PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
DR. YAHIA FARES UNIVERSITY OF MEDEA



Faculty of **T**echnology

Department of **E**lectrical **E**ngeneering

Master's End of Study Project

Branch: Telecommunications.
Option: Telecommunication Systems.

Person Re-identification Based CNN and Score Normalization

Presented by:

BENZAIBAK Afaf.

BOUDELLAL Chahrazed.

Proposed and supervised by:

Dr. CHOUCHANE Ammar.

Class 2020/2021

Dedication

To all my Family.

To all my Friends.

To all oppressed people who cannot voice themselves.

BENZAIBAK Afaf

I dedicate this humble thesis to my dearest people who supported me along my academic path. I owe my success to their sacrifice:

To my beloved parents, who bestowed me with all forms of love and support therefore no form of thanks or gratitude is enough to do them justice. May god protect them and bless them with good health and long life. I love you to the moon and back.

To my dearest Nihel Rayen twin, she's my other self, my all, my favorite soul, my confident and my beloved sister.

To my dear brother Anis, my right hand. I love so much.

To my grandparents.

To my whole respectful family BOUDELLAL and BENMOUSSA, (my lovely aunts and dear uncles, especially my closest Malika).

To those who shared this life experience with me, not forgetting the little birds of the family.

BOUDELLAL Chahrazed

Acknowledgments

We would like to thank "God" the Almighty for giving us the courage, the strength and the will to make this work.

The presentation of this work has given us the opportunity to express our gratitude to the people who have helped and advised us in an invaluable way during the preparation of this dissertation.

We would like to express our highest esteem to Mr. CHOUCHANE Ammar our supervisor for his assistance, encouragement and guidance he has given.

Many thanks go to the teachers and staff of Electrical Engineering department of Medea University.

We would like to thank again all the people who contributed directly or indirectly to the realization of this brief.

We hope that this brief will be useful to all future ones.

ملخّص

لإعادة تحديد هوية الأشخاص (Re-ID) أهمية كبيرة في مجال الأمن الذكي. فهي مشكلة تحديد الأشخاص على الصور الملتقطة بواسطة كاميرات مراقبة مختلفة دون تداخل مجالات الرؤية. في بيئة العالم الحقيقي ، تعد إعادة تعريف الأشخاص مهمة صعبة بسبب الاختلافات في ظروف الإضاءة وزوايا الرؤية والوضعية و الانسدادات. هناك مشكلتان أساسيتان ضروريتان لإعادة تعريف الأشخاص تمثيل الخصائص والتعلم المتري. يجب أن يكون التمثيل الفعال للخصائص قويًا للإضاءة والتغييرات في المنظور، ويجب تعلم مقياس مميز لمطابقة الصور المختلفة للأشخاص. في هذا العمل، سنستخدم تمثيلًا فعالًا للميزات يسمى الشبكة العصبية التلافيفية (CNN) لاكتشاف الأنماط والأشخاص، وطريقة تعلم متري تسمى التحليل التربيعي المتقاطع (XQDA) مع المسافة mahalanobis لتحسين نتائج التصنيف. بعد ذلك، سنطبق تطبيع النتيجة على مسافة mahalanobis الناتجة على أربعة مجموعات بيانات صعبة لإعادة تحديد هوية الأشخاص ,VIPeRCUHK01 و PRID450S و GRID. توضح النتائج التجريبية على مجموعة البيانات الأربعة المذكورة فعالية الوظيفة المقترحة.

الكلمات المفتاحية: إعادة تحديد هوية الشخص, ,CNN,XQDA تطبيع النتيجة, mahalanobis.

Abstract

Person re-identification (Re-ID) is an important task in the field of intelligent security. It's defined as the problem of identifying people across images that have been captured by different surveillance cameras without overlapping fields of view. In a real world environment, Re-ID is a challenging task due to variations in lighting conditions, viewing angles, pose and occlusions. Two fundamental problems are critical for person re-identification, feature representation and metric learning. An effective feature representation should be robust to illumination and viewpoint changes, and a discriminant metric should be learned to match various person images. In this work we are going to use an effective feature representation called convolutional neural network (CNN) to detect patterns to distinct persons, and metric learning method called cross-view quadratic discriminant analysis (XQDA) with the mahalanobis distance to improve classification results. Then we will apply a score normalization on the resulting mahalanobis distance on a four challenging person re-identification datasets VIPeR, GRID, CUHK01 and PRID450S. Experimental results on the four mentioned datasets demonstrate the effectiveness of the proposed feature.

Keywords: Person re-identification, CNN, XQDA, score normalization, mahalanobis.

Résumé

La ré-identification des personnes (Re-ID) est une tâche importante dans le domaine de la sécurité intelligente. Il s'agit du problème d'identification des personnes sur des images capturées par différentes caméras de surveillance sans chevauchement des champs de vision. Dans un environnement réel, Re-ID est une tâche difficile en raison des variations des conditions d'éclairage, des angles de vision, de la pose et des occlusions. Deux problèmes fondamentaux sont essentiels pour la ré-identification des personnes, la représentation des caractéristiques et l'apprentissage métrique. Une représentation efficace des caractéristiques doit être robuste à l'éclairage et aux changements de point de vue, et une métrique discriminante doit être apprise pour correspondre à diverses images de personnes. Dans ce travail, nous allons utiliser une représentation efficace des caractéristiques appelée réseau de neurones convolutionnels (CNN) pour détecter des modèles à des personnes distinctes, et une méthode d'apprentissage métrique appelée cross-view quadratic discriminant analysis (XQDA) avec la distance de mahalanobis pour améliorer les résultats de classification. Ensuite, nous appliquerons une normalisation du score sur la distance de mahalanobis résultante sur quatre ensembles de données de ré-identification de personnes difficiles VIPeR, GRID, CUHK01 et PRID450S. Les résultats expérimentaux sur les quatre ensembles de données mentionnés démontrent l'efficacité de la fonctionnalité proposée.

Les mots clés: ré-identification des personnes, CNN, XQDA, normalisation du score, mahalanobis.

Table of content

Dedication ..... **I**

Acknowledgments ..... **II**

ملخّص ..... **III**

Abstract ..... **IV**

Résumé ..... **V**

List of Figures ..... **VIII**

List of Tables ..... **IX**

List of Abbreviations ..... **X**

GeneralIntroduction **1**

Chapter 1 Biometrics and Person Re-identification **3**

1.1 Introduction 3

1.2 What is biometrics? 3

1.3 Biometric systems' modalities 3

1.4 Applications of biometric systems 8

1.5 Person re-identification 9

1.6 Re-identification pipeline 9

1.7 Challenges 11

1.8 Benchmark Datasets 14

1.9 Applications of person re-identification 15

1.10 Conclusion 17

Chapter 2 Artificial Intelligence and Deep Learning **18**

2.1 Introduction 18

2.2 Artificial intelligence 18

2.3 Machine learning 19

2.4 Classification of machine learning algorithms 20

2.4.1 Supervised learning 20

2.4.2 Unsupervised learning 22

2.4.3 Semi - supervised learning 22

2.4.4 Reinforcement learning 23

2.4.5 Transfer learning 24

2.5 Neural network 24

2.5.1 The biological neuron 24

2.5.2 The artificial neuron 25
2.6 Network architecture 27
2.7 Learning process in neural network 30
2.8 Convolutional neural networks 33
2.8.1 Convolution layer 34
2.8.2 Pooling layer 35
2.8.3 Fully-connected layer 36
2.9 ImageNet 37
2.10 Conclusion 38
Chapter 3 Implementations and Results 39
3.1 Introduction 39
3.2 General description of the proposed approach 39
3.3 Features projection using XQDA algorithm 41
3.4 Implementations and results 42
3.4.1 Databases and protocols 42
3.4.2 Discussions 47
3.4.3 Comparison with the stat of the art 49
3.5 Conclusion 49
General Conclusion 50
References 52

List of Figures

Figure 1.1 Biometric modalities; (a) face; (b) fingerprint; (c) hand-geometry; (d) gait; (e) iris; (f) 3D face; (g) voice; (h) ear structure..... 8
Figure 1.2 Pipeline of an end-to-end person re-id system 11
Figure 1.3 Example of some person re-identification challenges. Each pair of image shows the same person except (g). (a) Viewpoint variation, (b) pose variation, (c) illumination changes, (d) partial occlusion, (e) inaccurate pedestrian detection, (f) accessory change, (g) low resolution and (h) different people with similar clothing..... 14
Figure 2.1 A biological Neuron 25
Figure 2.2 The structure of an Artificial Neural Networks..... 26
Figure 2.3 A single layer perceptron..... 28
Figure 2.4 A multilayer perceptron with two hidden layers 29
Figure 2.5 A basic CNN Architecture 33
Figure 2.6 A schematic illustration of a convolutional operation. The convolutional kernel shifts over the source layer (image), filling the pixels in the destination layer (feature map).. 34
Figure 2.7 Max pooling..... 35
Figure 2.8 A fully connected layer..... 36
Figure 2.9 The architecture of the alexnet (CNN) 37
Figure 3.1 The extraction of a 4,096 dimensional feature vector for each image on the dataset. 40
Figure 3.2 Person re-identification system consisting of four stages: training data, feature extraction, metric learning and ranking. 41
Figure 3.3 sample image pairs from the PRID450S dataset. 43
Figure 3.4 Sample image pairs from the VIPeR dataset. 43
Figure 3.5 A four random images from the gallery and the probe of the GRID dataset. 44
Figure 3.6 Sample image pairs from the CUHK01 dataset..... 44
Figure 3.7 CMC curve of CUHK01 dataset..... 45
Figure 3.8 CMC curve of GRID dataset. 45
Figure 3.9 CMC curve of PRID450S dataset..... 46
Figure 3.10 CMC curve of VIPeR dataset. 46
Figure 3.11 the rank-1 performance on the four datasets CUHK01, PRID450S, GRID and VIPeR, with and without normalization..... 48

List of Tables

Table 1.1 Example of different datasets and its characteristics. 15
Table 3.1 CMC scores at r = 1, 5, 10, 15 and 20 of the used databases; VIPeR, GRID, CUHK01, and PRID450s with and without normalization. 47
Table 3.2 Comparison with the state-of-the-art of rank-1 identification rates (%) on the CUHK01, PRID450s, GRID and VIPeR databases. 49

List of Abbreviations

ANN_S: Artificial Neural Networks.

AI: Artificial Intelligence.

CCTV: Closed Circuit Television.

CMC: Cumulative Matching Characteristics Curves.

CMDA: Criterion Multilinear Discriminant Analysis.

CMLP: Criterion Multilinear Locality Preserved.

CMMF: Criterion Multilinear Marginal Fisher.

CNN_S: Convolutional Neural Networks.

ID: Identification.

ISO: International Standardization Organization.

IT: Information Technology.

LDA: Linear Discriminant Analysis.

MLP: Multilayer perceptron.

PCA: Principal component analysis.

Re-ID: Re-Identification.

ReLU: Rectified Linear Unit.

SLP: Signal Layer Perceptron.

SSL: Semi-Supervised Learning.

SVM: Support Vector Machine.

TSVM: Transductive Support Vector Machine.

XQDA: Cross-view Quadratic Discriminant Analysis.

General Introduction

Person re-identification is the task of identifying the same individual appeared in non-overlapping surveillance cameras with different viewpoints [1], when a pedestrian disappears from one camera's field of view and reappears in another camera's field of view, the Re-ID task must be able to identify the pedestrian in the new view based on the characteristics of the pedestrian in the previous view.

An appearance descriptor to reflect the person and a matching feature to compare certain appearance descriptors are typically used in person re-identification approaches. Contributions have been made over time to strengthen both the representation and the matching algorithm in order to increase robustness to the variations in pose, lighting, and background inherent to the problem. Recently, machine learning and deep learning, especially deep convolutional neural networks approaches have been applied to person re-identification system and achieved great results.

The re-identification process typically takes the two following steps:

- Extraction of person descriptors from multiple cameras and that include detection, tracking, feature extraction and descriptors generation.
- Establish correspondence (descriptor similarity ranking based ID assignments).

The main contributions of our work in order to enhance the efficiency of person re-identification system are:

1. We used CNN for an efficient feature representation.
2. To improve discriminative power for person re-identification we applied a metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA). Then we used the mahalanobis distance function as a distance metric.
3. we applied a score normalization function on the four datasets that we used VIPeR [2], GRID [3], CUHK01[4], and PRID450s [5], to optimize our results.

4. Finally, we evaluated and compared our results against the state of-the-art methods using the four challenging person re-identification databases that are mentioned previously.

The structure of our work is organized as follows: in chapter 1 we present biometrics and person re-identification and their applications. Deep learning and artificial intelligence are conducted in chapter 2. Finally, the implementation part which contains person re-identification based CNN and score normalization are presented in chapter 3.

Chapter 1 Biometrics and Person Re-identification

1.1 Introduction

Nowadays biometric systems become the most popular technique due to its liability. It is a branch of computer science that deals with the identification and verification of individuals. The goal of developing this kind of systems is to enhance the safety and information security. In this chapter, we examine the different modalities used in biometrics and its applications, with focusing on our application person re-identification, its pipeline, datasets, challenges and applications.

1.2 What is biometrics?

The international standardization organization (ISO) defines the term biometrics, as being "the automated recognition of individuals based on their biological and behavioral characteristics" [6]. The 'recognition' aims to associate an identity with an individual based on some physical characteristics exhibited intrinsically by his body parts and/or some behavioral characteristics created by the body. These characteristics are called "identifiers" or "traits". Examples of physical characteristics include among others: fingerprints, face, iris, etc. On the other hand, behavioral characteristics may include: signature, voice, keystroke dynamics [7].

1.3 Biometric systems' modalities

Biometric systems exploit a variety of biometric characteristics (or modalities) including fingerprint, face, ear, iris, retina, palmprint, vein, voice, signature, gait, and odor. The most leading biometric modalities are described below:

- **Fingerprint:** Fingerprint is one of oldest and the most popular biometrics that has proven its effectiveness and reliability in user identification. Fingerprint is composed of ridges and valleys found on the finger surface [8]. Fingerprint identification system consists of three phases:
    - Image acquisition and processing.
    - Feature extraction.

- Pattern recognition [9].

- **Face Recognition:** during the whole history of humanity, people used face to distinguish one person from the other. Facial (face) recognition is a computer application that automatically identifies or verifies a person with the help of a digital image or a video frame from a video source. One of the ways to do this is to compare the given example with the examples in the database [10]. In face biometrics, mainly visible imagery is used that is not affected by temperature and goggles [9].

  With the facial recognition modality, a digital image is used to analyze facial characteristics such as the distances between eyes, mouth, and nose. These measurements are stored in a database and used for comparison with a subject standing before a camera. The facial-scan modality is based on the standard biometric sequence of image acquisition, image processing, distinctive characteristic location, template creation, and matching. An optimal image is captured through a high-resolution camera, with moderate lighting and users directly facing a camera. The enrollment images define the facial characteristics to be used in all future verifications, thus a high-quality enrollment is essential. Challenges that occur in the image acquisition process include distance from user, angled acquisition, and lighting. Distance from the camera reduces facial size and thus image resolution. Users not looking directly at the camera, positioned vertically or horizontally more than 15 degrees away from ideal positioning are less likely to have images acquired. Lighting conditions, which cause an image to be underexposed or glared, may cause challenges. In addition, two-dimensional recognition is affected by changes in person's hair, age, and eye glasses. This modality requires appropriate camera equipment for user identification. On the other hand, this modality has the advantages of being non-intrusive and relatively less expensive [11].

- **Hand geometry:** geometry is the use of geometric shape of the hand for recognition purposes. This method was rather popular 10 years ago but nowadays it is seldom used. The method is based on the fact that the shape of the hand of one person differs from the shape of the hand of another person and does not change after certain age. But it is not unique [10].

A variety of measurements of the human hand, including its shape, and lengths and widths of the fingers, can be used as biometric characteristics. Hand geometry based biometric systems are installed at many locations around the world. The technique is very simple, relatively easy to use, and inexpensive. Operational environmental factors such as dry weather, or individual anomalies such as dry skin, generally have no negative effects on identification accuracy. A main disadvantage of this technique is its low discriminative capability. Hand geometry information may not be invariant over the lifespan of an individual, especially during childhood. In addition, an individual's jewelry or limitations in dexterity (e.g., from arthritis) may pose further challenges in extracting the correct hand geometry information. Lastly, because the physical size of a hand geometry-based system is large, it is difficult to use in certain applications [11].

- **Iris recognition:** The iris is a thin circular diaphragm, which lies between the cornea and the lens of the human eye. The iris is perforated close to its centre by a circular aperture known as the pupil. The function of the iris is to control the amount of light entering through the pupil. Iris is the elastic, pigmented, connective tissue that surrounds the pupil of the eye. Iris biometric is more reliable and accurate as compared to other biometric trait such as finger print. Iris texture is stable throughout life and is highly secure. Iris is less prone to attacks. Iris of the eye has different pattern for left and right eye. They are even unique for the identical twins.

  The structure of iris remains constant and stable throughout the life and serves as a good biometric trait for biometric recognition. The advantage of using iris for biometric recognition involves fewer margins of errors and fast speed and immune to the contact lenses and aging. These patterns are very rarely affected by the environment and hardly lost and believed that the trait revealed best, efficient and accurate results in the recognition process [9].

  Localization of the iris is an important step in iris recognition because, if done improperly, resultant noise (such as eyelashes, reflections, pupils, and eyelids) in the image may lead to poor performance. Iris imaging requires use of a high quality digital camera. Today's commercial iris cameras typically use infrared light to illuminate the iris without causing harm or discomfort to the subject [11].

Iris is used for various authentication and security applications that include identity cards and passports, prison security, database access and computer login, border control and Government programs [10].

- **Voice recognition:** Voice, like many other characteristics that are used for biometric methods, is unique; it takes quite little time to analyze the voice and to identify the person. Voice in biometrics or "voice print" is presented as a numerical model of the sound [10].

  Voice recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to verify the identity of a person accessing a system. That is, access control by voice in various services. These services include banking transactions over a telephone network, telephone shopping, database access services, information services, voice mail, security control for confidential information areas and remote access to computers. Most applications in which voice is used in user recognition are speaker verification systems.

  The advantages of this modality include its non-intrusive nature, high social acceptability, verification time is relatively low, and required technology is relatively inexpensive. On the other hand, it has some limitations. A person's voice can be easily recorded and used for unauthorized access. It has low accuracy, where an illness or the surrounding environment can change a person's voice, making absolute identification difficult or impossible [11].

- **Signature scanning:** Another behavioral biometric is signature by which the data can be extract by the signature of that particular person [10]. Each person has a unique style of handwriting, and no two signatures of different persons are identical. However, the variations of a typical signature also depend upon the physical and emotional state of a person. The identification accuracy of systems based on this highly behavioral biometric is reasonable but does not appear to be sufficiently high to lead to large-scale recognition. There are two approaches to identification based on signature; static and dynamic. Static signature identification uses only the geometric (shape) features of a signature, whereas dynamic (online) signature identification uses

both the geometric (shape) features and the dynamic features such as acceleration, velocity, pressure, and trajectory profiles of the signature.

An inherent advantage of a signature-based biometric system is that the signature has been established as an acceptable form of personal identification and can be incorporated transparently into the existing business processes requiring signatures such as credit card transactions. This modality requires a small verification time and relatively inexpensive equipment. Nonetheless, signature verification is designed to verify subjects based on the traits of their unique signature. As a result, individuals who do not sign their names in a consistent manner may have difficulty enrolling and verifying in signature verification [11].

- **Gait:** provides efficient means of verification. This modality has very sensitive limitation, such as collecting input data suited for the recognition process. This modality cannot be used in sensitive applications or daily life systems because of its high error rate [11].

  Ben et al [12] proposed the framework for the biometric system using gait modality which is a high dimensional modality and employed tensor analysis method. It was observed that when the view angles changes, it was complex to recognize cross-view gait even by using learning multilinear projection matrices. To handle such complexities a general tensor framework for cross-view gait recognition was proposed. To do so, a tensorial couple of mapping, with the coupling of multilinear locality-preserved criterion (CMLP), multi-linear marginal fisher criterion (CMMF) and multi-linear discriminant analysis criterion (CMDA) was used. For the experiment, CASIA and OU-ISIR gait datasets were selected and the results showed superior performance [9].
- **Ear:** This modality is more stable than facial expressions, making it appealing for biometric identification. Ear recognition may be done using a digital image, a thermographic image or an ear print. An ear print can be taken by pressing the ear against flat paper. This modality has some powerful advantages over many other modalities, such as its accuracy, allowing high enrolment and verification rates, not requiring an expert to interpret the comparison result, and usability with existing cameras and image capturing devices [11].

The most leading biometric modalities are illustrated in figure 1.1 below:

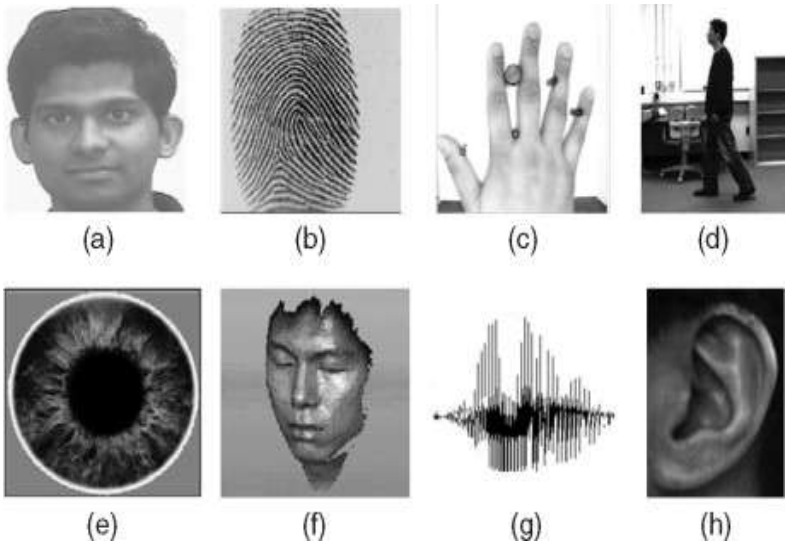


Figure 1.1 Biometric modalities; (a) face; (b) fingerprint; (c) hand-geometry; (d) gait; (e) iris; (f) 3D face; (g) voice; (h) ear structure.

1.4 Applications of biometric systems

There are numerous applications of biometric recognition systems used for the surveillance, law enforcement, time and attendance, logic access control and physical access control. Other applications include:

- Biometric systems are used to **secure the systems** for the surveillance and also for the physical buildings for various applications such as access control, computer networks, verification, investigations, authentication, parenthood determination, online banking, border control, e-commerce, medical records management, security monitoring and welfare disbursement.
- For **improving confidence level** after the verification of authenticated individual with all rights and privileges.
- To **identify an individual and control access** to information, physical spaces and services.
- **Improving safety,** reducing fraud and national security**.**
- **Provides uniqueness**, strong immunity and stability to forgery [9].

The applications of biometrics can be divided into the main groups such as: commercial applications, government applications, forensic applications, border

management, civil applications, customer verification at the point of sale, key applications include border management, interface for criminal and civil applications, and first responder verification. The enterprise solutions require the oversight of people, processes and technologies. Network infrastructure has become essential to functions of business, government, and web based business models. Consequently securing access to these systems and ensuring one's identity is essential. It uses to prevent stealing of possessions that mark the authorized person's identity e.g. security badges, licenses, or properties and to prevent fraudulent acts like faking ID badges or licenses, to ensure safety and security, thus decrease crime rates [13].

1.5 Person re-identification

Person re-identification (Re-ID) has been widely studied as a specific person retrieval problem across non-overlapping cameras [14, 15]. Given a query person-of interest, the goal of Re-ID is to determine whether this person has appeared in another place at a distinct time captured by a different camera [16]. The query person can be represented by an image [17, 18], a video sequence [19, 20], and even a text description [21, 22]. Due to the urgent demand of public safety and increasing number of surveillance cameras in university campuses, theme parks, streets, etc, person Re-ID is imperative in intelligent video surveillance system designs. Given its research impact and practical importance, Re-ID is a fast growing vision community [23].

The person re-identification is based on the pedestrian detection task. First, pedestrian gallery set is formed by collecting the cropped pedestrian images or extracted pedestrian image signature from each camera scene in the network. Next, the similarity or distance is measured to the query image. Finally, the best matched image according to the measured similarity is considered [24].

1.6 Re-identification pipeline

An automated re-identification mechanism takes as input either tracks or bounding boxes containing segmented images of individual persons, as generated by a localized tracking or detection process of a visual surveillance system. To automatically match people at different locations over time captured by different camera views [25], a re-identification process typically takes the following steps:

- **Extracting imagery features** that are more reliable, robust and concise than raw pixel data.
- **Constructing a descriptor** or representation, e.g. a histogram of features, capable of both describing and discriminating individuals.
- **Matching** specified probe images or tracks against a gallery of persons in another camera view by measuring the similarity between the images, or using some model-based matching procedure. A training stage to optimize the matching parameters may or may not be required depending on the matching strategy.

Such processing steps raise certain demands on algorithm and system design. This has led to both the development of new and the exploitation of existing computer vision techniques for addressing the problems of feature representation, model matching and inference in context in order to improve the efficiency of the Re-Identification systems [25].

Representation contemporary approaches to re-identification typically exploit low-level features such as color [26], texture, spatial structure [27], or combinations thereof [17, 25, 28, 29]. This is because these features can be relatively easily and reliably measured, and provide a reasonable level of inter-person discrimination together with inter-camera invariance. Such features are further encoded into fixed-length person descriptors, e.g. in the form of histograms, covariances or fisher vectors [25].

We take for example the matching; once a suitable representation has been obtained, nearest-neighbor [27]or model-based matching algorithms such as support-vector ranking [29] may be used for re-identification. In each case, a distance metric (e.g. Euclidean or Bhattacharyya) must be chosen to measure the similarity between two samples. Model based matching approaches [30, 31] and nearest-neighbor distance metrics [32, 33] can both be discriminatively optimized to maximize re-identification performance given annotated training data of person images. Bridging these two stages, some studies [3, 34, 35] have also attempted to learn discriminative low-level features directly from data [25].

Complementary aspects of the re-identification problem have also been pursued to improve performance, such as improving robustness by combining multiple frames worth of features along a trajectory tracklet [28, 36], set based analysis [37, 38], considering

external context such as groups of persons [39], and learning the topology of camera networks [40, 41] in order to reduce the matching search space and hence reduce false-positives [25]. The general pipeline of an end-to-end person re-id system is shown in figure 1.2.

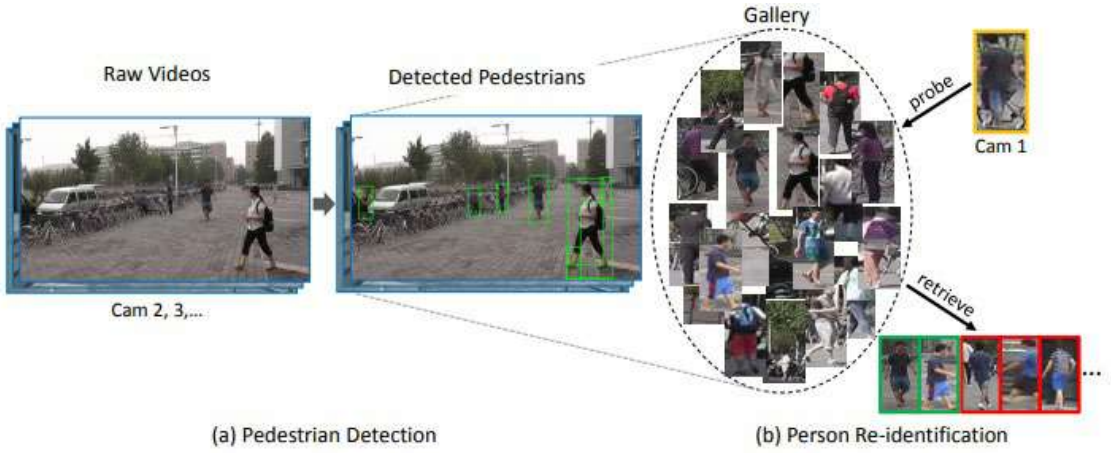


Figure 1.2 Pipeline of an end-to-end person re-id system

1.7 Challenges

Person Re-ID is a challenging task due to many factors such as presence of different viewpoints, varying low-image resolutions, illumination changes, unconstrained poses, occlusions, heterogeneous modalities, etc. Some challenging examples are shown in figure 1.3.The challenging factors and their effects are explained below.

- **Illumination variation:** Illumination conditions can vary in different camera scenes or during the day. The same person observed under different lightening conditions can have a color different on the appearance. This increases the intra-class variation [24].
- **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 [24], which may cause failure to re-ID systems trained on the source domain only when directly testing on a target domain, as the model cannot learn target camera-invariant features [42].

- **Pose variation:** The large pose deformations on human body are usually caused by the relative motions among rigid body parts. During the motion, some body parts like arms, legs and trunk can be approximately regarded as rigid regions, and their motions can be modeled by affine transformations [43]. In a typical surveillance scenario, the pose of a person with respect to a camera vary due to the camera viewpoint and the person's walking direction. As a result, the information extracted from the images both static and dynamic body cues may change drastically both across cameras and a long time. Hence, the pose/gait direction of the subject is a very significant criteria to be considered in the gait-based Re-ID scheme [44].
- **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 [24].
- **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 [24].
- **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 [24].

- **Partial occlusion:** The types of facial occlusion that occur may vary significantly. Generally, there are two major types of partial occlusion: systematic and temporary. Systematic occlusions are caused by the existence of individual facial components (e.g., hair, mustache, or a scar), or by people wearing adornments (e.g. glasses, clothes, a hat or surgical mask, or mark-ups). Temporary occlusions arise from a portion of the face being temporarily obscured by other objects (e.g., people moving across the face or hands covering the face), or self-occlusion due to changes in head pose or temporarily placing objects in front of the face [45].
- **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 [24].
- **Small sample size:** In general a re-identification module may be required to match single probe images to single gallery images. This means from a conventional classification perspective, there is likely to be insufficient data to learn a good model of each person's intra-class variability. 'One-shot' learning may be required under which only a single pair of examples is available for model learning. For this reason, many frameworks treat re-id as a pair wise binary classification (same versus different) problem [29, 31] instead of a conventional multi-class classification problem [14].
- **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 [24].
- **Data labeling:** If exploiting a supervised learning strategy to train a good model robust to cross-camera view variations, persons from each view annotated with identity or binary labels depicting same versus different are required. Consequently, models which can be learned with less training data are preferred

since for a large camera network, collecting extensive labeled data from every camera would be prohibitively expensive [14].

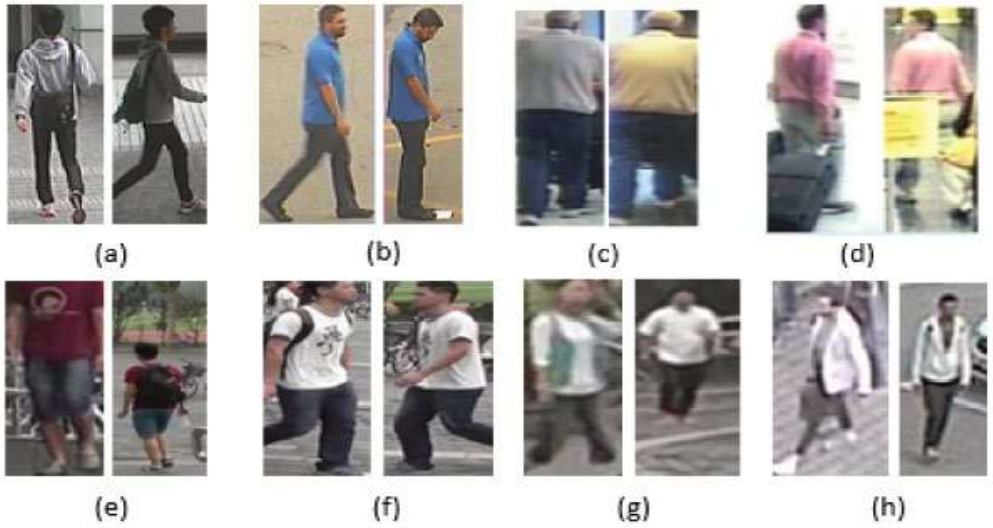


Figure 1.3 Example of some person re-identification challenges. Each pair of image shows the same person except (g). (a) Viewpoint variation, (b) pose variation, (c) illumination changes, (d) partial occlusion, (e) inaccurate pedestrian detection, (f) accessory change, (g) low resolution and (h) different people with similar clothing.

1.8 Benchmark datasets

Multiple benchmark datasets for person re-identification have been published in recent years. There are multiple factors to be considered when creating a benchmark dataset [46], such as:

- Number of pedestrians.
- Number of images per person in one camera view.
- Variations of resolutions, lightings, poses, occlusion and background in the same camera view and across camera views.
- Number of camera views.

Some public datasets that have been tested for person re-identification algorithms and its characteristics are illustrated in Table 1.1.

Table 1.1 Example of different datasets and its characteristics.

Datasets	Release time	# Identities	# Image	# Cameras	Crop Size
VIPeR	2007	632	1264	2	128*48
GRID	2009	1025	1275	8	Vary
PRID2011	2011	934	24541	2	128*64
CUHK01	2012	971	3884	2	160*60
CUHK02	2013	1816	7264	10	160*60
CUHK03	2014	1467	13164	10	Vary
iLIDS-VID	2014	300	42495	2	Vary
Market1501	2015	1501	32217	6	128*64
MARS	2016	12611	1191003	6	256*128
AIRPORT	2017	9651	39902	6	128*64
LPW	2018	2,731	592,438	3,4,4	-
ThermalWo-rld	2018	516	15,118	20	-

Note. In our work we use VIPeR, GRID, PRID450s and CUHK01 datasets for the evaluation of our system.

1.9 Applications of person re-identification

Person re-identification strategies have a lot of potential in a lot of different fields, from defense and surveillance to retail and health care.

- **Cross-camera person tracking:** Computer vision requires the ability to monitor people through several cameras, conduct crowd movement analysis, and recognize movements in order to comprehend a scene [24]. Typically, cross-camera tracking consists of two steps, object tracking in single camera and data association between multiple cameras. Consequently, besides the difficulty in single camera tracking, the main challenge of cross-camera tracking is to associate the trajectories of human movement in different views [47].

- **Tracking by detection:** Individual re-identification may be useful even in a single-camera tracking setup. Multiple person tracking is a difficult job, particularly in complex and crowded scenarios with frequent occlusions and individual interaction. 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 [24]. This stems directly from the development of powerful discriminative methods in machine learning, and their application to detection with offline training. Many of these algorithms can be adapted for online training, where each successful detection provides more information about the target [48].
- **Person retrieval:** Re-identification is linked to a recognition task in this case.
  A specific query with a target individual is given, and all of the corresponding instances in a broad database are searched. For image retrieval, the re-identification task is used, and it typically returns ranked lists, similarly related objects, and so on [24].
- **Human-machine interaction:** in the recent decades the scientist have concentrated more in the developing technologies that help in the improving the interactions between the humans and the machines by integrating novel resourceful procedures with the computers. This caused a significant breakthrough in the computer-vision and enabled the computers to operate or control the machine by monitoring the gestures and the expression of the humans [49].
- **Long-term human behavior and activity analysis:** Analyzing consumer shopping habits, for example, by watching them touch, interview, and try items in stores under various surveillance cameras. Another example is geriatric health care research, which looks at the long-term actions of elderly people to help physicians make more precise diagnosis [24].

1.10 Conclusion

Person re-identification is a recent area of research that combines biometrics and computer vision, but it performs admirably. In this chapter, we have discussed the field of biometrics along with its different modalities and their applications. We also discussed person re-identification, its pipeline, challenges and the benchmark databases. Several ongoing studies are constantly developing new and efficient methods to address individual re-ID challenges and increase recognition accuracy by incorporating powerful methods for feature extraction, dimensionality reduction, fusion, and classification

Chapter 2 Artificial Intelligence and Deep Learning

2.1 Introduction

The study of intelligent machines and neuroscience are both extremely active fields so as artificial intelligence. Artificial intelligence regulates several sub-disciplines such as machine learning and neural networks. Machine learning has progressed dramatically over the past two decades. It is important both for addressing fundamental scientific and engineering questions; it has produced and fielded across many applications. The aim of this chapter is to provide an overview of the basic concepts in machine learning literature, including their types and implementations, as well as their various algorithms. We also go over the convolutional neural networks.

2.2 Artificial intelligence

Artificial intelligence is wide-ranging branch of computer science concerned with building smart machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving, reasoning, discovering meaning, generalizing, or learn from past experiences. For example, discovering proofs for mathematical theorems or playing chess with great proficiency [50].

Still, despite continuing advances in computer processing speed and memory capacity, there are as yet no programs that can match human flexibility over wider domains or in tasks requiring much everyday knowledge. On the other hand, some programs have attained the performance levels of human experts and professionals in performing certain specific tasks**,** so that artificial intelligence in this limited sense is found in applications as diverse as medical diagnosis, computer search engines, and voice or handwriting recognition [51].

AI is an interdisciplinary science with multiple approaches, but advancements in machine learning and deep learning are creating a paradigm shift in virtually every sector of the tech industry.AI now consists many sub-fields, using a variety of techniques, such as [52]:

- Neural networks – e.g. brain modeling, time series prediction, classification.
- Evolutionary computation – e.g. genetic algorithms, genetic programming.
- Vision – e.g. object recognition, image understanding.
- Robotics – e.g. intelligent control, autonomous exploration.
- Expert systems – e.g. decision support systems, teaching systems.
- Speech processing– e.g. speech recognition and production.
- Natural language processing – e.g. machine translation.
- Planning – e.g. scheduling, game playing.
- Machine learning – e.g. decision tree learning, version space learning.

2.3 Machine learning

Machine learning is a sub-area of artificial intelligence, whereby the term refers to the ability of IT (Information Technology) systems to independently find solutions to problems by recognizing patterns [53].Or we can define machine learning as a subset of AI, which build a mathematical model based on sample data, known as "training data," in order to make predictions or decisions without being explicitly programmed to perform the task [54].

In order to enable the software to independently generate solutions, the prior action of people is necessary. For example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock must be defined. Once these two steps have been completed, the system can perform the following tasks by machine learning [54]:

- Finding, extracting and summarizing relevant data.
- Making predictions based on the analysis data.
- Calculating probabilities for specific results.
- Adapting to certain developments autonomously.
- Optimizing processes based on recognized patterns.

In machine learning, we are usually given a training set and a test set. By the training set, it will mean the union of the labeled set and the unlabeled set of examples available to machine learners. In comparison, test set consists of examples never seen before. Machine learning aims to establish a regressor or classifier through learning the

training set, and then to evaluate the performance of the regressor or classifier through the test set [54].

Machine learning works in a similar way to human learning. For example, if a child is shown images with specific objects on them, they can learn to identify and differentiate between them. Machine learning works in the same way: Through data input and certain commands, the computer is enabled to "learn" to identify certain objects (persons, objects, etc.) and to distinguish between them. For this purpose, the software is supplied with data and trained. For instance, the programmer can tell the system that a particular object is a human being ("human") and another object is not a human being ("no human"). The software receives continuous feedback from the programmer. These feedback signals are used by the algorithm to adapt and optimize the model. With each new data set fed into the system, the model is further optimized so that it can clearly distinguish between "humans" and "non-humans" in the end.

2.4 Classification of machine learning algorithms

According to the nature of training data, we can classify machine learning algorithms to five grand methods; supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and transfer learning.

2.4.1 Supervised learning

In supervised learning, the target is to infer a function or mapping from training data that is labeled. The training data consist of input vector X and output vector Y of labels or tags. A label or tag from vector Y is the explanation of its respective input example from input vector X. Together they form a training example. In other words, training data comprises training examples. If the labeling does not exist for input vector X, then X is unlabeled data.

Why such learning is called supervised learning? The output vector Y consists of labels for each training example present in the training data. These labels for output vector are provided by the supervisor. Often, these supervisors are humans, but machines can also be used for such labeling. Human judgments are more expensive than machines, but the higher error rates in data labeled by machines suggest superiority of human judgment. The manually labeled data is a precious and reliable resource for supervised

learning. However, in some cases, machines can be used for reliable labeling [55]. Given an example of four categories of supervised learning algorithm:

- **Decision tree**

A decision tree is a logically learning supervised algorithm that is mostly used for classification and provides a series of decision sequences that, if followed, will predict the label of unlabeled data [56].

- **Naïve bayes**

A classification method that estimates conditional class probabilities by applying Bayes' theorem under the (naive) assumption that the attribute values are independent given the class [57].

- **Support vector machine**

SVM is an advanced supervised method that can handle both regression and classification tasks, while classification is preferable. It has the ability to handle both continuous and categorical instances [56].

- **Linear discriminant analysis**

Linear Discriminant Analysis (LDA) is the most commonly used dimensionality reduction technique in supervised learning. Basically, it is a preprocessing step for pattern classification and machine learning applications. It projects the dataset into moderate dimensional-space with a genuine class of separable features that minimize overfitting and computational costs[58].

While focusing on projecting the features in higher dimension space onto a lower dimensional space, LDA achieve this via three step process [58];

- To compute the separate ability among various classes, i.e. the distance between the mean of different classes, that is also known as between-class variance.
- To compute the distance among the mean and sample of each class, that is also known as the within class variance.
- To create the lower dimensional space that maximizes the between class variance and minimizes the within class variance.

LDA has been successfully used in various applications, as far as a problem is transformed into a classification problem, this technique can be implemented. For

example, LDA can be used as a classification task for speech recognition, face recognition, image retrieval, biometrics, chemistry, etc [58].

2.4.2 Unsupervised learning

In unsupervised learning, we lack supervisors or training data. In other words, all what we have is unlabeled data. The idea is to find a hidden structure in this data. There can be a number of reasons for the data not having a label. It can be due to unavailability of funds to pay for manual labeling or the inherent nature of the data itself. With numerous data collection devices, now data is collected at an unprecedented rate. The variety, velocity, and the volume are the dimensions in which Big Data is seen and judged. To get something from this data without the supervisor is important. This is the challenge for today's machine learning practitioner [55]. Given an example of two types of unsupervised learning algorithms:

- **K-Means clustering**

Clustering or grouping is a type of unsupervised learning technique that when initiates, creates groups automatically. The items which possess similar characteristics are put in the same cluster [59].

- **Principal component analysis**

Principal component analysis (PCA) is a technique for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not a priori, hence making PCA an adaptive data analysis technique. It is adaptive in another sense too, since variants of the technique have been developed that are tailored to various different data types and structures [60]. In PCA, the dimension of the data is reduced to make the computations faster and easier.

2.4.3 Semi - supervised learning

In this type of learning, the given data are a mixture of classified and unclassified data. This combination of labeled and unlabeled data is used to generate an appropriate model for the classification of data. In most of the situations, labeled data is scarce and unlabeled data is in abundance (as discussed previously in unsupervised

learning description). The target of semi-supervised classification is to learn a model that will predict classes of future test data better than that from the model generated by using the labeled data alone. The way we learn is similar to the process of semi-supervised learning [55]. There are many categories of semi-supervised learning, four of which are mentioned below:

- **Generative models**

Generative models are one of the oldest semi-supervised learning method assumes a structure like$p(x,y) = p(y)p(x|y)$ where $p(x|y)$ is a mixed distribution e.g. Gaussian mixture models [59].

- **Low-density separation**

One of the important intuitions of semi-supervised learning is the low density separation, which states that the class boundary passes through a low density region [61].

- **Graph-based methods**

Graph-based SSL is among the most popular SSL methods. As the name suggests, graph plays an important role in this family of algorithms. The data in these problems appear in the form of a weighted graph [62].

- **Self-training**

In self-training, a classifier is trained with a portion of labeled data. The classifier is then fed with unlabeled data. The unlabeled points and the predicted labels are added together in the training set [59].

2.4.4 Reinforcement learning

The reinforcement learning method aims at using observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk. In order to produce intelligent programs (also called agents), reinforcement learning goes through the following steps:

- Input state is observed by the agent.
- Decision making function is used to make the agent perform an action.
- After the action is performed, the agent receives reward or reinforcement from the environment.
- The state-action pair information about the reward is stored.

Using the stored information, policy for particular state in terms of action can be fine-tuned, thus helping in optimal decision making for our agent [55].

2.4.5 Transfer learning

Transfer learning is an important tool in machine learning to solve the basic problem of insufficient training data. It tries to transfer the knowledge from the source domain to the target domain by relaxing the assumption that the training data and the test data must be. This will leads to a great positive effect on many domains that are difficult to improve because of insufficient training data [63].

2.5 Neural network

Neural networks are a computational model that shares some properties with the humain brain in which many simple units are working in parallel with no centralized control unit. The weights between the units are the primary means of long-term information storage in neural networks. Updating the weights is the primary way the neural network learns new information. The behavior of neural networks is shaped by its network architecture. A network's architecture can be defined by the Number of neurons, Number of layers and Types of connections between layers.

2.5.1 The biological neuron

The human brain is a collection of more than 10 billion interconnected neurons. Each neuron is a cell that uses biochemical reactions to receive, process, and transmit information. Tree like networks of nerve fibers called dendrites are connected to the cell body or soma, where the cell nucleus is located. Extending from the cell body a single long fiber called the axon, which eventually branches into strands and sub-strands, and is connected to other neurons through synaptic terminals or synapses. The transmission of signals from one neuron to another at synapses is a complex chemical process in which specific transmitter substances are released from the sending end of the junction. The effect is to raise or lower the electrical potential inside the body of the receiving cell. If the potential reaches a threshold, a pulse is sent down the axon and the cell is 'fired' [64]. Figure 2.1 illustrates a biological neuron.

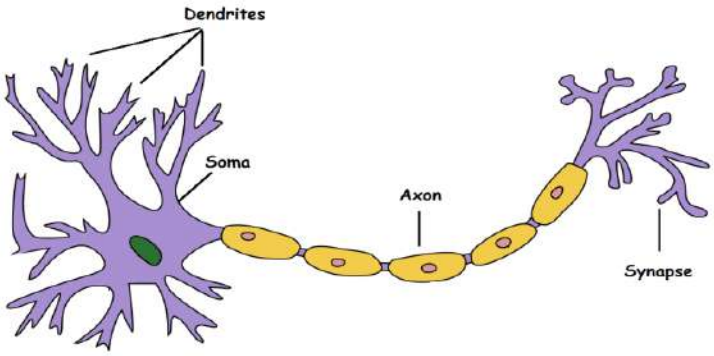


Figure 2.1 A biological Neuron

- **Synapses :**

Synapses are the connecting junction between axon and dendrites. The majority of synapses send signals from the axon of a neuron to the dendrite of another neuron. The exceptions for this case are when a neuron might lack dendrites, or a neuron lacks an axon, or a synapse, which connects an axon to another axon [65].

- **Dendrites :**

Dendrites have fibers branching out from the soma in a bushy network around the nerve cell. Dendrites allow the cell to receive signals from connected neighboring neurons and each dendrite is able to perform multiplication by that dendrite's weight value. Here multiplication means an increase or decrease in the ratio of synaptic neurotransmitters to signal chemicals introduced into the dendrite [65].

- **Axons :**

Axons are the signals, long fibers extending from the main soma. They stretch out longer distances than dendrites and measure generally 1 centimeter in length (100 times the diameter of the soma). Eventually, the axon will branch and connect to other dendrites. Neurons are able to send electrochemical pulses through cross-membrane voltage changes generating action potential. This signal travels along the cell's axon and activates synaptic connection with other neurons [65].

2.5.2 The artificial neuron

Inspired by biological neural networks, ANNs are massively parallel computing systems consisting of an extremely large number of simple processors with many interconnections. ANN models attempt to use some "organizational" principles believed to be used in the human. One type of network sees the nodes as 'artificial neurons'. These

are called artificial neural networks (ANNs). An artificial neuron is a computational model inspired in the natural neurons as shown in figure 2.2. Since the function of ANNs is to process information, they are used mainly in fields related with it. There are a wide variety of ANNs that are used to model real neural networks, and study behavior and control in animals and machines, but also there are ANNs that are used for engineering purposes, such as pattern recognition, forecasting, and data compression. These basically consist of inputs (like synapses), which are multiplied by weights. Weights assigned with each arrow represent information flow. These weights are then computed by a mathematical function which determines the activation of the neuron. Another function (which may be the identity) computes the output of the artificial neuron (sometimes in dependence of a certain threshold). The neurons of this network just sum their inputs. Since the input neurons have only one input, their output will be the input they received multiplied by a weight [66].

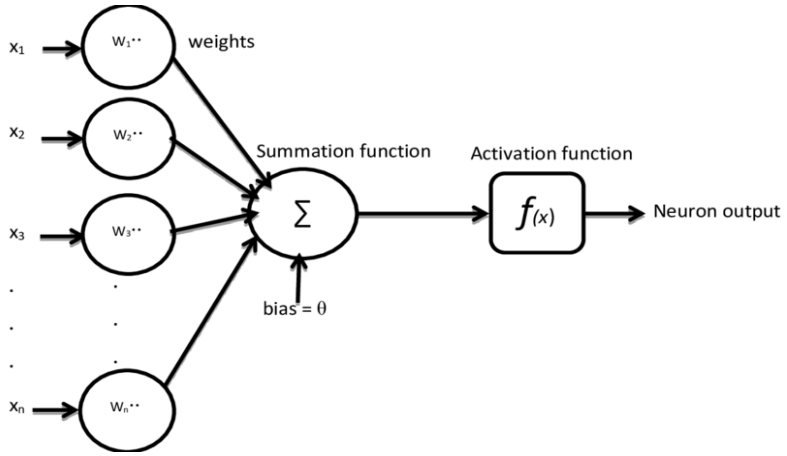


Figure 2.2 The structure of an Artificial Neural Networks

If the weight is high, then input will be strong. By adjusting the weights of an artificial neuron we can obtain the output we want for specific inputs. But when we have an ANN of hundreds or thousands of neurons, it would be quite complicated to find by hand all the necessary weights. But we can find algorithms which can adjust the weights of the ANN in order to obtain the desired output from the network. This process of adjusting the weights is called learning or training. The training begins with random weights, and the goal is to adjust them so that the error will be minimal.

The signal flow from inputs $x_1 ... x_n$ is considered to be unidirectional, which are indicated by arrows, as is a neuron's output signal flow (O). The neuron output signal O is given by the following relationship:

$$O = f(net) = f\left(\sum_{j=1}^{n} w_j x_j\right). \quad \textbf{(2.1)}$$

Where w_j is the weight vector, and the function f(net) is referred to as an activation (transfer) function. The variable net is defined as a scalar product of the weight and input vectors,

$$net = w^T x = w_1 x_1 + \ldots + w_n x_n. \quad \textbf{(2.2)}$$

Where T is the transpose of a matrix, and, in the simplest case, the output value O is computed as:

$$O = f(net) = \begin{cases} 1 \ if \ w^T \geq \theta \\ 0 \ otherwise \end{cases}. \quad \textbf{(2.3)}$$

Where θ is called the threshold level, and this type of node is called a linear threshold unit.

2.6 Network architecture

Based on the connection pattern network architecture can be grouped into two categories:

- Feed-forward networks, in which graphs have no loops.
- Recurrent (or feedback) networks, in which loops occur because of feedback connections.

Generally, Feed-forward networks are static, that is, they produce only one set of output values rather than a sequence of values from a given input. Feed-forward networks are memory-less in the sense that their response to an input is independent of the previous network state [66].

a. Single layer perceptron

The simplest of the neural network models, Single Layer Perceptron was designed by researchers McCulloch and Pitts. In the eyes of many machine learning scientists, SLP is viewed as the beginning of artificial intelligence and provided inspiration in developing other neural network models and machine learning models. The SLP architecture is such

that a single neuron is connected by many synapses, each of which contains a weight [67] as illustrated in figure 2.3.

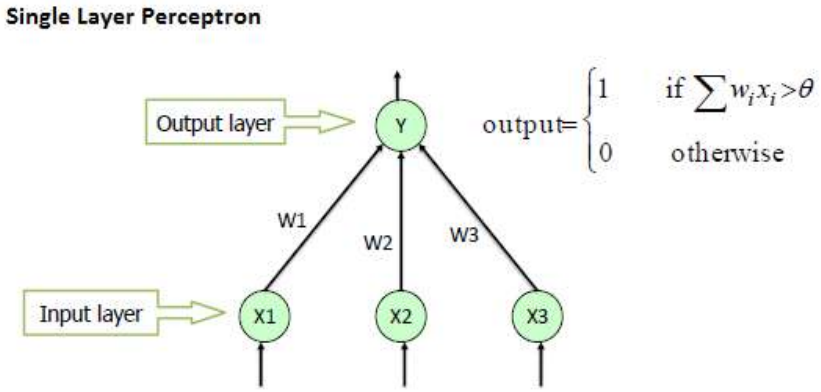


Figure 2.3 A single layer perceptron

b. Multilayer perceptron

Inspired by the Single Layer Perceptron, the multilayer perceptron consists of an input layer, hidden layers and an output layer. The multilayer perceptron model features multiple layers that are interconnected in such a way that they form a feed-forward neural network. Each neuron in one layer has directed connections to the neurons of a separate layer. One of the key distinguishing factors in this model and the single layer perceptron model is the back-propagation algorithm, a common method of training neural networks. Back-propagation passes the error calculated from the output layer to the input layer such that we can see each layer's contribution to the error and alter the network accordingly. Here, we use a gradient descent algorithm to determine the degree to which the weights should change upon each iteration. Gradient descent another popular machine learning/optimization algorithm is simply the derivative of a function such that we find a scalar (a number with magnitude as its only property) value those points in the direction of greatest momentum. By subtracting the gradient, this leads us to a solution that is more optimal than the one we currently are at until we reach a global optimum [67]. Figure 2.4 shows the architecture of a simple feed forward neural network, it consists of an input layer, two hidden layers and an output layer. Each circle represents an artificial neuron or a node; the neurons are connected with each other by connections that represent the weights.

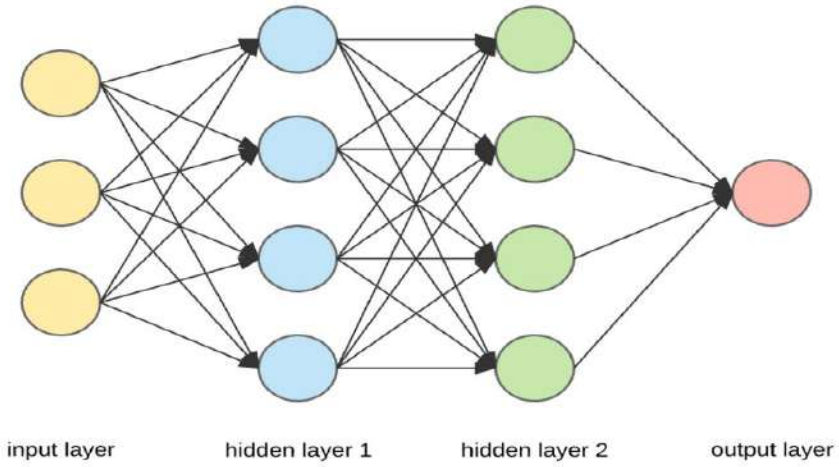


Figure 2.4 A multilayer perceptron with two hidden layers

c. Radial basis function network

A radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis of the inputs and neuron parameters. Radial basis function networks are used for function approximation, time series prediction, and system control [66].

➢ **Activation functions**

Activation functions reside within neural network layers and change the data they receive before transferring it to the next layer. Activation functions give neural networks their power allowing them to model complex non-linear relationships. Neural networks can model highly complex relationships between features by changing inputs with non-linear functions. ReLu and Sigmoid are two common activation functions. The properties of activation functions are generally as follows:

- **Non-linear:** We can only use a prediction equation that looks like a straight line in linear regression. This works well for simple datasets with a one-to-one relationship between inputs and outputs, but what if our dataset contains non-linear patterns? (For example, x2, sin, and log). A non-linear prediction equation is needed to model these relationships. This non-linearity is provided by activation functions.
- **Continuously differentiable:** To use gradient descent to improve our model, we need a nice slope in our output so that we can compute error derivatives with

respect to weights. We wouldn't know which way to update our weights to reduce our error if our neuron instead outputted 0 or 1 (perceptron).

- **Fixed Range:** The input data is usually squashed into a narrow range by activation functions, which makes training the model more stable and effective [68].

2.7 Learning process in neural network

In the sense of ANNs (Artificial Neural Networks), a learning process can be thought of as the problem of upgrading network architecture and connection weights so that a network can perform a particular task efficiently. The connection weights are typically learned by the network from available training patterns. Iteratively updating the weights in the network improves performance over time. The willingness of Artificial Neural Networks to learn from examples makes them appealing and exciting. ANNs tend to learn underlying rules (like input-output relationships) from a set of representative examples rather than following a set of rules defined by human experts. One of the most important benefits of neural networks over conventional expert systems is this [66].

- **Learning paradigms:** Three main learning paradigms exist, each of which corresponds to a specific abstract learning activity. Supervised learning, unsupervised learning, and reinforcement learning are the three types of learning.
  - Supervised learning the network is provided with a correct answer (output) for every input pattern: Weights are determined to allow the network to produce answers as close as possible to the known correct answers.
  - Unsupervised learning does not require a correct answer associated with each input pattern in the training data set: It explores the underlying structure in the data, or correlations between patterns in the data, and organizes patterns into categories from these correlations. Hybrid learning combines supervised and unsupervised learning. Parts of the weights are usually determined through supervised learning, while the others are obtained through unsupervised learning.

- Reinforcement learning: is a variant of supervised learning in which the network is provided with only a critique on the correctness of network outputs, not the correct answers themselves [66].

➢ **Error-Correction Rules:** The network is given a desired output for each input pattern in the supervised learning paradigm. The actual output produced by the network during the learning process can differ from the desired output. The basic principle of error-correction learning rules is to adjust the connection weights based on the error signal (target output-obtained output) to gradually reduce the error [66].

➢ **Perceptron learning rule:** The perceptron is a supervised classification algorithm that converts an input into one of several non-binary outputs. The perceptron learning algorithm is an online algorithm, meaning it processes elements in the training set one by one. A perceptron is analogous to a linear neuron in artificial neural networks, except that it performs classification rather than regression. A linear neuron adjusts its weights to minimize a real valued prediction error, while a perceptron adjusts its weights to divide the training examples into their respective classes. To differentiate it from a multilayer perceptron, which is a misnomer for a more complex neural network, the perceptron algorithm is also known as the single-layer perceptron. The single-layer perceptron is the simplest feedforward neural network as a linear classifier [66].

➢ **Boltzmann learning:** Boltzmann learning is derived from information-theoretic and thermodynamic principles. The objective of Boltzmann learning is to adjust the connection weights so that the states of visible units satisfy a particular desired probability distribution. Boltzmann learning can be viewed as a special case of error-correction learning in which error is measured not as the direct difference between desired and actual outputs, but as the difference between the correlations among the outputs of two neurons under clamped and free running operating conditions [66].

➢ **Hebbian rule:** "Cells that fire together, wire together," is a popular summary of the theory. Hebb's theory is a technique for figuring out how to change the

weights between model neurons. When two neurons fire at the same time, their weight increases; when they activate separately, it decreases. Strong positive weights are assigned to nodes that are both positive and negative at the same time, while strong negative weights are assigned to nodes that are opposite [66].

$$Wi(new) = Wi(old) + XiY. \quad \textbf{(2.4)}$$

- **Competitive wing rules:** Unlike Hebbian learning, competitive-learning output units compete among themselves for activation. As a result, only one output unit is active at any given time. This phenomenon is known as winner-take-all. Competitive learning has been found to exist in biological neural network. Competitive learning often clusters or categorizes the input data. Similar patterns are grouped by the network and represented by a single unit. This grouping is done automatically based on data correlations. The simplest competitive learning network consists of a single layer of output units (Only the weights of the winner unit get updated).

  This learning rule has the effect of bringing the stored pattern in the winner unit (weights) closer to the input pattern. We can see from the competitive learning rule that unless the learning rate is zero, the network will continue to learn (update weights). During learning, a particular input pattern may fire different output units at different iterations. This brings up the question of a learning system's stability. After a finite number of learning iterations, the method is said to be stable if no pattern in the training data changes its category. One way to achieve stability is to gradually reduce the learning rate as the learning process progresses.

  However, this artificial freezing of learning causes another problem termed plasticity, which is the ability to adapt to new data. This is known as Grossberg's stability plasticity dilemma in competitive learning [66].

- **The back propagation algorithm:** In layered feed-forward ANNs, the back propagation algorithm is used. This means that the artificial neurons are arranged in layers and transmit signals "forward," with errors propagated backwards. Neurons in the input layer provide input to the network, and neurons in the output layer provide the network's output. There may be one or more intermediate layers.

Backpropagation is a supervised learning algorithm. We can adjust the weights using the gradient descendent method after we've found the error.

Since the time needed to train the networks grows exponentially, ANNs implementing the backpropagation algorithm do not have too many layers. The backpropagation algorithm has also been improved, allowing for faster learning [66].

2.8 Convolutional neural networks

Convolutional neural network is a type of deep learning model for processing data that has a grid pattern, such as images, which is inspired by the organization of animal visual cortex and designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. CNN is a mathematical construct that is typically composed of three types of layers (or building blocks): convolution, pooling, and fully connected layers [69].A typical architecture (see figure 2.5) consists of repetitions of a stack of several convolution layers and a pooling layer, followed by one or more fully connected layers. The step where input data are transformed into output through these layers is called forward propagation.

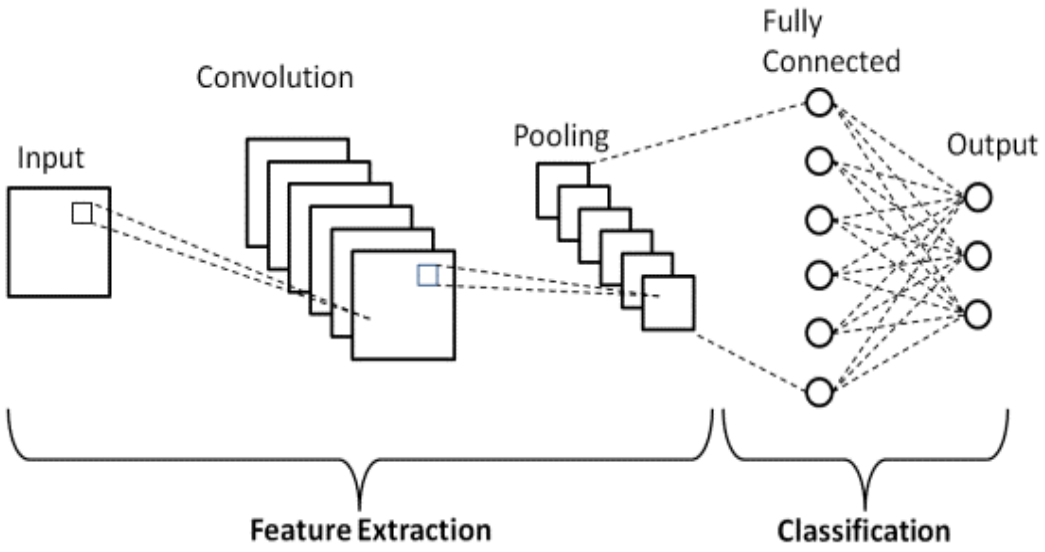


Figure 2.5 A basic CNN Architecture

2.8.1 Convolution layer

A convolution layer is a fundamental component of the CNN architecture that performs feature extraction, which typically consists of a combination of linear and nonlinear operations, i. e., convolution operation and activation function [69].

- **Convolution:** Convolution is a specialized type of linear operation used for feature extraction, where a small array of numbers, called a kernel, is applied across the input, which is an array of numbers, called a tensor. An element-wise product between each element of the kernel and the input tensor is calculated at each location of the tensor and summed to obtain the output value in the corresponding position of the output tensor, called a feature map. This procedure is repeated applying multiple kernels to form an arbitrary number of feature maps, which represent different characteristics of the input tensors; different kernels can, thus, be considered as different feature extractors. Two key hyper parameters that define the convolution operation are size and number of kernels. The former is typically 3 × 3, but sometimes 5 × 5 or 7 × 7. The latter is arbitrary, and determines the depth of output feature maps[69]. Figure 2.6 shows the convolution operation. For a two-dimensional image X as input, the convolution operator is defined as [24]:

$$(W * X)(i,j) = \sum_m \sum_n X(m,n)W(i-m, j-n). \qquad \textbf{(2.5)}$$

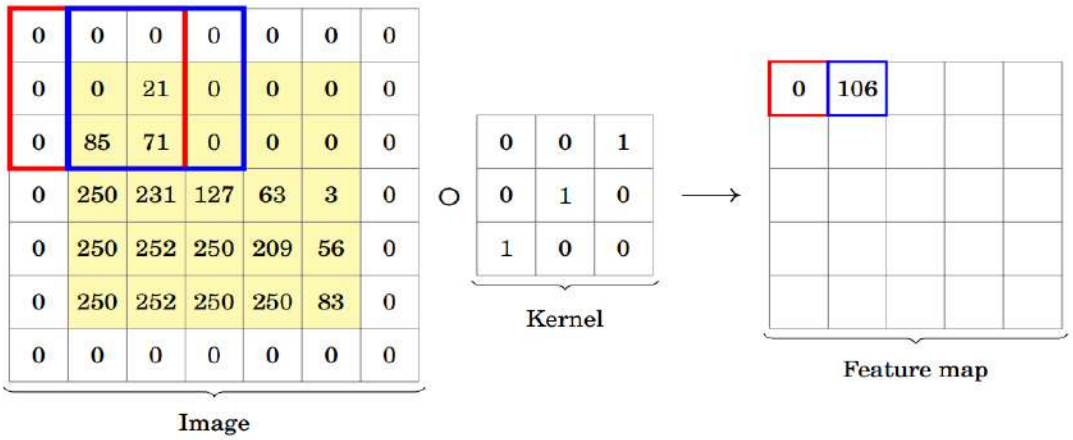


Figure 2.6 A schematic illustration of a convolutional operation. The convolutional kernel shifts over the source layer (image), filling the pixels in the destination layer (feature map).

- **Nonlinear activation function:** The outputs of a linear operation such as convolution are then passed through a nonlinear activation function. Although smooth nonlinear functions, such as sigmoid or hyperbolic tangent [69].

2.8.2 Pooling layer

The aim of pooling layers as shown in figure 2.7 is to progressively reduce the dimensionality of the representation, lowering the number of parameters and the model's computational complexity.

The pooling layer scales the dimensionality of each activation map in the input using the "MAX" function. In most CNNs, these take the form of **max-pooling layers** with kernels with a dimensionality of 2x2 and a stride of 2 along the input's spatial dimensions. This reduces the size of the activation map to 25% of its original size while keeping the depth volume at its original size. Figure 2.7 shows an example of 2x2 max pooling.

There are only two commonly observed methods of max-pooling due to the harmful aspect of the pooling layer. The stride and filters of the pooling layers are usually both set to 2x2, allowing the layer to stretch within the input's spatial dimensionality. Overlapping pooling can also be used, with the stride set to 2 and the kernel size set to 3. Because of the disruptive nature of pooling, having a kernel size greater than 3 would typically result in a substantial reduction in model efficiency.

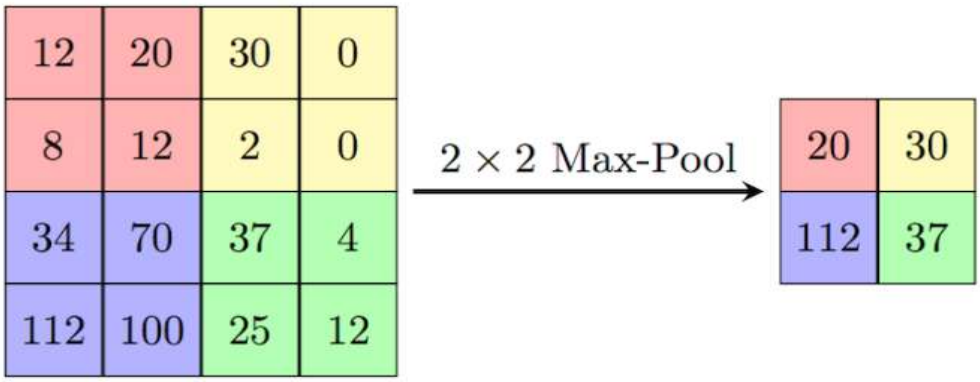


Figure 2.7 Max pooling

It is also important to understand that beyond max-pooling, CNN architectures may contain general-pooling. General pooling layers are comprised of pooling neurons

that are able to perform a multitude of common operations including L1/L2-normalisation, and average pooling [69].

2.8.3 Fully-connected layer

The output feature maps of the final convolution or pooling layer is typically flattened, i.e., transformed into a one-dimensional (1D) array of numbers (or vector), and connected to one or more fully connected layers, also known as dense layers, in which every input is connected to every output by a learnable weight. Once the features extracted by the convolution layers and down sampled by the pooling layers are created, they are mapped by a subset of fully connected layers to the final outputs of the network, such as the probabilities for each class in classification tasks. The final fully connected layer typically has the same number of output nodes as the number of classes. Each fully connected layer is followed by a nonlinear function [69]. The fully connected layer is illustrated figure 2.8.

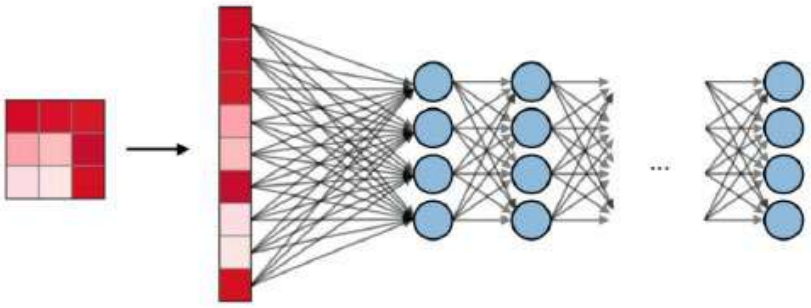


Figure 2.8 A fully connected layer

- **Last layer activation function:** The activation function applied to the last fully connected layer is usually different from the others. An appropriate activation function needs to be selected according to each task. An activation function applied to the multiclass classification task is a softmax function which normalizes output real values from the last fully connected layer to target class probabilities, where each value ranges between 0 and 1 and all values sum to 1 [69].

2.9 ImageNet

The dataset imagenet is a dataset of over 15 million labeled high-resolution images belonging to roughly 22,000 categories [70]. The images were collected from the web and labeled by human labelers using Amazon's Mechanical Turk crowd-sourcing tool. Starting in 2010, as part of the Pascal Visual Object Challenge, an annual competition called the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) has been held. ILSVRC uses a subset of ImageNet with roughly 1000 images in each of 1000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images. ILSVRC-2010 is the only version of ILSVRC for which the test set labels are available. On ImageNet, it is customary to report two error rates: top-1 and top-5, where the top-5 error rate is the fraction of test images for which the correct label is not among the five labels considered most probable by the model [71].

AlexNet is the name of a convolutional neural network which famously won the 2012 Imagenet LSVRC-2012 competition by a large margin. The architecture is depicted in Figure 2.9.

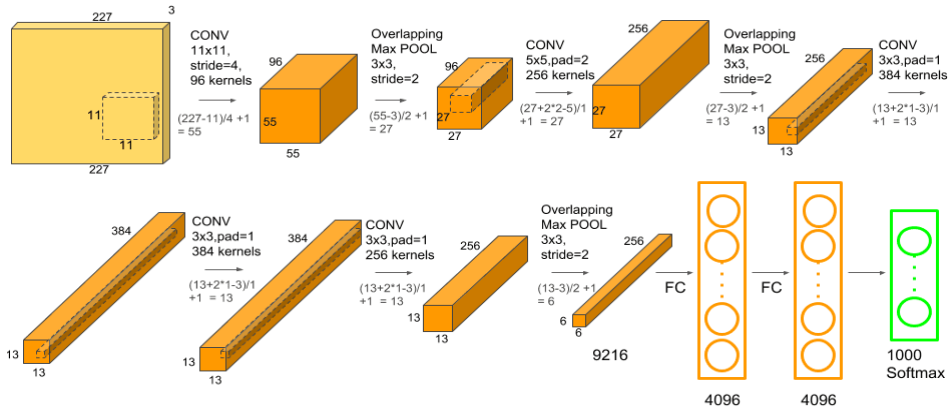


Figure 2.9 The architecture of the alexnet (CNN)

The alexnet has eight weighted layers, the first five of which are convolutional and the last three of which are fully connected. The last fully-connected layer's output is sent into a 1000-way softmax, which generates a distribution across the 1000 class labels. The network maximizes the multinomial logistic regression objective, which is equivalent to maximizing the average across training cases of the log-probability of the correct label

under the prediction distribution. Only those kernel mappings in the preceding layer that are on the same GPU are connected to the kernels of the second, fourth, and fifth convolutional layers. The kernels of the third convolutional layer are connected to all kernel maps in the second layer. The neurons in the fully-connected layers are connected to all neurons in the previous layer. Response-normalization layers follow the first and second convolutional layers. Max-pooling layers, of the kind described in Section 3.4, follow both response-normalization layers as well as the fifth convolutional layer. The relu non-linearity is applied to the output of every convolutional and fully-connected layer. The first convolutional layer filters the 224×224×3 input image with 96 kernels of size 11×11×3 with a stride of 4 pixels. The second convolutional layer takes as input the (response-normalized and pooled) output of the first convolutional layer and filters it with 256 kernels of size 5 × 5 × 48. The third, fourth, and fifth convolutional layers are connected to one another without any intervening pooling or normalization layers. The third convolutional layer has 384 kernels of size 3 × 3 × 256 connected to the (normalized, pooled) outputs of the second convolutional layer. The fourth convolutional layer has 384 kernels of size 3 × 3 × 192, and the fifth convolutional layer has 256 kernels of size 3 × 3 × 192. The fully-connected layers have 4096 neurons each [71].

In our work we used this architecture of alexnet which is the architecture of the convolutional neural network.

2.10 Conclusion

In this chapter, we reviewed the principles, approaches, and implementations of machine learning, as well as its algorithms. We also discussed the artificial neural networks that has been considered to be one of the most powerful tools over the last few decades, and has become very popular in the literature as it is able to handle a huge amount of data. And we reviewed as well one of the most popular deep neural networks which is the CNN. That has an excellent performance in machine learning problems, specially the applications that deal with image data.

Chapter 3 Implementations and Results

3.1 Introduction

The Most methods used in person re-identification are based on metric learning and feature representation. Many efforts have been made in both directions to tackle the challenge of person re-identification. An effective feature representation should be robust to illumination and viewpoint changes, and a discriminant metric should be learned to match various person images.

In this chapter we are going to discuss and evaluate and compare the performances of a person re-identification system using the efficient feature representation based on convolutional neural network, and an efficient metric learning method called c-view quadratic discriminant analysis (XQDA). We are going to run a set of experiments with different settings and parameters on four challenging person re-identification databases, VIPeR[2], GRID[3], CUHK01 [4] and PRID450s [5].

3.2 General description of the proposed approach

Person re-identification is the problem of identifying a person from a gallery with the same identity as the probe. There are two main critical problems for person re-identification, feature representation and metric learning. In our work we used convolutional neural network based imagenet dataset, and metric learning method called cross-view quadratic discriminant analysis. This method is shown to be effective and efficient through our experiments on four databases, VIPeR [2], GRID [3], CUHK01 [4] and PRID450s [5].

On a target person re-identification dataset, we used CNN features. From the learned CNN's fully connected layer (FC6), we extracted a 4,096 dimensional feature vector for each image on the dataset as shown in figure 3.1. The target dataset is then subjected to metric learning XQDA to improve discriminative power for person re-identification [72]. The CNNs has specialization for being able to detect patterns and make sense of them; this pattern detection makes CNN useful for image analysis.

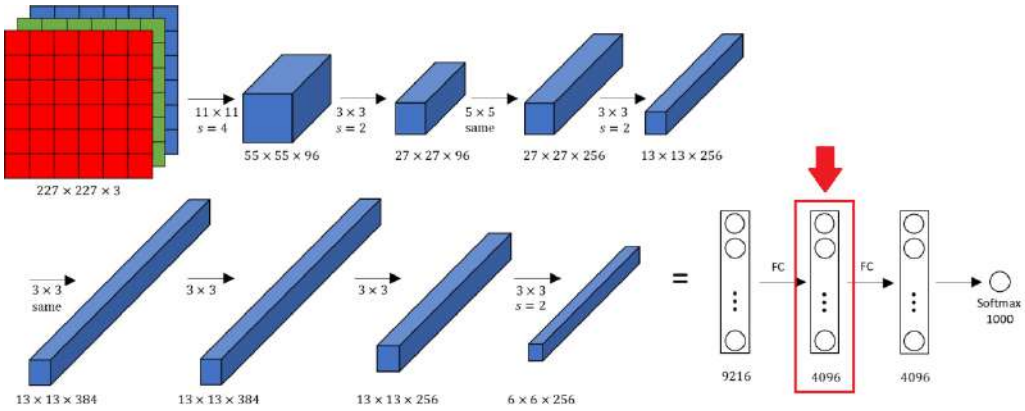


Figure 3.1 The extraction of a 4,096 dimensional feature vector for each image on the dataset.

As the variances in samples are typically hard to explore simply with the extracted high-dimensional features, it is essential for re-identification tasks to find a good distance metric. We used the mahalanobis distance function; it is a prominent and widely used approach for improving classification results by exploiting the structure of the data [73]. Given n data points $\boldsymbol{x}_i \in \mathbb{R}^m$, the goal is to estimate a matrix $\boldsymbol{M}$ such that the distance metric can be written as

$$d_M(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j). \qquad \textbf{(3.1)}$$

Where $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are two vectors (samples), and $\boldsymbol{M}$ is a positive semidefinite matrix.

Because the mahalanobis distance is calculated based on global samples, mutual interference between variables is ruled out. However, when calculating the mahalanobis distance, the required training samples are more than the sample dimensions, otherwise the resulting covariance matrix will be biased [74]. Then, we applied a score normalization function, it is a processing step that is applied to remove nuisance bias and variation that would otherwise influence comparison scores in diverse environmental conditions [75], which optimized our results and it is defined as:

$$N = \frac{x - x_{min}}{x_{max} - x_{min}}. \qquad \textbf{(3.2)}$$

Where $\boldsymbol{x}$ is a vector (sample).

Finally, for evaluating the performance of our work, we used the average cumulative matching characteristic curves to show the rank matching rates. CMC are one of the most used metric to evaluate identification performance. A rank r matching rate indicates the percentage of probe images with correct matches found in the top r ranks against the gallery images. Also it

helps in the visual comparison of different methods. The overall system of person re-identification is illustrated in figure 3.2.

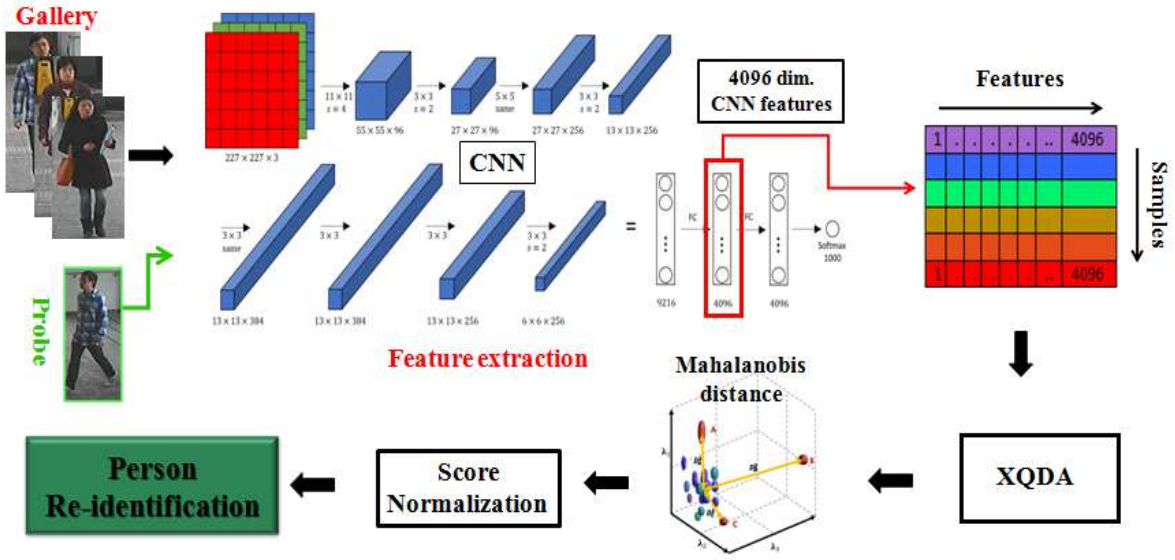


Figure 3.2 Person re-identification system consisting of four stages: training data, feature extraction, metric learning and ranking.

3.3 Features projection using XQDA algorithm

The cross-view quadratic discriminant analysis metric learning is an improvement version of the kiss metric [76], which is derived from a perspective of probability statistics. The XQDA considers learning a discriminant low dimension feature subspace with Fisher criterion [77]. Usually, the original feature dimensions $d$is large, and a low dimensional space $\mathbb{R}^r (r < d)$ is preferred for classification. Let $\boldsymbol{X_s} \epsilon \mathbb{R}^{d \times n}$ be a set of intra-personal difference of the n similar pairs, and be $\boldsymbol{X_D} \epsilon \mathbb{R}^{d \times n}$ a set of extra-personal difference of the n dissimilar pairs, where d is the dimension of the feature vectors. In$\boldsymbol{X_s}$, each column is a sample difference of a similar pair, while each column is a sample difference of a dissimilar pair in $\boldsymbol{X_D}$. The covariance matrices can be calculated as:

$$\sum_s = \frac{1}{n} X_s X_s^T \ , \ \sum_D = \frac{1}{n} X_D X_D^T \ . \qquad \textbf{(3.3)}$$

Given any training samples $\boldsymbol{u}, \boldsymbol{v}$ from two camera views, the XQDA metric calculates the distance as follow

$$d(u,v) = (u-v)^T W (\sum_{S'}^{-1} - \sum_{D'}^{-1}) W^T (u-v). \qquad \textbf{(3.4)}$$

Where $\boldsymbol{W}$ is a subspace projection matrix that to be learned.

$\sum_{S'} = W^T \sum_S W$ And$\sum_{D'} = W^T \sum_D W$ are the covariance matrices in subspace. Similar to the KISS metric, the variances of subspace can still be used to distinguish intra-personal and extra-personal difference. Therefore, the projection direction $\boldsymbol{W}$ can be optimized by maximizing

$$J(W) = \frac{W^T \sum_S W}{W^T \sum_D W}. \quad \textbf{(3.5)}$$

Equation (3-5) can be converted into the generalized Eigen value decomposition problem as LDA. The largest $\boldsymbol{r}$ corresponding eigenvectors of $\sum_S^{-1} \sum_D$ constitute the projection matrix $W = [w_1, w_2, \ldots, w_r]$. The final metric matrix is denoted as $M = \sum_{S'}^{-1} - \sum_{D'}^{-1}$. which can be used to calculate the distance of any two data points in the subspace [77].

3.4 Implementations and results

3.4.1 Databases and protocols

We evaluated the performance of the CNN features using four challenging person re-identification databases, VIPeR[2], GRID[3], CUHK01[4], and PRID450s[5] that will be described below.

- **PRID450S dataset [5]:** Person Re-ID 450S dataset was captured from 2 camera views. Contains 450 image pairs recorded from two different, static surveillance cameras. Additionally, the dataset also provides an automatically generated, motion based foreground/background segmentation as well as a manual segmentation of parts of a person. All images are scaled to 128 × 48 pixels. Each subject appears in single image at each camera (single-shot).

The main challenges are related to changes in viewpoint, pose as well as significant differences in background and illumination as shown in Figure 3.3.



Figure 3.3 sample image pairs from the PRID450S dataset.

- **VIPeR dataset [2]:** The VIPeR dataset is a challenging person re-identification database that has been widely used for benchmark evaluation [78]. It was captured from two camera views. Contains 1264 images of 632 identities. Each pair contains two images of the same individual seen from different viewpoints. Each image pair was taken from an arbitrary viewpoint under varying illumination conditions. The data were collected in an academic setting over the course of several months. Each image has been scaled to be 128x48 pixels.

  Figure 3.4 shows sample image pairs from the VIPeR dataset. The widely adopted experimental protocol on this database is to randomly divide the 632 pairs of images into half for training and the other half for testing, and repeat the procedure 10 times to get an average performance.



Figure 3.4 Sample image pairs from the VIPeR dataset.

- **GRID dataset [3]:** The GRID dataset was captured from 8 disjoint camera views. It contains 250 pedestrian image pairs, with each pair contains two images of the same person from different camera views (see figure 3.5). Besides, there are 775 additional images that do not belong to the 250 persons which can be used to enlarge the gallery set. An experimental setting of 10 random trials is provided for the GRID dataset. For each trial, 125 image pairs are used for training, and the remaining 125 image pairs, as well as the 775 background images are used for test [78].



Figure 3.5 A four random images from the gallery and the probe of the GRID dataset.

- **CUHK01 dataset [4]:** The CUHK01 dataset was captured from two camera views, with higher resolution, containing 3884 images of 971persons, and each person has two images in each view. We randomly choose 486 pairs of images for training and leave the rest for testing [79]. Each image has been scaled to be 160x60 pixels. Figure 3.6 shows random images from the CUHK01 database of two persons that has two images in each view.



Figure 3.6 Sample image pairs from the CUHK01 dataset.

We divide each dataset into training set and test set randomly, with each set containing half of the available individuals. In all datasets, the number of probe images is equal to the number of gallery images. Note that we've added an extra 775 images to the gallery set for the GRID dataset. The evaluation procedure is repeated 10 times and we report the (CMC) curves to evaluate the performance of the CNN features using four person re-identification datasets; VIPeR [2], GRID [3], CUHK01 [4], and PRID450s [5]. Figure 3.7, 3.8, 3.9 and 3.10 shows the CMC curves of the four datasets that we used without normalization and with normalization.

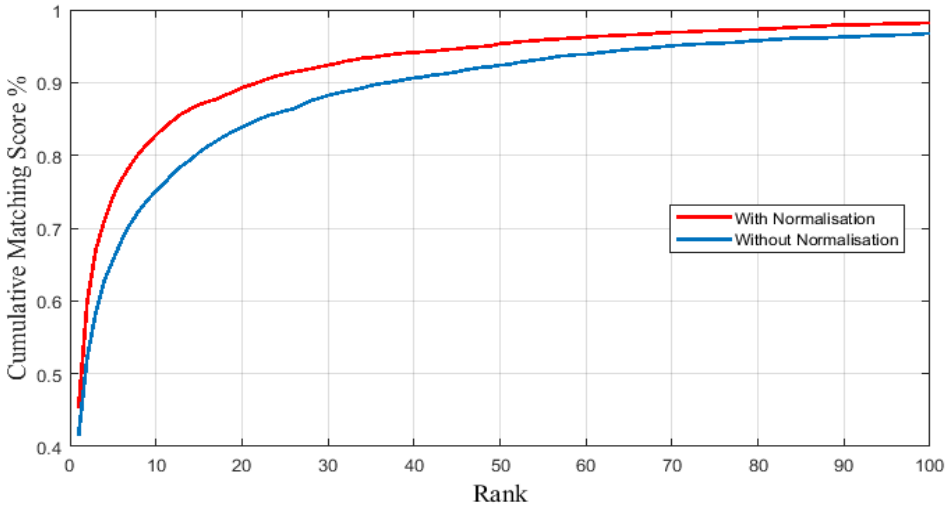


Figure 3.7 CMC curve of CUHK01 dataset.

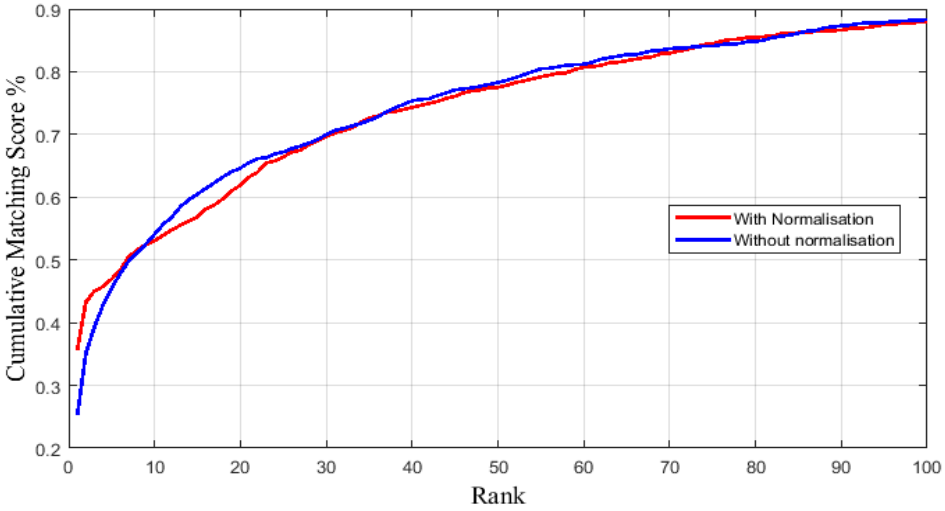


Figure 3.8 CMC curve of GRID dataset.

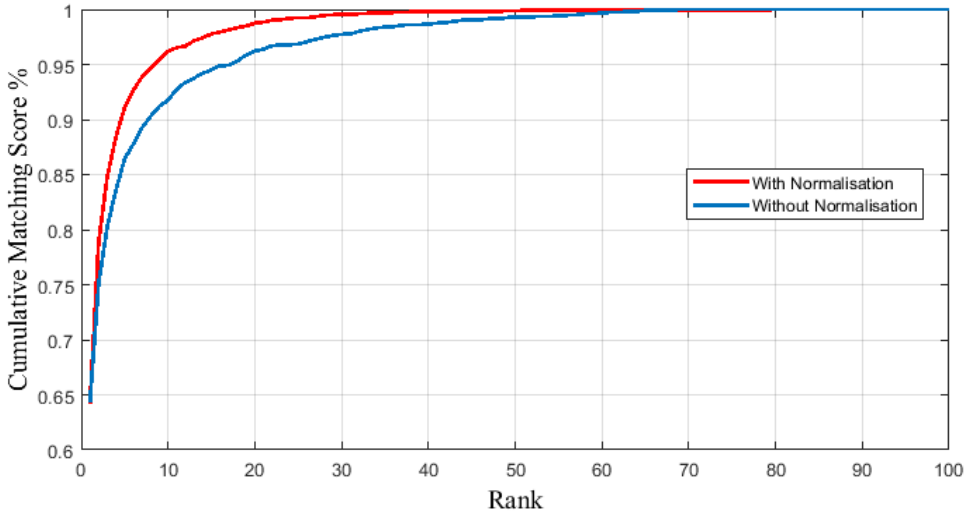


Figure 3.9 CMC curve of PRID450S dataset.

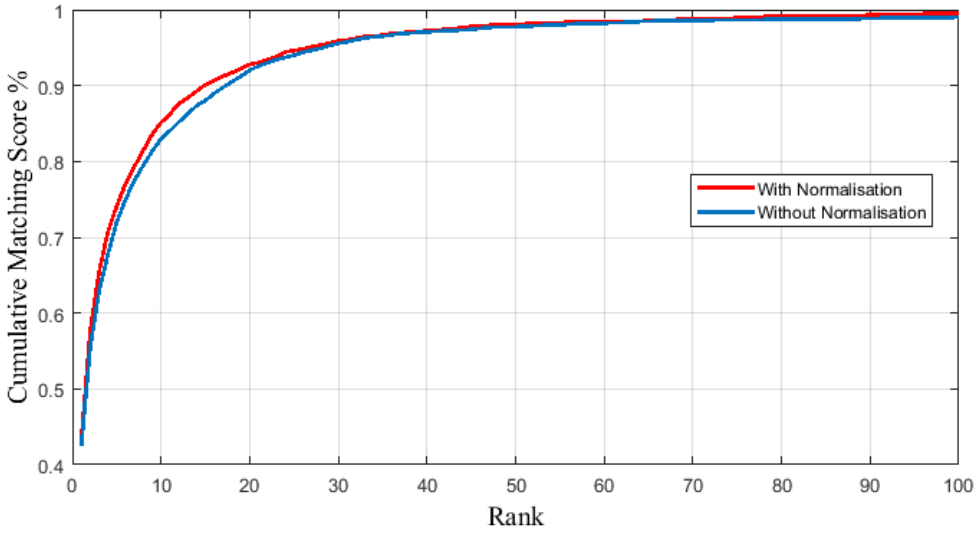


Figure 3.10 CMC curve of VIPeR dataset.

The CMC score measures the quality of identifying the correct match at each rank. In addition, the matching rate of rank-1 represents the probability that the top-ranked search result contains the correct images.

The experimental results comprising of re-identification accuracies attained for all given datasets at different ranks (rank-1, rank 5, rank-10 and rank-20) are in table 3.1.

Table 3.1 CMC scores at r = 1, 5, 10, 15 and 20 of the used databases; VIPeR, GRID, CUHK01, and PRID450s with and without normalization.

	CUHK01				
	RANK=1	RANK=5	RANK=10	RANK=15	RANK=20
Without Normalization	41.51%	65.61%	75.11%	80.39%	83.90%
With Normalization	**45.24%**	74.35%	82.78%	86.95%	89.30%
	GRID				
	RANK=1	RANK=5	RANK=10	RANK=15	RANK=20
Without Normalization	25.20%	45.52%	54.08%	60.48%	64.64%
With Normalization	**35.68%**	46.96%	53.04%	56.88%	61.92%
	PRID450S				
	RANK=1	RANK=5	RANK=10	RANK=15	RANK=20
Without Normalization	64.31%	86.44%	91.78%	94.53%	96.22%
With Normalization	**64.22%**	91.11%	96.18	97.78%	98.76%
	VIPeR				
	RANK=1	RANK=5	RANK=10	RANK=15	RANK=20
Without Normalization	42.16%	71.99%	83.01%	88.04%	92.03%
With Normalization	**43.16%**	74.11%	85.13%	90.13%	92.78%

3.4.2 Discussions

The experiments that we conducted using the proposed approach XQDA metric learning along with the mahalanobis distance metric to make use of its advantages for improving classification by exploiting the structure of the data. Then, to extract features from the four datasets VIPeR[2], GRID[3], CUHK01[4], and PRID450s[5], We have used the CNN that shows quite satisfactory results. Then a score normalization has been applied to remove nuisance bias and variation that would influence comparison scores on the four mentioned datasets, which lead to efficient results in the CMC curves shown previously. In order to evaluate the performance of the score normalization on the four mentioned datasets, we created graphic columns that demonstrates rank-1 performance of the four datasets VIPeR [2], GRID [3], CUHK01 [4], and PRID450s [5], with and without normalization, and it is illustrated below in figure 3.11.

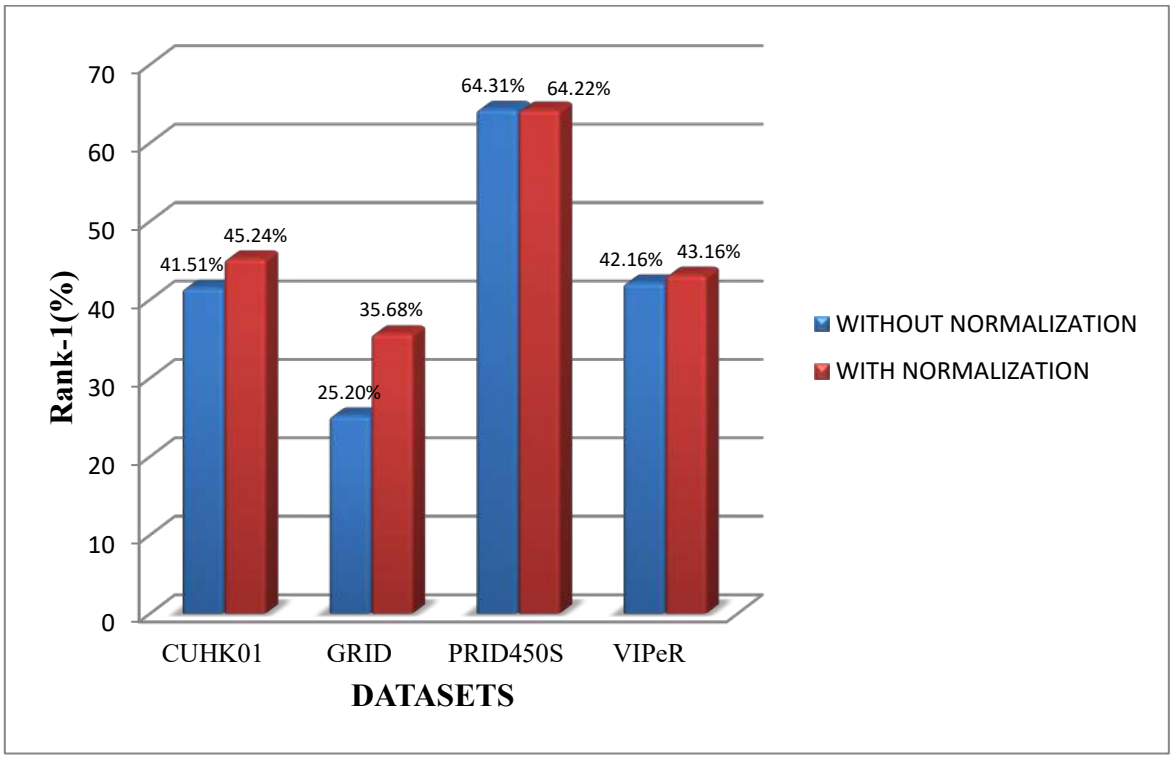


Figure 3.11 the rank-1 performance on the four datasets CUHK01, PRID450S, GRID and VIPeR, with and without normalization.

When we applied a score normalization on the resulting Mahalanobis distance, as we can see through the graph above, the rank-1 rates of VIPeR[2], GRID[3], CUHK01[4], and PRID450s[5] databases improves on three out of four databases, the results without normalization were respectfully 42.16%, 25.20%, 41.51% and 64.31%, then with normalization the results improves to become respectfully 43.16%, 35.68%, 45.24% and 64.22%, for PRID450s database as we can see the result of rank-1didn't improve, it decreased by 0.09% which is a small amount compared to the effectiveness of the score normalization on the other three databases, but its performance on the other ranks; rank-5, rank-10, rank-15 and rank-20 on the same database were remarkable as we can see through table 3-1 above. GRID also shows quite an improvement on rank-1 after the score normalization, it improved by 10.48%, which is a huge amount considering its complexity compared to the other three databases. We speculate that the use of score normalization on our datasets largely contributes to improve the results.

3.4.3 Comparison with the stat of the art

We compared the performance of the proposed approach to the state-of-the-art rank-1 results reported on the four datasets VIPeR [2], GRID [3], CUHK01 [4], and PRID450s [5], which are summarized in table 3-2 below. From Table 3.2 it can be observed that the Comprehensive experimental results on the used Re-ID datasets in terms of mean accuracy gave an optimistic performance especially the result of the dataset GRID [3] that gave the best performance (35.68%)compared to the other methods that are certified by the Institute of electrical and electronics engineers (IEEE).

Table 3.2 Comparison with the state-of-the-art of rank-1 identification rates (%) on the CUHK01, PRID450s, GRID and VIPeR databases.

Methods	**VIPeR**	**CUHK01**	**PRID450S**	**GRID**
	Rank-1			
FT-CNN+XQDA 2016 [72]	42.50%	46.80%	58.20%	25.20%
HIPHOP+LOMO 2017 [80]	54.20%	78.80%	-	26.00%
Hybrid+XQDA 2018 [81]	50.50%	68.30%	66.50%	25.50%
Kernel X-CRC 2019 [82]	51.60%	61.20%	68.80%	26.60%
EML 2020 [83]	44.37%	67.37%	63.58%	19.47%
CNN+XQDA with score normalization 2021 (Ours)	**43.16%**	**45.24%**	**64.22%**	**35.68%**

3.5 Conclusion

In this chapter we have presented an effective method for person re-identification. We've used an efficient descriptor called CNN and a metric learning approach called XQDA. In order to improve the performance of person re-identification system we applied score normalization on the resulting mahalanobis distance of the four challenging person re-identification databases, VIPeR, GRID, CUHK01 andPRID450s. Finally, to evaluate our results we've reported the Cumulative Matching Characteristic curves and we compared with the stat of the art.

General Conclusion

Person re-identification is a new field that was developed to ensure security and safety to public. We can follow a person's path and ensure that nothing illegal or inappropriate is done by employing re-identification and tracking models. Re-identification is a difficult problem since humans are frequently captured by many low-resolution cameras, under occlusion conditions, in poor lighting, and in a variety of stances. In this case, robust modeling of a person's full body appearance is required, especially when other traditional biometric indicators are unavailable or impossible to capture due to sensor resolution or frame rate limitations. Researchers are trying to improve the performance of the person re-identification system by developing a robust feature representation and a discriminant metric learning.

The goal of our work was to get a better approach in terms of accuracy for the person re-identification system by applying the score normalization function on the Mahalanobis distance on four challenging datasets VIPeR [2], GRID [3], CUHK01 [4] and PRID450s [5], after using the Convolutional Neural Network as a feature representation due to its powerful feature learning abilities and fitting capacity and the Cross-view Quadratic Discriminant Analysis as a metric learning method.

Our project can be summarized as follows:

- A general description of biometrics, biometrics' modalities and its applications, also the literature of person re-identification along with its pipeline, challenges and its application.
- An overview of Artificial Intelligence and machine learning, the classification of its algorithms and its applications.
- The use of a discriminant metric learning called Cross-view Quadratic Discriminant Analysis where it considers learning a discriminant low dimension feature subspace.
- The CNN was used as a feature representation and it shows optimistic results due to its ability and capacity on feature learning.

- The use of a score normalization function on the Mahalanobis distance that shows a great impact on the re-identification process on the four challenging datasets VIPeR [2], GRID [3], CUHK01 [4] and PRID450s [5].

Due to the promising performances of the score normalization process, it would be interesting -as a future work- to apply it with other feature representations like LOMO (Local Maximal Occurrence) and the GOG (Gaussian Of Gaussian) descriptor. It is also interesting to evaluate our approach based on XQDA to other cross-view matching problems, like the heterogeneous face recognition.

References

1. Guo, J., et al., Person re-identification by weighted integration of sparse and collaborative representation. IEEE Access, 2017. **5**: p. 21632-21639.
2. Gray, D., S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. in Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS). 2007. Citeseer.
3. Liu, C., et al. Person re-identification: What features are important? in European Conference on Computer Vision. 2012. Springer.
4. Zhao, Z., B. Zhao, and F. Su, Person re-identification via integrating patch-based metric learning and local salience learning. Pattern Recognition, 2018. **75**: p. 90-98.
5. Bischof, P.M.R.a.M.H.a.M.K.a.C.B.a.H., Person Re-Identification. Advances in Computer Vision and Pattern Recognition. 2014, London, United Kingdom: Springer. 247--267.
6. ISO, I., Information technology–vocabulary–part 37: biometrics. International Standard ISO/IEC, 2012: p. 2382-37.
7. Belhadj, F., Biometric system for identification and authentication, 2017, Ecole nationale Supérieure en Informatique Alger.
8. Hassanien, A.E., et al., Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016. Vol. 533. 2016: Springer.
9. Dargan, S. and M. Kumar, A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. Expert Systems with Applications, 2020. **143**: p. 113114.
10. Sharma, A.K., A. Raghuwanshi, and V.K. Sharma, Biometric System-A Review. International Journal of Computer Science and Information Technologies, 2015. **6**(5): p. 4616-4619.
11. Ibrahim, D.R., A.A. Tamimi, and A.M. Abdalla. Performance analysis of biometric recognition modalities. in 2017 8th international conference on information technology (ICIT). 2017. IEEE.
12. Ben, X., et al., A general tensor representation framework for cross-view gait recognition. Pattern Recognition, 2019. **90**: p. 87-98.
13. Ghayoumi, M. A review of multimodal biometric systems: Fusion methods and their applications. in 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS). 2015. IEEE.
14. Gong, S., et al., The re-identification challenge, in Person re-identification. 2014, Springer. p. 1-20.
15. Zheng, L., Y. Yang, and A.G. Hauptmann, Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984, 2016.
16. Gheissari, N., T.B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. in 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). 2006. IEEE.
17. Gray, D. and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. in European conference on computer vision. 2008. Springer.

18. Zheng, L., et al. Scalable person re-identification: A benchmark. in Proceedings of the IEEE international conference on computer vision. 2015.
19. Wang, T., et al. Person re-identification by video ranking. in European conference on computer vision. 2014. Springer.
20. Zheng, L., et al. Mars: A video benchmark for large-scale person re-identification. in European Conference on Computer Vision. 2016. Springer.
21. Ye, M., et al. Specific person retrieval via incomplete text description. in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 2015.
22. Li, S., et al. Identity-aware textual-visual matching with latent co-attention. in Proceedings of the IEEE International Conference on Computer Vision. 2017.
23. Ye, M., et al., Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
24. Chen, Y., Person re-identification in images with deep learning, 2018, Lyon.
25. Gong, S. and T. Xiang, Person re-identification, in Visual Analysis of Behaviour. 2011, Springer. p. 301-313.
26. Madden, C., E.D. Cheng, and M. Piccardi, Tracking people across disjoint camera views by an illumination-tolerant appearance representation. Machine Vision and Applications, 2007. **18**(3): p. 233-247.
27. Farenzena, M., et al. Person re-identification by symmetry-driven accumulation of local features. in 2010 IEEE computer society conference on computer vision and pattern recognition. 2010. IEEE.
28. Bazzani, L., et al., Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters, 2012. **33**(7): p. 898-903.
29. Prosser, B.J., et al. Person re-identification by support vector ranking. in BMVC. 2010.
30. Zheng, W.-S., S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. in CVPR 2011. 2011. IEEE.
31. Avraham, T., et al. Learning implicit transfer for person re-identification. in European Conference on Computer Vision. 2012. Springer.
32. Hirzer, M., et al. Person re-identification by descriptive and discriminative classification. in Scandinavian conference on Image analysis. 2011. Springer.
33. Zheng, W.-S., S. Gong, and T. Xiang, Reidentification by relative distance comparison. IEEE transactions on pattern analysis and machine intelligence, 2012. **35**(3): p. 653-668.
34. Layne, R., et al. Person re-identification by attributes. in Bmvc. 2012.
35. Zhao, R., W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.
36. Bazzani, L., M. Cristani, and V. Murino, Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding, 2013. **117**(2): p. 130-144.
37. Li, W., R. Zhao, and X. Wang. Human reidentification with transferred metric learning. in Asian conference on computer vision. 2012. Springer.

38. Zheng, W.-S., S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. in 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012. IEEE.
39. Zheng, W.-S., S. Gong, and T. Xiang. Associating Groups of People. in BMVC. 2009.
40. Loy, C.C., T. Xiang, and S. Gong, Time-delayed correlation analysis for multi-camera activity understanding. International Journal of Computer Vision, 2010. **90**(1): p. 106-129.
41. Loy, C.C., T. Xiang, and S. Gong, Incremental activity modeling in multiple disjoint cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011. **34**(9): p. 1799-1813.
42. Ren, C.-X., et al., Domain adaptive person re-identification via camera style generation and label propagation. IEEE Transactions on Information Forensics and Security, 2019. **15**: p. 1290-1302.
43. Li, J., et al., Pose-guided representation learning for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
44. Nambiar, A., A. Bernardino, and J.C. Nascimento, Gait-based person re-identification: A survey. ACM Computing Surveys (CSUR), 2019. **52**(2): p. 1-34.
45. Zhang, L., et al., Facial expression analysis under partial occlusion: A survey. ACM Computing Surveys (CSUR), 2018. **51**(2): p. 1-49.
46. Wang, X. and R. Zhao, Person re-identification: System design and evaluation overview, in Person Re-Identification. 2014, Springer. p. 351-370.
47. Nie, W., et al., Single/cross-camera multiple-person tracking by graph matching. Neurocomputing, 2014. **139**: p. 220-232.
48. Henriques, J.F., et al. Exploiting the circulant structure of tracking-by-detection with kernels. in European conference on computer vision. 2012. Springer.
49. Suma, V., Computer vision for human-machine interaction-review. Journal of trends in Computer Science and Smart technology (TCSST), 2019. **1**(02): p. 131-139.
50. investopedia. Artificial Intelligence (AI). 2018; Available from: https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp.
51. Encyclopedia Britannica, Artificial intelligence. 2020, August 11.
52. Bullinaria, J.A., IAI: The Roots, Goals and Sub-fields of AI. URL https://www. cs. bham. ac. uk/~ jxb/IAI/w2. pdf, 2005.
53. spotlightmetal, Machine Learning - Definition and application examples. 2021.
54. Zhang, X.-D., Machine learning, in A Matrix Algebra Approach to Artificial Intelligence. 2020, Springer. p. 223-440.
55. Mohammed, M., M.B. Khan, and E.B.M. Bashier, Machine learning: algorithms and applications. 2016: Crc Press.
56. Sen, P.C., M. Hajra, and M. Ghosh, Supervised classification algorithms in machine learning: A survey and review, in Emerging technology in modelling and graphics. 2020, Springer. p. 99-111.
57. Geurts, P., A. Irrthum, and L. Wehenkel, Supervised learning with decision tree-based methods in computational and systems biology. Molecular Biosystems, 2009. **5**(12): p. 1593-1605.

58. analytic steps. Introduction to Linear Discriminant Analysis in Supervised Learning. 2021; Available from: https://www.analyticssteps.com/blogs/introduction-linear-discriminant-analysis-supervised-learning.
59. Dey, A., Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies, 2016. **7**(3): p. 1174-1179.
60. Jolliffe, I.T. and J. Cadima, Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016. **374**(2065): p. 20150202.
61. Narayanan, H., M. Belkin, and P. Niyogi. On the relation between low density separation, spectral clustering and graph cuts. in NIPS. 2006.
62. Zhou, X. and M. Belkin, Semi-supervised learning, in Academic Press Library in Signal Processing. 2014, Elsevier. p. 1239-1269.
63. Tan, C., et al. A survey on deep transfer learning. in International conference on artificial neural networks. 2018. Springer.
64. Abraham, A., Artificial neural networks. Handbook of measuring system design, 2005.
65. Patterson, J. and A. Gibson, Deep learning: A practitioner's approach. 2017: " O'Reilly Media, Inc.".
66. Gupta, N., Artificial neural network. Network and Complex Systems, 2013. **3**(1): p. 24-28.
67. Beysolow II, T., Introduction to deep learning using R: A step-by-step guide to learning and implementing deep learning models using R. 2017: Apress.
68. GitHub, Machine Learning Glossary. 2017.
69. Yamashita, R., et al., Convolutional neural networks: an overview and application in radiology. Insights into imaging, 2018. **9**(4): p. 611-629.
70. Krizhevsky, A., I. Sutskever, and G.E. Hinton, ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017. **60**(6): p. 84-90.
71. Krizhevsky, A., I. Sutskever, and G.E. Hinton, Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012. **25**: p. 1097-1105.
72. Matsukawa, T. and E. Suzuki. Person re-identification using CNN features learned from combination of attributes. in 2016 23rd international conference on pattern recognition (ICPR). 2016. IEEE.
73. Roth, P.M., et al., Mahalanobis distance learning for person re-identification, in Person re-identification. 2014, Springer. p. 247-267.
74. Wu, W., et al., Deep features for person re-identification on metric learning. Pattern Recognition, 2021. **110**: p. 107424.
75. Nautsch, A., et al., Privacy-preserving speaker recognition with cohort score normalisation. arXiv preprint arXiv:1907.03454, 2019.
76. Koestinger, M., et al. Large scale metric learning from equivalence constraints. in 2012 IEEE conference on computer vision and pattern recognition. 2012. IEEE.
77. Yang, Z., et al. Incremental XQDA Metric Learning for Person Reidentification. in 2018 IEEE International Conference on Information and Automation (ICIA). 2018. IEEE.

78. Liao, S., et al. Person re-identification by local maximal occurrence representation and metric learning. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
79. Zhao, C., et al., Similarity learning with joint transfer constraints for person re-identification. Pattern Recognition, 2020. **97**: p. 107014.
80. Chen, Y.-C., et al., Person re-identification by camera correlation aware feature augmentation. IEEE transactions on pattern analysis and machine intelligence, 2017. **40**(2): p. 392-408.
81. Perwaiz, N., M.M. Fraz, and M. Shahzad, Person re-identification using hybrid representation reinforced by metric learning. IEEE Access, 2018. **6**: p. 77334-77349.
82. Prates, R. and W.R. Schwartz, Kernel cross-view collaborative representation based classification for person re-identification. Journal of Visual Communication and Image Representation, 2019. **58**: p. 304-315.
83. Ma, W., et al., A New Date-Balanced Method Based on Adaptive Asymmetric and Diversity Regularization in Person Re-Identification. International Journal of Pattern Recognition and Artificial Intelligence, 2020. **34**(09): p. 2056004.