end, we minimise the following softmax cross-entropy loss on the given classification task:

$$E_{identification} = -\sum_{k=1}^{N} y_k log(P(y_k = 1|x)) \ , \tag{6.1}$$

$$\text{with} \quad P(y_j = 1|x) = \frac{e^{W_j^T x + b_j}}{\sum_{k=1}^{N} e^{W_k^T x + b_k}} \ , \tag{6.2}$$

where y is the one-hot-coded identity label, x is the input to the last fully-connected layer, W and b are weights and bias of the last fully-connected layer and $P(y_j = 1|x)$ is the predicted probability that the input x corresponds to identity j. We suppose that the embedding learned by using classification loss can be generalized and encode the features with which we can calculate a distance between arbitrary pedestrian images as siamese or triplet loss. And The intuition is that the resulting feature representation can be transferred to the task of group re-identification.

After training the model, we discard the FC layer and represent the activation map of the last convolutional layer as a set of K 2D feature channel responses $\mathscr{X} = \mathscr{X}_i, i = 1...K$, where $\mathscr{X}_i$ is the 2D map representing the responses of the i^{th} feature channel. A ReLU activation function is applied as a last step to guarantee that all elements are non-negative. A final location-invariant representation, called Maximum Activations of Convolutions (MAC) [108], is constructed by a spatial max-pooling over all locations concatenated in a K-dimensional vector:

$$f = [f_1...f_i...f_K]^\top, with f_i = \max_{x \in \mathscr{X}_i}(x) \ . \tag{6.3}$$

When applying the model for group association, a group image resized to 224×244 pixels is given as input (corresponding to a single-person size of roughly 64×128 pixels in the image). The distance between two images is measured with the cosine distance between the feature vectors produced as described above. This feature representation does not encode the location of the activations unlike the activations of fully connected layers, due to the

max-pooling operated over the whole last convolution layer feature map. It encodes the maximum "local" response of each of the convolutional filters. Thus, it offers translation invariance to the resulting representation.

6.3.2 Group-assisted person re-identification

In our setting, the input data is composed of group images with annotated individual identities and corresponding bounding boxes. As shown in Fig. 6.2, to explicitly capture both *person* and *group context* features, we divide the input image **I** into two input images to process them separately.

First, a query person image **P** is obtained from the raw group image by using the given annotated bounding box. Second, its group context image **G** is obtained from the raw group image by covering the query person image region with the pixels of the mean image colour. Then, these two images are given to the CNN model explained in Section 6.3.1. Two parallel branches of this network are employed to extract the feature embeddings for respectively the group context input image and the person image. The two branches are almost identical, except for the last layer, where the MAC feature representation is used for the (larger) group context image and the input vector of the discarded fully connected layer is used for the single-person image. As illustrated in Fig. 6.2, after processing the query image, the same procedure is applied to the gallery images in order to compute a distance measure on the two representation between query and candidate image (resulting in 4 feature vectors in total).The advantage of this method is that it can be easily combined with any CNN-based single person re-identification approach. For a given query and candidate image, the cosine distance is used to separately compute a group context distance D_{gr} between the two group images and a person distance D_{id} between the two single-person images.

$$D_{id}(P_i, P_j) = 1 - cos(F(f_{conv}(P_i)), F(f_{conv}(P_j))) \ . \tag{6.4}$$

$$D_{gr}(G_i, G_j) = 1 - cos(GMP(f_{conv}(G_i)), GMP(f_{conv}(G_j))) \ . \tag{6.5}$$

Where f_{conv} is the projection of the convolution layers of the network, GMP is the global max-pooling operation, F is the operations after the convolution layers in the CNN models. The final distance measure is simply the sum of these two distances:

$$D(I_i, I_j) = D_{id}(P_i, P_j) + D_{gr}(G_i, G_j) \ . \tag{6.6}$$

We use the sum of the group and single person distance as the combination. a more advanced way of combination could be considered such as a weighted sum or fusion by a fully connected layer in CNN. The reason why we did not perform them here is that there are only a few images having both identity and group annotation, for weighted loss it needs a number of images as validation to determine the weight. As well for fusion by fully-connected layer, a number of training images are needed. Thus in the approach, we retrain the simple sum of distance way which could also be easier to extend to other CNN-based re-identification approaches.

6.4 Experiments

6.4.1 Datasets

The **Market-1501 Dataset** [137] is used to train the CNN model. The dataset is split into 751 identities for training and 750 identities for test. In our experiments, we used only the training set of Market-1501 to train the CNN model.

The **Ilids-group Dataset** is extracted by Zheng et al. [138] from the i-LIDS MCTS dataset. It contains 274 images of 64 groups taken from airport surveillance cameras. Most of the groups have 4 images, either from different camera views or at different times. Some example images are shown in Fig. 6.3.

The **OGRE Dataset** [67] contains 1279 images of 39 groups acquired by three disjoint cameras pointing at a parking lot. This is a challenging dataset with many different viewpoints and self-occlusions. We manually annotated a subset of this dataset with 450 bounding boxes and 75 identities.

The Cumulative Match Curve (CMC) is employed as evaluation measure for both group association and person re-identification. For the group association test, we follow the test protocol in [138, 67]. That is, for each group, one randomly selected image is included in the gallery, all the remaining images form the probe set. The test is repeated 10 times, then the average scores are computed. For the person re-identification test, the images with person bounding boxes are used. We take each person bounding box as query image in turn, and the rest of the images as gallery set. The final result is the average CMC score over all queries.

6.4.2 Experimental setting

To show that our approach can be applied with different CNNs, we used ResNet-50 and a CNN composed of 5 convolution layers and 2 fully-connected layers (see Tab. 6.1, architecture is similar to the one used in Chapters 3 and 4, and we denote it Convnet-5 here). The

Fig. 6.3 Some example images from people group datasets: (a) OGRE dataset (b) Ilids-group dataset. One can note the challenging viewpoint variation and the resulting variations in relative positions if individuals as well as partial occlusions by objects, the image border or among individuals

weights pre-trained on the ImageNet dataset are used as initialization for Resnet-50, and the Convnet-5 is trained from scratch. For training, data augmentation is performed by randomly flipping the images and cropping central regions with random perturbation. Dropout is applied to the fully connected layers to reduce the risk of over-fitting. The optimization is performed by Stochastic Gradient Descent with a learning rate of 0.001 for Resnet and 0.005 for Convnet-5, a momentum of 0.9 and a batch size of 50.

layer	type	filter size	output size
C1	Convolution	5×5	128×48×32
P1	Max-Pooling	2×2	64×24×32
C2	Convolution	3×3	64×24×64
P2	Max-Pooling	2×2	32×12×64
C3	Convolution	3×3	32×12×128
P3	Max-Pooling	2×2	16×6×128
C4	Convolution	3×3	16×6×256
P5	Max-Pooling	2×2	8×3×256
C5	Convolution	3×3	8×3×512
fc1	Fully-connected	-	512
fc2	Fully-connected	-	751

Table 6.1 The architecture of Convnet-5.

6.4.3 Group association results

The comparison with the state-of-the-art method on the Ilids-group and OGRE dataset is shown in the Table 6.2. The shown CMC score is the success rate of the true match group in the top n results in the ranking list. We compared not only with the group association methods in [138, 67], but also with two encoding techniques, namely IFV [94] and VLAD [44], applied by [67] in group association as well our CNN models with global average pooling (GAP).

Our method based on Resnet-50 outperforms the best state-of-the-art method PREF [67] in terms of the Rank 1 score by a margin of 5.6% and 6.1% points on Ilids-group and OGRE datasets, respectively. The results of Convnet-5 are near the state-of-the-art result on OGRE dataset, but inferior to the state-of-the-art on Ilids group dataset. One possible reason is that the deeper architecture of Resnet-50 can extract more effective features and the model has a better generalization ability by using the Imagenet pretrained weight initialization, since this is a cross-task and cross-dataset test. The better performance of Convnet-5 on OGRE with respect to the state of the art can be explained by the fact that the model is trained from scratch on Market-1501, which is similar to OGRE scene but not to the airport scene in Ilids group. An important advantage of our method is that we do not use any pedestrian detection data or method as in PREF, IFV, VLAD.

Compared to the GAP-based model, using GMP increased the Rank 1 score on the two datasets by 3.9% and 2.9% points with Resnet-50 and by 3% and 1.1% points with Convnet-5, respectively. This demonstrates the benefit of the invariance property of the GMP for group association.

Method	Ilids-Group			OGRE		
	Rank 1	Rank 10	Rank 25	Rank 1	Rank 10	Rank 25
CRRRO+BRO [138]	22.5*	57.0*	**76.0***	-	-	-
IFV [94]	26.1	60.2	75.8	14.6	43.3	76.8
VLAD [44]	26.0	57.0	75.0	13.0	41.1	74.3
PREF [67]	31.1	60.3	75.5	15.1	41.6	75.8
Ours with Convnet-5+GAP	22.8	45.9	61.9	13.7	38.1	71.1
Ours with Convnet-5+GMP	25.8	44.8	61.5	14.8	37.8	71.8
Ours with Resnet-50+GAP	32.8	56.0	70.7	18.3	46.8	79.7
Ours with Resnet-50+GMP	**36.7**	**60.5**	73.7	**21.2**	**50.4**	**82.2**

Table 6.2 Comparison with group association state-of-the-art methods on the Ilids-group and OGRE dataset. CMC scores are used as evaluation measures. *: figures extracted from a curve.

6.4.4 Group-assisted person re-identification results

The result of person re-identification is shown in Table 6.3. The shown CMC score is the success rate of the true match person in the top n results in the ranking list. We compare the person re-identification results with some variants of our method. *Sum feature* and *Concatenate feature* represent variants that first sum or concatenate the single-person feature representation and the group feature representation and then compute the distance measure on these vectors. We compared also to a variant that retains the query or candidate person image in the group image without covering the corresponding region with the mean colour.

The results show that the method proposed in this chapter (i.e. covering the person in the group image and summing the person and group distance) achieved the best re-identification results on both CNN models. Covering the person image improves the Rank 1 score by 2.7% points and 1.5% points respectively with Resnet-50 and Convnet-5. Since some persons from the same group share very similar context, covering the query or candidate person can better discriminate persons in the same or similar group context.

Finally, the combined distance achieves better results than only using the single person distance and the group distance. Overall, our proposed method based on Resnet-50 and Convnet-5 increases respectively the result by 9.6% points and 2.1% points with respect to only using single-person images. The Convnet-5 gets less improvement due to the inferior group association performance. Anyway, This improvement clearly shows that group context has the ability to considerably reduce the appearance ambiguity and our method can be easily combined with any CNN-based single person re-identification approach. The Fig. 6.4 shows a example ranking result improved by using group context.

Variant	Resnet-50			Convnet-5		
	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10
Single person only	47.2	69.3	78.8	51.6	70.4	77.9
Group context only	26.2	57.2	66.3	12.7	42.3	53.0
Sum features	41.1	69.9	77.7	16.3	50.3	61.7
Concatenate features	51.9	**75.1**	81.1	16.4	52.6	61.9
Dist sum w/o mean img cover	54.1	73.7	80.8	52.2	**70.6**	78.1
Dist sum w/ mean img cover	**56.8**	73.7	**81.7**	**53.7**	70.4	**78.6**

Table 6.3 Person re-identification accuracy on CMC scores (in %) on the OGRE dataset.

Fig. 6.4 Ranking result example by using Resnet-50. The leftmost image is the query. The images with green tick are true matches. The rest images from left to right are in decreasing order of similarity (a) using only single person appearance (b) using single person appearance and group context. We can see that true matches advance in the ranking list with group context.

6.5 Conclusion

In this chapter, we presented an effective deep learning-based group association and group-assisted person re-identification approach. The deep group feature representation is extracted by a CNN and global max-pooling is applied to achieve location-invariance of individuals in group images. We also proposed a method improving single-person re-identification by incorporating the group context, defining a combined distance metric. This method can be combined with any CNN-based single person re-identification approach. We experimentally showed that our method outperforms the state-of-the-art in group association and that the deep group feature representation considerably enhances the person re-identification performance.

Chapter 7

Conclusion and Perspectives

In this thesis, we first presented the general context of automatic video analysis in a network of surveillance cameras with non-overlapping views, and then specifically studied the person re-identification problem from video streams. We presented several novel methods for image-based person re-identification using deep learning and we thoroughly evaluated these methods by experiments. We conclude our work by pointing out the different contributions and also some limitations of the proposed approaches. Finally, we discuss future perspectives of this work.

7.1 Contributions

At the very beginning, we discussed the issues and challenges in person re-identification as well as the limitation of existing approaches. Then, we proposed to apply specific deep learning models and training strategies to tackle with these issues regarding four different aspects, which constitute the contributions of this thesis.

CNN aggregating attribute and identity information

The first aspect involves semantic pedestrian attributes. Since detailed biometrics are often not available in images captured from surveillance cameras, attributes that are semantic human descriptions are used to enhance the lower-level feature representation and to better deal with pose and viewpoint variations. To exploit this mid-level descriptions which are complementary to appearance-only features, we proposed a method with two contributions: the fusion of low-level features and CNN features is introduced for attribute recognition and a CNN framework aggregating attribute and identity information.

Orientation specific CNN

The second aspect of this thesis concers the use of body pose information to alleviate the viewpoint variance problem. The novelty of our proposed method is that body orientation estimation and deep feature extraction are performed jointly in a unified CNN framework. An orientation gate module is introduced to steer the training of the orientation specific layers in the neural network. In that way, orientation-dependent signatures are calculated by combining learned orientation-specific feature representations.

Novel listwise loss function

The third aspect concerns the learning of an improved similarity metric. A novel loss function is proposed in order to learn a more robust person representation metric. The re-identification can be considered as a ranking problem within a list. The contribution of this loss is two-fold. Firstly, ranking lists instead of image pairs or triplets are used for training, thus integrating more explicitly the order of similarity and relations between sets of images. Secondly, a weighting is incoporated in the loss function based on the mean average precision and the Rank 1 score in order to directly optimize these global evaluation measures.

Exploiting group context for person re-identification with deep learning

Finally, as the last contribution, we proposed a deep learning based method using group context to improve person re-identification. This method aims to reduce the ambiguity from similar appearance (e.g. clothing) among a large gallery set. Location-invariance of members in the group is achieved by global max-pooling to better associate groups. And we propose to combine the single person appearance and the group context by summing the group distance and the single person distances. This method can be applied to any CNN based person re-identification method to exploit group context.

7.2 Limitations

We have demonstrated that the proposed methods have advantages with respect to existing methods and achieve state-of-the-art performance. But the proposed methods suffer from some limitations. This section presents these limitations and some of them could be eventually solved by the ideas of future work in the next section.

Data annotation

The first limitation is the data annotation. Both our attribute based and body orientation-based approach need additional data annotation of the pedestrian dataset. To train an effective deep neural network, a relative large amount of data is necessary. Due to the extremely extensive manual work for annotation, multi-label annotated datasets (with identity, body orientation, body parts labels etc.) are difficult to construct.

Computation

All the proposed methods are based on deep learning. The computation for the training step with backpropagation is more expensive than classical methods. In most cases, a powerful GPU is advisable for training. Especially, as we mentioned, the Rank-triplet loss needs a large batch size to show a superior performance, and more computation and memory resources are thus necessary. In the case of applications with real-time constraints and without GPU, a very deep network may be not suitable for inference.

Group constraint

Our group assisted re-identification method works only when people are walking in group and when the group context is available. So the method depends on people grouping and group determination methods. This raises some additional difficulties. For example, persons in the background can sometimes cause errors for people grouping. And this would further affect the person re-identification performance.

Partial occlusion/Bounding box misalignment

We did not propose explicit solutions for the occlusion and misalignment problem in the proposed methods. If the training set includes some images with such problems, we believe that the learned CNN models could deal with the light partial occlusion or misalignment. But heavily affected images could still easily lead to some errors. Besides improving pedestrian detection, more local information should maybe be considered to achieve the robustness.

7.3 Future work and perspectives

A very natural idea about future work is to put the 4 approaches in a united framework to see the contribution of each component and if the improvements could be complementary. Moreover, we also have some ideas about new directions to further improve our methods:

Exploit temporal information

All our proposed methods are based on images. However exploiting temporal information could further improve our methods. For example, the body orientation could be determined by the appearance and the walking direction. By selecting specific images in a video, we can avoid person images with non-visible attributes or occlusion. Also group association could be combined with the method that detects groups in video. Finally, recurrent neural network like LSTM could be applied to exploit spatial-temporal information for re-identification.

CNN architectures integrating attention

An attention mechanism consist in finding and processing a salient subset of features, frames in image sequence or spatial locations in one image subject to the task. The attention can be learned jointly with the feature representation. We might perform the pose and body part detection in an unsupervised way with attention directly integrated in a CNN , which could partially solve the data annotation problem. Paying attention to local salient regions could also be a solution to the misalignment and occlusion problem.

Mutual improvement with detection

In this thesis, only the re-identification problem is investigated. At present, detection and re-identification are relatively independent tasks. The mutual improvement between the two tasks could be interesting. For example, the pedestrian detection feature could help the body segmentation or orientation classification; The active feature map for person re-identification may help correct the misalignment errors; Detection and re-identification may be unified in multi-task learning framework, for example.

Semi-supervised person re-identification with deep learning

Most deep learning-based methods are trained in a supervised way. With automatic pedestrian detection and tracking methods, large amounts of data could be collected. It would be interesting if we can make use of these unlabeled data in an unsupervised or semi-supervised manner to enhance the existing methods.

References

[1] Ahmed, E., Jones, M., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3908–3916.

[2] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307.

[3] Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014). Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599.

[4] Bak, S., Corvee, E., Bremond, F., and Thonnat, M. (2010). Person re-identification using spatial covariance regions of human body parts. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*.

[5] Bak, S., Zaidenberg, S., Boulay, B., and Bremond, F. (2014). Improving person re-identification by viewpoint cues. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 175–180. IEEE.

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

[7] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

[8] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

[9] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In *ICML*, pages 89–96.

[10] Burges, C. J., Ragno, R., and Le, Q. V. (2007). Learning to rank with nonsmooth cost functions. In *NIPS*, pages 193–200.

[11] Cai, Y. and Medioni, G. (2014). Exploring context information for inter-camera multiple target tracking. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 761–768. IEEE.

[12] Cai, Y., Takala, V., and Pietikainen, M. (2010). Matching groups of people by covariance descriptor. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2744–2747. IEEE.

[13] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136.

[14] Chen, W., Chen, X., Zhang, J., and Huang, K. (2017a). Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

[15] Chen, W., Chen, X., Zhang, J., and Huang, K. (2017b). A multi-task deep network for person re-identification. In *AAAI*, pages 3988–3994.

[16] Chen, Y., Zhu, X., and Gong, S. (2017c). Person re-identification by deep learning multi-scale representations. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2590–2600.

[17] Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344.

[18] Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., and Murino, V. (2011). Custom pictorial structures for re-identification. In *Bmvc*, volume 1, page 6. Citeseer.

[19] Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE.

[20] Chum, O., Philbin, J., Sivic, J., Isard, M., and Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

[21] Deng, Y., Luo, P., Loy, C. C., and Tang, X. (2014). Pedestrian attribute recognition at far distance. In *Proc. of the ACM international conference on Multimedia*, pages 789–792. ACM.

[22] Deng, Y., Luo, P., Loy, C. C., and Tang, X. (2015). Learning to recognize pedestrian attribute. *arXiv preprint arXiv:1501.00901*.

[23] Dikmen, M., Akbas, E., Huang, T. S., and Ahuja, N. (2010). Pedestrian recognition with a learned metric. In *Asian conference on Computer vision*, pages 501–512. Springer.

[24] Ding, S., Lin, L., Wang, G., and Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003.

[25] Duan, K., Parikh, D., Crandall, D., and Grauman, K. (2012). Discovering localized attributes for fine-grained recognition. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3474–3481.

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

[27] Fogel, I. and Sagi, D. (1989). Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113.

[28] García, J., Martinel, N., Foresti, G. L., Gardel, A., and Micheloni, C. (2014). Person orientation and feature distances boost re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4618–4623. IEEE.

[29] Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person reidentification using spatiotemporal appearance. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1528–1535. IEEE.

[30] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[31] Gordo, A., Almazán, J., Revaud, J., and Larlus, D. (2016). Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257.

[32] Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. of International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3. Citeseer.

[33] Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 262–275.

[34] Haque, A., Alahi, A., and Fei-Fei, L. (2016). Recurrent attention models for depth-based person identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1229–1238. IEEE.

[35] Håstad, J. and Goldmann, M. (1991). On the power of small-depth threshold circuits. *Computational Complexity*, 1(2):113–129.

[36] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[37] Herbrich, R. (2000). Large margin rank boundaries for ordinal regression. *Advances in large margin classifiers*, pages 115–132.

[38] Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

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

[40] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

[41] Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025.

[42] Javed, O., Rasheed, Z., Shafique, K., and Shah, M. (2003). Tracking across multiple cameras with disjoint views. In *Proceedings of the Ninth IEEE International Conference on Computer Vision-Volume 2*, page 952. IEEE Computer Society.

[43] Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317.

[44] Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311.

[45] Kalantidis, Y., Mellina, C., and Osindero, S. (2016). Cross-dimensional weighting for aggregated deep convolutional features. In *European Conference on Computer Vision*, pages 685–701.

[46] Khamis, S., Kuo, C.-H., Singh, V. K., Shet, V. D., and Davis, L. S. (2014). Joint learning for attribute-consistent person re-identification. In *ECCV Workshops (3)*, pages 134–146.

[47] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICML*.

[48] Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2288–2295.

[49] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proc. of Advances in neural information processing systems (NIPS)*, pages 1097–1105.

[50] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *Proc. of the International Conference on Computer Vision (ICCV)*, pages 365–372. IEEE.

[51] Layne, R., Hospedales, T. M., and Gong, S. (2014). Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer.

[52] Layne, R., Hospedales, T. M., Gong, S., and Mary, Q. (2012). Person re-identification by attributes. In *Proc. of the British Machine Vision Conference (BMVC)*, volume 2, page 8.

[53] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[54] Li, A., Liu, L., and Yan, S. (2014a). Person re-identification by attribute-assisted clothes appearance. In *Person Re-Identification*, pages 119–138. Springer.

[55] Li, D., Chen, X., and Huang, K. (2015). Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. *Proc. of the Asian Conference on Pattern Recognition (ACPR)"*.

[56] Li, D., Chen, X., Zhang, Z., and Huang, K. (2017a). Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393.

[57] Li, H., Chen, J., Lu, H., and Chi, Z. (2017b). Cnn for saliency detection with low-level feature integration. *Neurocomputing*, 226:212–220.

[58] Li, S., Bak, S., Carr, P., and Wang, X. (2018a). Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378.

[59] Li, W. and Wang, X. (2013). Locally aligned feature transforms across views. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3594–3601.

[60] Li, W., Zhao, R., and Wang, X. (2012). Human reidentification with transferred metric learning. In *ACCV*.

[61] Li, W., Zhao, R., Xiao, T., and Wang, X. (2014b). Deepreid:deep filter pairing neural network for person re-identification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159.

[62] Li, W., Zhu, X., and Gong, S. (2017c). Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference on Artificial Intelligence*.

[63] Li, W., Zhu, X., and Gong, S. (2018b). Harmonious attention network for person re-identification. In *CVPR*, volume 1, page 2.

[64] Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206.

[65] Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., and Li, S. Z. (2010). Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1301–1306.

[66] Lin, Z. and Davis, L. S. (2008). Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *International symposium on visual computing*, pages 23–34. Springer.

[67] Lisanti, G., Martinel, N., Del Bimbo, A., and Foresti, G. L. (2017). Group re-identification via unsupervised transfer of sparse features encoding. In *Proc. of the IEEE International Conference on International Conference on Computer Vision (ICCV)*.

[68] Lisanti, G., Masi, I., and Del Bimbo, A. (2014). Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*, page 10. ACM.

[69] Liu, H., Feng, J., Qi, M., Jiang, J., and Yan, S. (2017). End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*.

[70] Liu, J., Kuipers, B., and Savarese, S. (2011). Recognizing human actions by attributes. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3344.

[71] Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., and Hu, J. (2018). Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108.

[72] Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., and Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3557.

[73] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee.

[74] Loy, C. C., Xiang, T., and Gong, S. (2009). Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1988–1995. IEEE.

[75] Lumini, A., Nanni, L., and Ghidoni, S. (2016). Deep featrues combined with hand-crafted features for face recognition. *International Journal of Computer Research*, 23(2):123.

[76] Ma, B., Su, Y., and Jurie, F. (2012). Local descriptors encoded by fisher vectors for person re-identification. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 413–422.

[77] Ma, L., Liu, H., Hu, L., Wang, C., and Sun, Q. (2016). Orientation driven bag of appearances for person re-identification. *arXiv preprint arXiv:1605.02464*.

[78] Ma, L., Yang, X., and Tao, D. (2014). Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670.

[79] Makris, D., Ellis, T., and Black, J. (2004). Bridging the gaps between cameras. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE.

[80] Man, J. and Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322.

[81] Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372.

[82] Matsukawa, T. and Suzuki, E. (2016). Person re-identification using cnn features learned from combination of attributes. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2428–2433. IEEE.

[83] McLaughlin, N., del Rincon, J. M., and Miller, P. C. (2017). Person reidentification using deep convnets with multitask learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):525–539.

[84] McLaughlin, N., Martinez del Rincon, J., and Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334.

[85] Mignon, A. and Jurie, F. (2012). Pcca: A new approach for distance learning from sparse pairwise constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2672. IEEE.

[86] Munaro, M., Basso, A., Fossati, A., Van Gool, L., and Menegatti, E. (2014a). 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 4512–4519. IEEE.

[87] Munaro, M., Fossati, A., Basso, A., Menegatti, E., and Van Gool, L. (2014b). One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*, pages 161–181. Springer.

[88] Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59.

[89] Pedagadi, S., Orwell, J., Velastin, S., and Boghossian, B. (2013). Local fisher discriminant analysis for pedestrian re-identification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325.

[90] Porikli, F. (2003). Inter-camera color calibration by correlation model function. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–133. IEEE.

[91] Prosser, B. J., Zheng, W.-S., Gong, S., Xiang, T., and Mary, Q. (2010). Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6.

[92] Radenović, F., Tolias, G., and Chum, O. (2016). Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*, pages 3–20.

[93] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533.

[94] Sánchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245.

[95] Sarfraz, M. S., Schumann, A., Wang, Y., and Stiefelhagen, R. (2017). Deep view-sensitive pedestrian attribute inference in an end-to-end model. In *Proc. of the British Machine Vision Conference (BMVC)*.

[96] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

[97] Schumann, A. and Stiefelhagen, R. (2017). Person re-identification by deep learning attribute-complementary information. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1435–1443. IEEE.

[98] Schwartz, W. R. and Davis, L. S. (2009). Learning discriminative appearance-based models using partial least squares. In *Conference on Computer Graphics and Image Processing (SIBGRAPI)*, pages 322–329. IEEE.

[99] Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., and Wang, J. (2015). Person re-identification with correspondence structure learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3200–3208.

[100] Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., and Li, S. Z. (2016). Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, pages 732–748. Springer.

[101] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[102] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

[103] Su, C., Zhang, S., Xing, J., Gao, W., and Tian, Q. (2016). Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491. Springer.

[104] Subramaniam, A., Chatterjee, M., and Mittal, A. (2016). Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, pages 2667–2675.

[105] Sudowe, P., Spitzer, H., and Leibe, B. (2015). Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95.

[106] Sun, Y., Zheng, L., Deng, W., and Wang, S. (2017). Svdnet for pedestrian retrieval. In *Proc. of the IEEE International Conference on International Conference on Computer Vision (ICCV)*.

[107] Tian, Y., Luo, P., Wang, X., and Tang, X. (2015). Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5087.

[108] Tolias, G., Sicre, R., and Jégou, H. (2016). Particular object retrieval with integral max-pooling of cnn activations. In *International Conference on Learning Representations*.

[109] Tuzel, O., Porikli, F., and Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In *European conference on computer vision*, pages 589–600. Springer.

[110] Ukita, N., Moriguchi, Y., and Hagita, N. (2016). People re-identification across non-overlapping cameras using group features. *Computer Vision and Image Understanding*, 144:228–236.

[111] Vaquero, D. A., Feris, R. S., Tran, D., Brown, L., Hampapur, A., and Turk, M. (2009). Attribute-based people search in surveillance environments. In *Proc. of Workshop on Applications of Computer Vision (WACV)*, pages 1–8. IEEE.

[112] Varior, R. R., Haloi, M., and Wang, G. (2016a). Gated siamese convolutional neural network architecture for human re-identification. In *Proc. of the IEEE International Conference on European Conference on Computer Vision (ECCV)*, pages 791–808. Springer.

[113] Varior, R. R., Shuai, B., Lu, J., Xu, D., and Wang, G. (2016b). A siamese long short-term memory architecture for human re-identification. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 135–153.

[114] Verma, A., Hebbalaguppe, R., Vig, L., Kumar, S., and Hassan, E. (2015). Pedestrian detection via mixture of cnn experts and thresholded aggregated channel features. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 163–171.

[115] Wang, F., Zuo, W., Lin, L., Zhang, D., and Zhang, L. (2016). Joint learning of single-image and cross-image representations for person re-identification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1288–1296.

[116] Wang, J., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y., et al. (2014a). Learning fine-grained image similarity with deep ranking. In *CVPR*.

[117] Wang, J., Wang, Z., Gao, C., Sang, N., and Huang, R. (2017). Deeplist: Learning deep features with adaptive listwise constraint for person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):513–524.

[118] Wang, T., Gong, S., Zhu, X., and Wang, S. (2014b). Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer.

[119] Wang, X., Doretto, G., Sebastian, T., Rittscher, J., and Tu, P. (2007). Shape and appearance context modeling. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

[120] Wei, L. and Shah, S. K. (2015). Subject centric group feature for person re-identification. In *CVPR Workshops*, pages 28–35.

[121] Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88.

[122] Wu, L., Shen, C., and Hengel, A. v. d. (2016a). Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*.

[123] Wu, S., Chen, Y.-C., Li, X., Wu, A.-C., You, J.-J., and Zheng, W.-S. (2016b). An enhanced deep feature representation for person re-identification. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE.

[124] Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008). Listwise approach to learning to rank: theory and algorithm. In *ICML*, pages 1192–1199.

[125] Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

[126] Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., and Yang, X. (2016). Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer.

[127] Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., and Li, S. Z. (2014). Salient color names for person re-identification. In *European conference on computer vision*, pages 536–551. Springer.

[128] Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Deep metric learning for person re-identification. In *Proc. of the IEEE International Conference on International Conference on Pattern Recognition*, pages 34–39.

[129] You, J., Wu, A., Li, X., and Zheng, W.-S. (2016). Top-push video-based person re-identification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1345–1353.

[130] Yu, R., Zhou, Z., Bai, S., and Bai, X. (2017). Divide and fuse: A re-ranking approach for person re-identification. In *Proc. of the British Machine Vision Conference (BMVC)*.

[131] Zhang, L., Xiang, T., and Gong, S. (2016a). Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248.

[132] Zhang, Y., Li, B., Lu, H., Irie, A., and Ruan, X. (2016b). Sample-specific svm learning for person re-identification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1278–1287.

[133] Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., and Tang, X. (2017a). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085.

[134] Zhao, L., Li, X., Wang, J., and Zhuang, Y. (2017b). Deeply-learned part-aligned representations for person re-identification. In *Proc. of the IEEE International Conference on International Conference on Computer Vision (ICCV)*.

[135] Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3586–3593. IEEE.

[136] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer.

[137] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*.

[138] Zheng, W.-S., Gong, S., and Xiang, T. (2009). Associating groups of people. In *BMVC*.

[139] Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 649–656. IEEE.

[140] Zheng, Z., Zheng, L., and Yang, Y. (2017a). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13.

[141] Zheng, Z., Zheng, L., and Yang, Y. (2017b). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. of the IEEE International Conference on International Conference on Computer Vision (ICCV)*.

[142] Zhong, Z., Zheng, L., Cao, D., and Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *Proc. of the International Conference on Computer Vision (ICCV)*.

[143] Zhu, J., Liao, S., Lei, Z., and Li, S. Z. (2014). Improve pedestrian attribute classification by weighted interactions from other attributes. In *Asian Conference on Computer Vision*, pages 545–557. Springer.

[144] Zhu, J., Liao, S., Lei, Z., and Li, S. Z. (2017). Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229.

[145] Zhu, J., Liao, S., Lei, Z., Yi, D., and Li, S. (2013). Pedestrian attribute classification in surveillance: Database and evaluation. In *Proc. of the International Conference on Computer Vision (ICCV) Workshops*, pages 331–338.

[146] Zhu, J., Liao, S., Yi, D., Lei, Z., and Li, S. Z. (2015). Multi-label cnn based pedestrian attribute learning for soft biometrics. In *Proc. of the International Conference on Biometrics(ICB)*, pages 535–540. IEEE.

Appendix A

Image Examples from Person Re-identification Datasets

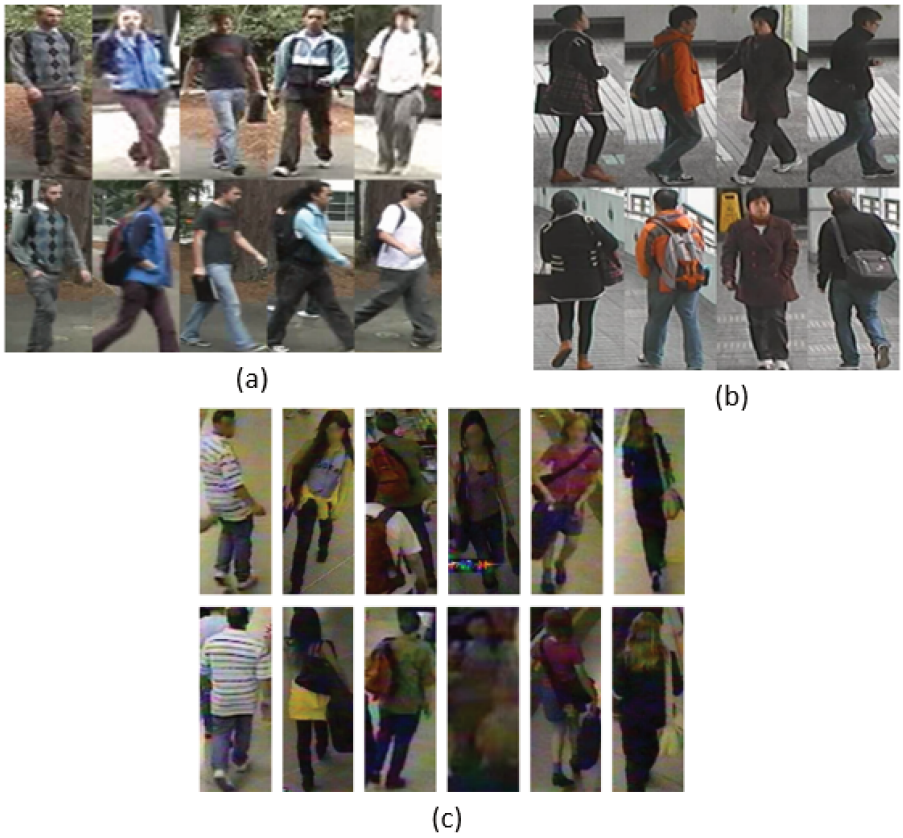


Fig. A.1 Single-shot person re-identification datasets: (a) VIPeR (b) CUHK01 (c) GRID.

Fig. A.2 Multi-shot person re-identification datasets: (a) CUHK03 (b) Market-1501 (c) DukeMTMC-Reid

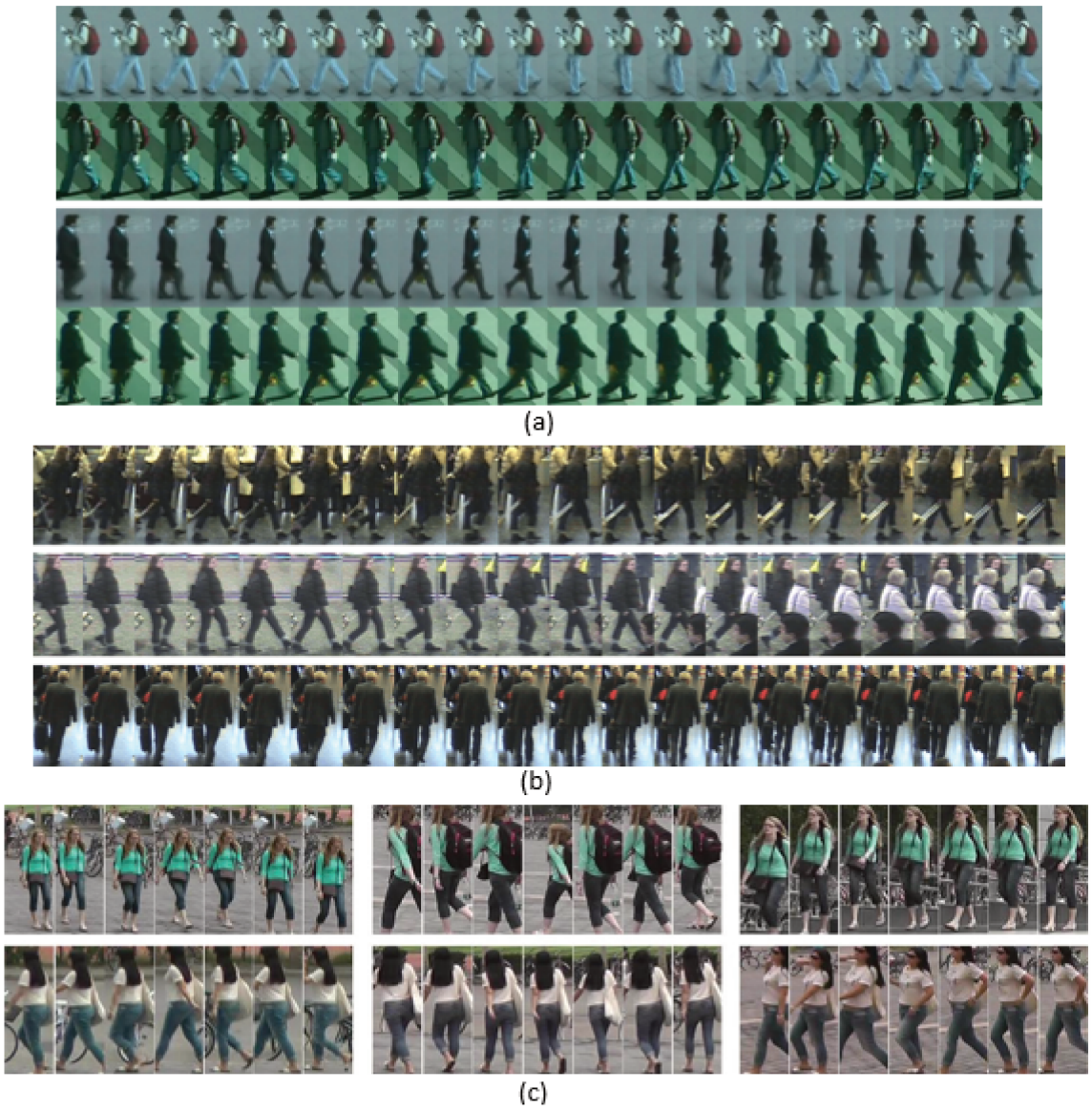


Fig. A.3 Video based person re-identification datasets: (a) PRID2011 (b) Ilids (c) MARS.

Appendix B

Complete Attribute Recognition Results on PETA

attribute	Accuracy (%)					Recall@FPR=0.1			AUC		
	MRFr2 [21]	DeepMar [55]	ikSVM [22]	mlcnn [144]	ours	ikSVM [22]	mlcnn [144]	ours	ikSVM [22]	ml-cnn [144]	ours
personalLess30	86.8	85.8	79.3	81.1	86.0	61.3	63.8	80.8	86.7	88.5	93.8
personalLess45	83.1	81.8	76.1	79.9	84.7	51.1	59.4	74.9	82.1	84.6	91.9
personalLess60	80.1	86.3	79.7	92.8	95.4	64.7	70.2	83.0	84.9	87.7	92.8
personalLarger60	93.8	94.8	96.1	97.6	98.9	89.1	90.7	94.6	95.3	94.9	96.8
carryingBackpack	70.5	82.6	76.4	84.3	85.5	46.2	58.4	70.2	84.5	85.2	91.9
carryingOther	73.0	77.3	76.2	80.9	85.7	38.6	46.9	65.1	74.1	77.7	88.4
lowerBodyCasual	78.2	84.9	85.5	90.5	92.1	53.7	56.2	76.1	85.6	87.5	93.1
upperBodyCasual	78.1	84.4	81.8	89.3	91.2	47.1	62.1	74.2	83.7	87.2	92.5
lowerBodyFormal	79.0	85.2	84.6	90.9	93.3	65.4	72.5	82.8	86.0	87.8	92.7
upperBodyFormal	78.7	85.1	87.1	91.1	93.4	62.4	70.5	83.4	85.2	87.6	92.9
accessoryHat	90.4	91.8	92.0	96.1	97.5	81.4	86.1	89.9	91.3	92.6	95.0
upperBodyJacket	72.2	79.2	88.8	92.3	94.7	53.9	53.4	77.4	83.3	81.0	92.1
lowerBodyJeans	81.0	85.7	78.6	83.1	87.6	57.2	67.6	83.2	85.0	87.7	94.5
footwearLeatherShoes	87.2	87.3	81.9	85.3	90.2	66.6	72.3	87.8	87.3	89.8	95.7
hairLong	80.1	88.9	79.3	88.1	91.3	56.0	76.5	88.3	84.2	90.6	95.6
personalMale	86.5	89.9	78.5	84.3	88.9	54.1	74.8	87.0	85.8	91.7	95.8
carryingMessengerBag	78.3	82.0	74.5	79.6	84.5	50.2	58.3	70.7	78.4	82.0	89.8
accessoryMuffler	93.7	96.1	94.8	97.2	98.8	90.7	88.4	93.6	95.1	94.5	96.2
accessoryNothing	82.7	85.8	78.9	86.1	89.0	35.4	52.6	71.5	81.8	86.1	92.1
carryingNothing	76.5	83.1	75.8	80.1	84.5	49.4	55.2	71.8	81.6	83.1	91.3

carryingPlasticBags	81.3	87.0	86.9	93.5	96.6	70.6	67.3	83.6	87.7	86.0	92.2
footwearShoes	78.4	80.0	72.3	75.8	80.8	46.9	52.8	68.3	79.2	81.6	89.4
upperBodyShortSleeve	75.8	87.5	83.1	88.1	90.7	68.2	69.2	86.2	89.9	89.2	94.5
footwearSneaker	75.0	78.7	78.0	81.8	85.7	45.5	52.0	73.0	83.3	83.2	92.0
lowerBodyTrousers	82.2	84.3	73.4	76.3	83.4	49.7	56.2	75.2	80.7	84.2	92.0
upperBodyTshirt	71.4	83.0	84.1	90.6	93.3	63.8	63.5	82.7	89.3	88.7	92.8
upperBodyOther	87.3	86.1	79.7	82.0	86.2	70.1	73.2	80.8	87.1	88.5	93.5
27 attributes average	80.8	85.4	81.6	86.6	90.0	58.9	65.6	79.9	85.2	87.0	93.0
upperBodyLogo	52.7	68.4			95.5			63.5			85.2
upperBodyPlaid	65.2	81.1			97.7			73.2			88.6
footwearSandals	52.2	67.3			97.6			82.1			89.4
lowerBodyShorts	65.2	80.4			96.8			85.5			93.3
lowerBodyShortSkirt	69.6	82.2			96.8			89.0			94.9
upperBodyThinStripes	51.9	66.5			98.2			61.0			81.6
accessorySunglasses	53.5	69.9			96.3			72.0			87.5
upperBodyVNeck	53.3	69.8			98.9			78.4			87.8
35 attributes average	75.6	82.6			91.7			78.9			92.0
footwearBlack			74.3	76.0	82.3	50.4	57.2	73.5	81.4	84.1	91.6
footwearBrown			82.4	92.1	95.4	63.2	65.8	82.5	84.7	85.3	93.3
footwearGrey			79.3	87.1	89.8	48.8	50.8	70.6	80.9	80.9	90.7
footwearWhite			79.0	85.9	89.3	52.3	62.7	78.8	83.8	86.2	93.0
hairBlack			84.8	87.8	90.9	75.5	81.0	89.6	91.9	93.6	96.4
hairBrown			84.2	89.6	92.4	72.2	77.4	89.3	89.8	91.3	95.7
hairGrey			92.2	95.3	96.8	71.1	74.9	87.6	87.8	89.4	94.7
hairShort			77.6	86.9	90.0	52.5	69.7	85.9	82.9	89.8	95.4
lowerBodyBlack			84.5	83.9	90.3	75.6	71.2	90.2	91.8	90.8	96.8
lowerBodyBlue			85.6	88.6	94.4	72.4	77.3	92.8	90.2	90.8	97.4
lowerBodyGrey			78.7	82.1	86.6	54.5	53.4	77.7	84.1	82.8	93.1
upperBodyBlack			85.8	86.2	90.8	81.4	80.1	91.0	93.2	93.1	96.9
upperBodyBlue			92.8	94.5	96.9	80.6	76.2	94.9	93.0	90.9	97.5
upperBodyBrown			89.4	93.3	96.5	72.1	68.6	89.1	88.9	87.6	95.8
upperBodyGrey			82.4	84.4	89.6	60.4	55.3	80.2	85.6	83.0	93.8
upperBodyRed			95.6	96.3	97.8	90.9	86.8	95.7	96.6	94.7	97.2
upperBodyWhite			87.0	88.8	92.8	76.2	75.3	91.4	92.3	91.2	96.5
upperBodyLongSleeve			84.8	87.9	90.2	76.5	74.3	88.2	90.9	90.0	94.9
45 attribute average			82.8	87.2	90.7	62.6	67.3	82.3	86.4	87.7	93.8
53 attributes average					91.7			81.3			93.0

Table B.1 Attribute recognition results on PETA

Appendix C

French Summary

Introduction

De nos jours, un grand nombre de caméras sont installées dans des lieux privés et publics pour faire face à l'augmentation de la délinquance et de la criminalité. L'analyse automatique de vidéos de surveillance est un domaine de recherche important et concurrentiel qui exploite une grande quantité de données produites par les caméras de surveillance. Dans ce contexte, la ré-identification de personnes dans les vidéos constitue l'un des enjeux importants. L'objectif de la ré-identification est de retrouver la même personne dans des vidéos enregistrés dans différentes zones d'un même lieu public durant une période de temps réduite. Par ailleurs, la ré-identification est également nécessaire dans des applications comme l'interaction humain-machine, ou l'indexation de contenu de vidéo, etc.

La ré-identification est une problématique difficile. L'apparence de la personne subit des variations significatives tels que la variation de poses, la variation de point de vue et la variation de l'éclairage. La résolution de la vidéo est généralement basse et des occultations sur les personnes sont fréquentes. Les visages ne sont souvent pas visibles. Les méthodes basées sur la biométrie comme la reconnaissance faciale ne sont donc pas applicables.

État de l'art (voir Chapter 2)

Etant donnée une image de requête, pour trouver la bonne correspondance de cet individu dans un grand ensemble d'images de galerie, il faut prendre en compte deux problèmes. D'abord, il est nécessaire d'avoir une représentation caractéristique des images de requête et des images de galerie. Deuxièmement, une métrique de distance est nécessaire pour déterminer si une image de requête et une image de galerie appartiennent à la même classe d'individus. Les travaux antérieurs consistent soit à construire une représentation caractéristique robuste, soit à chercher une meilleure métrique de similarité.

Un grand nombre de travaux ont été proposés pour l'extraction de caractéristiques. Gray et Tao [33] ont proposé d'utiliser Adaboost pour sélectionner les bonnes caractéristiques dans un ensemble de caractéristiques de couleurs et de textures. Farenzena et al. [26] ont proposé la méthode Symmetry-Driven Accumulation of Local Features (SDALF) dans laquelle la symétrie et l'asymétrie de l'image de personne sont prises en compte pour traiter le problème de point de vue. Ma et al. [76] ont encodé les caractéristiques locales comme un vecteur de Fisher pour créer une représentation globale d'une image. Cheng et al. [18] ont appliqué l'approche de "pictorial structure" pour localiser les parties du corps et extraire les couleurs. Liao et al. [64] ont appliqué une normalisation de couleur basée sur l'algorithme de Retinex pour produire une image avec des couleurs plus cohérentes. Les caractéristiques de couleurs et de textures sont ensuite extraites dans une fenêtre glissante. Pour avoir une invariance horizontale dans l'image, on garde seulement la valeur maximale de chaque dimension du vecteur caractéristique parmi les fenêtres d'une même bande horizontale.

Dans l'étape de mesure de similarité, pour la ré-identification de personne, la distance de Mahalanobis est la plus utilisée. Dans ce cas, nous cherchons à apprendre un sous-espace qui est défini par une matrice de projection W dans lequel les distances entre les exemples (y_i et y_j) reflètent mieux les similarités entre les personnes, c'est-à-dire $||y_i - y_j||_2 = (x_i - x_j)^T A(x_i - x_j)$ où $A = W^T W$. Li et al. [59] ont proposé d'apprendre une fonction Locally-Adaptive Decision (LADF) qui pourrait être considérée comme une fusion d'une distance métrique et d'un seuillage localement adapté. Koestinger et al. [48] ont présenté une métrique appelée KISSME (keep it simple and straight forward metric). La décision "une paire est similaire" est formulée comme un test de ratio de ressemblance. La distance est basée sur la différence entre l'inverse de la matrice de covariance pour les paires similaires et celle des paires dissimilaires. Liao et al. [64] ont proposé d'apprendre un sous-espace dans lequel la variance des paires similaires est minimisée et celle des paires dissimilaires est maximisée, et puis d'appliquer la métrique KISSME dans cet espace réduit.

Suivant le grand succès de l'apprentissage profond dans le domaine de vision par ordinateur, des méthodes basées sur les réseaux de neurones à convolution ont été proposées pour la ré-identification de personnes. Yi et al. [128] ont construit un réseau de neurones siamois à convolution. L'image est découpée en trois parties horizontales. Chaque partie est associée à un réseau de neurones à convolution. A la fin, les trois parties sont fusionnées au niveau de leurs scores. DeepReID [61] ont proposé une nouvelle architecture de réseau Filter Pairing Neural Network (FPNN). Ils ont utilisé une couche de "patch matching" pour modéliser le déplacement des parties du corps. Ahmed et al. [1] ont proposé un réseau qui a une paire d'images en entrée et un score de similarité en sortie. Dans leur réseau, ils ont calculé la différence de cartes caractéristiques pour capturer la relation locale entre deux

images d'entrée. Cheng et al. [17] ont proposé un réseau en triplet basé à la fois sur le corps entier et les parties du corps avec une fonction coût améliorée.

Les méthodes basées sur l'apprentissage profond ont montré une performance supérieure sur les grandes bases de données de ré-identification. Mais les méthodes classiques aves l'extraction de caractéristiques et l'apprentissage de métrique marchent mieux quand la taille de la base d'apprentissage est réduite.

Ré-identification de personne avec les attributs (voir Chapter 3)

Les attributs sont des propriétés sémantiques et observables dans des images. Une idée principale de cette thèse est de faciliter la ré-identification avec la prise en compte des attributs sémantiques de la personne : vêtements, objets portés, chapeau, lunettes, personnes accompagnantes, etc. L'utilisation de ce type d'attributs présente quelques avantages. L'attribut est plus invariant contre les variations des points de vue que les caractéristiques visuelles de bas niveau. L'autre avantage est de rendre le modèle intelligible par l'humain. Cela pourrait permettre de chercher une personne sans image mais avec seulement une description textuelle d'humain. Donc nous proposons un système de ré-identification basé sur CNN et assisté par les attributs.

Dans un premier temps, notre méthode effectue l'apprentissage des attributs. L'architecture du réseau de neurones convolutif est constituée de plusieurs couches de « convolution-pooling ». Et puis, les cartes de caractéristiques sont découpées verticalement en 4 parties. Sur chaque partie, différents noyaux de convolutions sont appliqués afin que le réseau apprenne les caractéristiques spécifiques à chaque partie du corps.

En plus, notre approche combine les caractéristiques CNN avec une caractéristique de bas niveau LOMO de sorte de régler le problème de grande variation d'apparence visuelle et de location des attributs aux changements de pose ou point de vue. Une fonction de perte « multi-label » est employée pour l'entrainement. Comme les nombres d'exemples positifs (présence des attributs) et négatifs (absence des attributs) sont généralement déséquilibrés, un terme de pondération est introduit dans la fonction de perte.

Nos expériences sur les trois benchmarks publics sur la reconnaissance d'attributs, montrent que la méthode proposée améliore l'état de l'art. De plus, en combinant les attributs et les caractéristiques de bas niveau, les résultats s'améliorent, montrant ainsi la complémentarité des deux types de caractéristiques. Pour les bases de données de taille limitée, nous avons proposé de pré-entrainer le modèle en architecture triplet avec des données de ré-identification pour éviter le problème de sur-apprentissage. Finalement, l'approche surpasse largement le résultat de l'état de l'art sur trois benchmark publics.

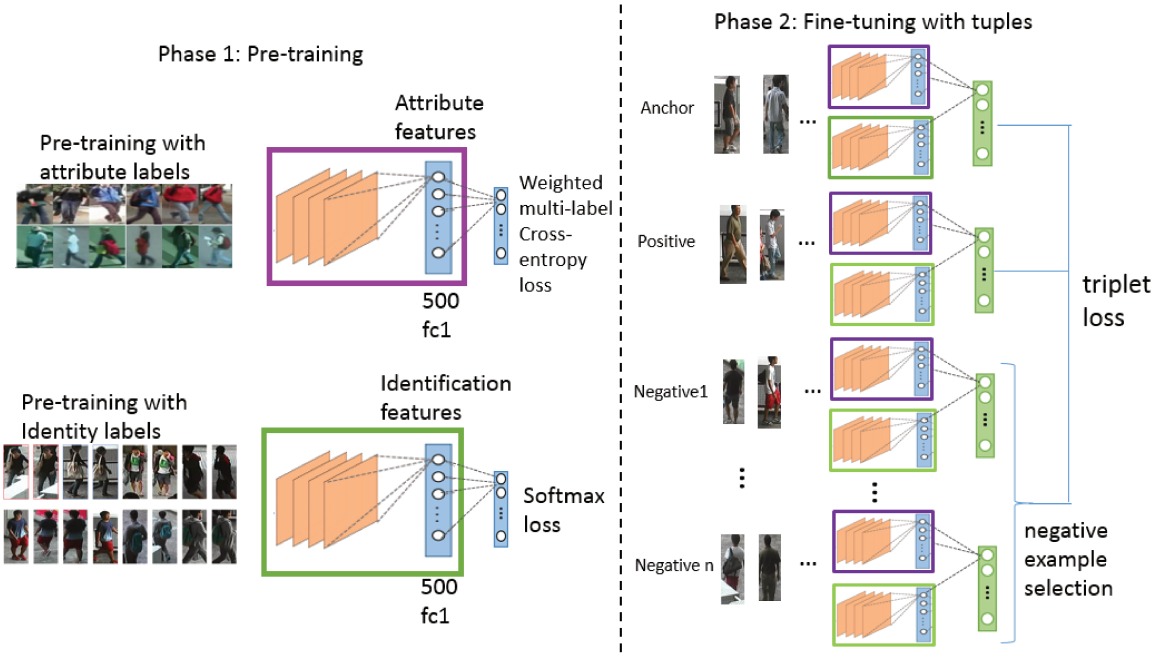


Fig. C.1 Aperçu de la ré-identification de personne avec attributs

Enfin, nous combinons les représentations des attributs appris et les représentations «caractéristiques d'identification» apprises par un autre CNN avec la fonction de perte de «classification softmax» (voir Fig. C.1). Tous les deux CNN sont appris de manière supervisée avec respectivement des «labels» attribut et identité. Pour intégrer les deux sous réseaux dans le même système, les couches de « feature embeddings » (la couche qui est devant la couche de sortie) de ces deux sous-réseaux sont concaténées et sont liées à une couche complètement connectée. L'idée est de combiner l'information d'identification et l'information d'attributs pour caractériser une personne et la ré-identifier. Nous faisons apprendre cette dernière couche complètement connectée dans une structure de triplet pour l'apprentissage. Le résultat montre que la prise en compte des attributs améliore la performance et nous obtenons un résultat équivalant à l'état de l'art.

Ré-identification de personne avec CNN spécifique à orientation (voir Chapter 4)

Un des plus grands défis pour la ré-identification de personnes est la variation de point de vue. Les images avec la même orientation du corps contiennent des silhouettes similaires. Par contre, les images d'une même personne présentent souvent des différences significatives sous les différentes orientations du corps. A notre connaissance, les méthodes existantes de ré-identification ne font pas de distinction dans l'orientation de la prise de vue et supposent que les images de personnes viennent d'un seul et unique domaine.

Dans notre approche, nous proposons de considérer que les images de piétons avec les différentes orientations sont de différents domaines. Nous proposons un réseau de neurones à convolution spécifique à chaque orientation qui effectue conjointement la régression d'orientation du corps et extrait une représentation profonde spécifique à orientation pour la ré-identification. L'aperçu de l'approche est montré dans la figure C.2. Le CNN est composé de deux branches. Dans la branche de ré-identification, il y a 3 couches de convolutions consécutives suivies par 4 couches complètement connectées (« fc layer ») en parallèle qui correspondent aux 4 différentes orientations : gauche, droit, frontal et arrière. L'idée est que le réseau de neurones apprend différentes projections de différents domaines d'orientation dans un espace caractéristique en commun.

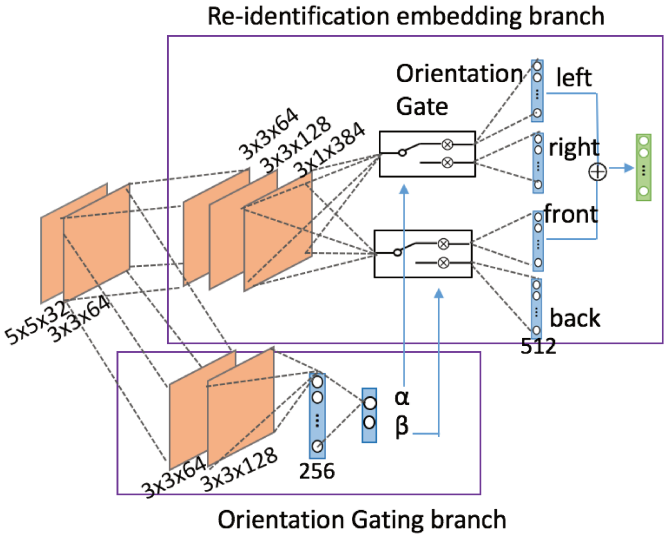


Fig. C.2 Aperçu du réseau convolutif spécifique à orientation

La branche de la régression d'orientation estime un vecteur de deux dimensions représentant l'angle de projection sur l'axe gauche-droite et l'axe frontal-derrière. Basé sur ce vecteur, la représentation caractéristique spécifique à l'orientation est construite en sélectionnant et pondérant la composant gauche ou droite et la composant frontal ou arrière.

L'apprentissage est effectué en deux étapes. Dans un premier temps, un apprentissage multi-tâche est réalisé. Deux fonctions de perte sont calculées. La perte euclidienne est utilisée pour la régression d'orientation et la perte de classification softmax est utilisée pour l'identification. La fonction de perte finale est une somme pondérée de ces deux pertes. Ensuite, pour spécialiser les 4 branches spécifiques à orientation, la dernière couche complètement connectée est dupliquée dans la branche ré-identification. La sélection et la pondération des 4 couches sont conduites par l'orientation estimée. L'apprentissage se fait dans une structure triplet.

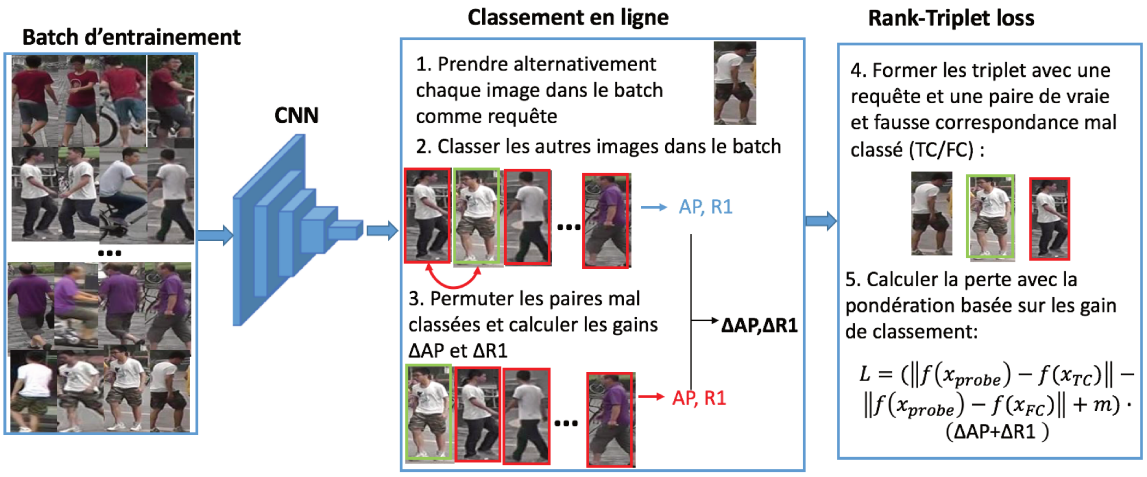


Fig. C.3 Aperçu de la fonction de perte Rank-Triplet

La représentation combinée des caractéristiques des différentes orientations pour obtenir une représentation robuste à variation de pointe de vue. Nous montrons expérimentalement que la prise en compte des orientations améliore la performance de la ré-identification sur les bases de données CUHK01 et Market.

Ré-identification de personne avec la fonction de perte Rank-Triplet (voir Chapter 5)

Dans la littérature, pour entrainer un réseau de neurones pour la ré-identification, les différentes fonctions de perte sont proposées comme la fonction de perte contrastive et celle de triplet. Contrairement à ces fonctions de perte existantes, nous proposons une fonction de perte appelée ''Rank-Triplet loss" basée sur une liste. Etant donnée une requête, la fonction est calculée sur le classement prédit par notre modèle et sur le classement de vérité terrain.

En plus, notre fonction de perte prend en compte directement les mesures d'évaluation. Comme montre la figure C.3, pendant l'apprentissage, chaque image dans le batch d'entrainement est alternativement utilisée comme une requête et les restes comme l'ensemble de galerie. Ensuite, les triplets sont formés par l'image de requête et une paire de vraie et fausse correspondances mal-classées. Tous les possible triplets sont formés. Et puis, nous estimons les importances des triplets en utilisant le gain sur les mesures d'évaluation pour corriger le classement de vraie et fausse correspondances dans les triplets. Donc les mesures d'évaluation de ré-identification : AP (average precision) and R1(Rang 1) sont calculées pour chaque requête. Puis, nous recalculons AP et R1 en permutant les positions de vraie et fausse correspondances dans la liste de classements. Donc nous obtenons une amélioration d'AP et de R1. La perte pour chaque triplet est pondérée par l'amélioration des mesure d'évaluation.

Cette pondération utilise mieux les triplets difficiles qui contribuent plus à l'amélioration de classement de liste. En même temps, l'utilisation de tous les triplets permet de stabiliser la procédure d'entrainement. Si la méthode se limitait à n'utiliser que des exemples les plus difficiles, cela pourrait mener à un mauvais minima local pendant l'apprentissage.

Ré-identification de personne avec le contexte de groupe (voir Chapter 6)

La plupart des approches existantes utilise seulement l'apparence visuelle d'une seule personne pour la ré-identification. Mais ceci pourrait entrainer des ambiguïtés quand les personnes sont très similaires (type ou couleur de vêtement par exemple). Ce problème devient plus grave s'il y a un grand nombre de candidats dans l'ensemble de galerie. Pour résoudre ce problème, l'information contextuelle des piétons pourrait être utilisée. Dans le cas réel, les gens marchent souvent en groupe de personnes. Donc, le contexte de groupe peut servir comme une information supplémentaire pour aider à déterminer si deux images avec les vêtements similaires viennent de la même personne.

Nous proposons d'extraire une représentation de groupe en utilisant le réseau de neurones à convolution profond. D'abord, nous faisons apprendre un modèle avec les données de ré-identification de personne. Puis nous le transférons au problème de l'appariement de groupe qui n'est pas une opération simple. En effet, le nombre de personnes et leurs positions relatives pourraient varier dans le temps et suivant la position de la caméra. Pour résoudre ce problème, nous appliquons une opération de ''Global max-pooling" sur les activations des convolutions pour avoir une invariance de translation dans la représentation résultante. Grâce à ''Global max-pooling", cette représentation caractéristique n'encode pas la location des activations contrairement à un modèle à couche complètement connectée. Autrement dit, elle encode la réponse locale maximale de chaque filtre convolutif.

En outre, nous combinons l'apparence de personne et le contexte de groupe pour améliorer la performance de la ré-identification (voir Fig. C.4). Pour ce faire, d'abord la distance basée sur l'apparence de la seule personne est calculée avec la représentation apprise sur la base de donnée de ré-identification. Ensuite, la distance de groupe est calculée avec le ''Global max-pooling" en couvrant dans l'image, la personne en requête avec des pixels de couleur moyenne. La distance finale est donc la somme des deux distances. L'avantage de notre approche est qu'il est facile de combiner avec n'importe quel modèle CNN. Nous montrons avec les expériences que la prise en compte de contexte de groupe améliore la performance de ré-identification.

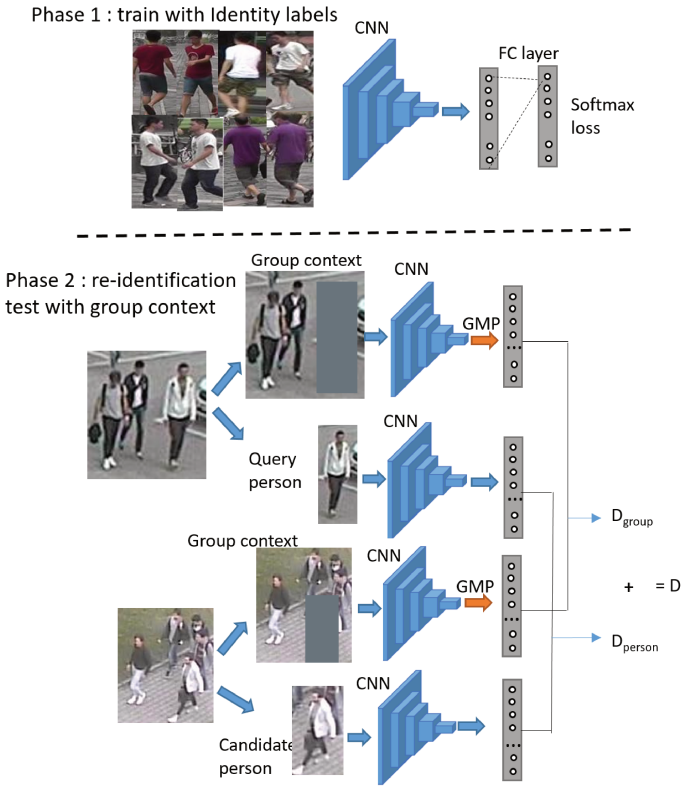


Fig. C.4 Aperçu de la ré-identification avec le contexte de groupe

Conclusion

Nous avons proposé 4 différentes approches basées sur les réseaux de neurones à convolution de différents aspects : attributs de piétons, orientation de corps, fonction de perte et le contexte de groupe. Chacune de ces méthodes améliore la performance de ré-identification. Comme possibles perspectives, nous pourrons envisager d'exploiter l'information temporelle dans les vidéos, de mettre en place le mécanisme de «l'attention» dans l'architecture du réseau (en conduisant par exemple, un apprentissage pondéré en fonction de la saillance locale) et d'effectuer l'apprentissage profond de manière semi-supervisée.

UNIVERSITÉ DE LYON

INSA

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : CHEN
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 12/10/2018

Prénoms : Yiqiang

TITRE : Ré-identification de personnes dans des images par apprentissage automatique

NATURE : Doctorat

Numéro d'ordre : 2018LYSEI074

Ecole doctorale : Infomaths

Spécialité : Informatique

RESUME :
La vidéosurveillance est d'une grande valeur pour la sécurité publique. En tant que l'un des plus importantes applications de vidéosurveillance, la ré-identification de personnes est définie comme le problème de l'identification d'individus dans des images captées par différentes caméras de surveillance à champs non-recouvrants. Cependant, cette tâche est difficile à cause d'une série de défis liés à l'apparence de la personne, tels que les variations de poses, de point de vue et de l'éclairage etc.

Pour régler ces différents problèmes, dans cette thèse, nous proposons plusieurs approches basées sur l'apprentissage profond de sorte d'améliorer de différentes manières la performance de ré-identification. Dans la première approche, nous utilisons les attributs des piétons tels que genre, accessoires et vêtements. Nous proposons un système basé sur un réseau de neurones à convolution(CNN) qui est composé de deux branches : une pour la classification d'identité et l'autre pour la reconnaissance d'attributs. Nous fusionnons ensuite ces deux branches pour la ré-identification. Deuxièmement, nous proposons un CNN prenant en compte différentes orientations du corps humain. Le système fait une estimation de l'orientation et, de plus, combine les caractéristiques de différentes orientations extraites pour être plus robuste au changement de point de vue. Comme troisième contribution de cette thèse, nous proposons une nouvelle fonction de coût basée sur une liste d'exemples. Elle introduit une pondération basée sur le désordre du classement et permet d'optimiser directement les mesures d'évaluation. Enfin, pour un groupe de personnes, nous proposons d'extraire une représentation de caractéristiques visuelles invariante à la position d'un individu dans une image de group. Cette prise en compte de contexte de groupe réduit ainsi l'ambigüité de ré-identification.

Pour chacune de ces quatre contributions, nous avons effectué de nombreuses expériences sur les différentes bases de données publiques pour montrer l'efficacité des approches proposées.

MOTS-CLÉS : vision par ordinateur, apprentissage profond, ré-identification de personne

Laboratoire (s) de recherche : LIRIS

Directeur de thèse: Atilla BASKURT

Président de jury :

Composition du jury :

BENOIS-PINEAU, Jenny	Prof.	Université de Bordeaux	Rapporteuse
BREMOND, François	DR	INRIA	Rapporteur
THOME, Nicolas	Prof.	CNAM	Examinateur
ACHARD, Catherine	MCF-HDR	Sorbonne Université	Examinatrice
DUFOR, Jean-Yves	Dr. Ing.	Thales ThereSIS Lab	Examinateur
BASKURT, Atilla	Prof.	INSA-LYON	Directeur de thèse
DUFFNER, Stefan	MCF	INSA-LYON	Co-directeur de thèse