APPLICATION OF RE-IDENTIFICATION ALGORITHMS TO EXTRACT TRAJECTORY OF PEDESTRIANS

by

MOHAMMAD NIROU JAZI

A thesis submitted to the
Department of Computer Science
in conformity with the requirements for
the degree of Master of Science

Bishop's University
Canada
January 2023

Abstract

Thousands of video frames and images are being recorded and captured every day. Because of the rapid expansion of the use of surveillance cameras, person re-identification (PRI) on different non-overlapped cameras and person tracking for trajectory extraction has become even more important. This thesis proposes a new model using CNN and transfer learning, namely extending ResNet50 and VGG16 as the base models, and modifying the structure to overcome aforementioned challenges. We call the new model as ResNet-PRI and VGG-PRI respectively and we extract deep features from these proposed models. We also propose another method that increases the person re-identification efficiency even more. The proposed method combines deep features extracted from ResNet-PRI and VGG-PRI with the hand-crafted features called Gaussian of Gaussian (GOG) descriptors, using both learnt and hand-crafted characteristics. The person re-identification (PRI) problem is then described as a similarity problem that output the most identical individuals, obviating the necessity for metric learning algorithms in the process. The results are evaluated on three databases, namely CUHK01, CUHK03 and GRID. Experimental results show that our proposed method achieves better precision in most ranks compared to the other state-of-the-art models. This thesis proposes clear methods for simplifying the process and increasing the accuracy of the problem of PRI and extracting the trajectory of pedestrian from videos captured by mounted cameras in the city by focusing on the appearance features of the target.

Keywords: Artificial Intelligence, Machine learning, Deep Learning, Person re-identification, Person tracking, CNN

Acknowledgements

I want to take some time to thank from the people and organizations that have helped me to come to this point of my life which I am grateful for. First, I want to thank my parents, and my brother for their true support and kindness and motivating me through my whole journey. It would not be possible to come to this end without them. Next, I would like to thank my supervisors, professor Madjid Allili and professor Mohammed Ayoub Alaoui Mhamdi for their support and guidance in writing my thesis. Then, I would like to thank Bishops university and specially the computer science department for giving me the opportunity to continue my graduate studies in the field that I am truly passionate about. Next, I would like to thank you the graduate entrance scholarship committee of Bishops university and MITACS foundation for the financial support that I received during writing my thesis which greatly helped me to focus on my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Importance of PRI algorithms to extract trajectory of pedestrians

Due to recent progress in object detection, tracking-by-detection has become the leading paradigm in multiple object tracking algorithms. When tracking a person, we first detect person and now we have a bounding box containing the person. Next step is comparing this bounding box which is containing the person with the extracted persons from previous frames or images from cameras that may have non-overlapping viewpoints. To track the person, if the person has been seen before, the already existing ID will be assigned to the new detection and if not, a new unique ID will be assigned to the detected person and it is considered as new detection.

Given a video, consider that we detect a new person in a current frame of the video. Then, we can focus on different aspects and measures to say whether the person has been seen in previous frames or not. One approach can be focusing on the bounding box coordinates and location within 2 or more frames, and this approach only can be used for videos. This approach will be discussed in more details in the next section. The second approach which is our main focus in this thesis, is focusing on the visual features of the person detected by the bounding box in order to track targets within consecutive frames or images of non-overlapping cameras. Actually, this approach suggests using a person re-identification (PRI) method for the tracking task. This approach which is our main focus in this thesis, is focusing on the visual features of the person detected by the bounding box in order to re-identify targets within consecutive frames or images of non-overlapping cameras. PRI algorithms can be used not only for tacking pedestrians in videos but also for re-identifying people from images of non-overlapping cameras. Please note that, focusing on visual features which is what is discussed in this thesis, is a more general approach as it can be used either for re-identifying people from different images of non-overlapping cameras or re-identifying pedestrians within different frames of video. It is also obvious that re-identifying person or pedestrians from non-overlapping images would be much more challenging for the PRI system as the system faces much more variations in viewpoint, illumination, and background cluttering compared to 2 consecutive frames within a video [8, 9].

In our case, we want to re-identify and track pedestrians, so that we can summarize the usage of our proposed method into:

- Given one query image, find the query person in gallery images (the collection of saved images till now which is basically a collection of all pedestrians that have been seen by camera up to now) from non-overlapping cameras

- Given a video, we want to find out which parts of the image depicts the same person in different frames. So, we want to automatically identify pedestrians in a video and interpret them as a set of trajectories with high accuracy.

To get more detailed information about what was said please read the following sections.

1.1.1 Visual representation VS. Statistical modeling for pedestrian tracking

When dealing with person tracking and re-identifying, somehow we need to learn about our target to be able to re-identify it and track it. In general, we can focus on two aspects. We can focus on either appearance or motion.

1. **Visual representation (Appearance):** It focuses on constructing robust features and representation that can describe the object. To simplify that, we need to know how the target looks like.

2. **Statistical modeling (Motion):** It uses statistical learning techniques to build mathematical models for object identification effectively. In other words, to make predictions of where the targets go. Motion estimation usually infers the predictive capability of the model to predict the object's future position accurately [8, 9].

Visual representation

Appearance representation deals with modeling the visual appearance of the object. Appearance modeling has to be conducted so that the algorithms can capture various changes and distortions introduced when the target object moves.

Visual object tracking is an important and fundamental subject of computer vision. [10]. It has diverse applications in numerous fields, such as video understanding, visual surveillance, augmented reality and human-computer interaction, etc. It is still a challenging task in some difficult scenes like illumination changes, deformation, background clutter, motion blur, occlusion, low-resolution, to name a few. [10]. Recently, Convolutional Neural Networks (CNN) have demonstrated the superior performance of feature extraction in various computer vision tasks including object classification detection, and segmentation. With the development of object tracking algorithms, deep features have also been developed to further pursue high performance. Compared to handcrafted features-based trackers show greater potential and have significant advantages in accuracy and robustness [10].

Statistical modeling approach

As the focus in this thesis is on the visual representation approach, in this section we briefly discuss the statistical modeling approach.

Motion estimation approximates the possible region where the object could most likely be present. Once the location of the object is approximated, we can then use a visual model to lock down the exact location of the target. In this thesis we focus on the visual representation which focuses on the object's visual appearance. Thanks to the advancements happening in the deep learning field, now we can take advantage of the deep models that can greatly help the overall performance of the tracking systems. These deep features can be solely used for the re-identification or can be used in combination with the statistical modeling algorithms to boost the performance of pedestrian tracking algorithms and systems.

Concerning this approach which is based on using the bounding box coordinates information, we can mention the Centroid Tracker algorithm [11] which works as described below:

1. Taking an initial set of object detections (such as an input set of bounding box coordinates)
2. Creating a unique ID for each of the initial detections
3. And then tracking each of the objects as they move around frames in a video, maintaining the assignment of unique IDs [11]

Basically, the Centroid algorithm considers a threshold for a circle around the detected object and if the detected object within the next frame falls under the defined threshold, then these bounding boxes are labeled as identical and the same ID is assigned to them. It is obvious that this algorithm can not handle the situations in which there is occlusion, and is not able to pick up objects it has lost in between frames. Also, this algorithm can not handle situations in which people are moving very closely or cross each other.

Another algorithm which falls into this category, is The Kalman filter [12, 13]. Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) solution of the least-squares method. The filter is very powerful in several aspects: it supports estimations of past, present, and even future states, and it can do so even when the precise nature of the modeled system is unknown. Kalman filter technique is used to estimate the state of a linear system where the state is assumed to be distributed as a Gaussian. In 1960, R.E. Kalman published a famous paper describing a recursive solution to the discrete-data linear filtering problem. Object tracking is performed by predicting the object's position from the previous information and verifying the existence of the object at the predicted position. Secondly, the observed likelihood function and motion model must be learnt by some sample of image sequences before tracking is performed [12, 13].

Considering all the above mentioned challenges, the researchers looked for methods to handle these barriers. We can overcome these issues by replacing the association metric with a more informed metric that combines motion and appearance information. In particular, it is possible to apply a convolutional neural network (CNN) that has been trained to discriminate pedestrians on a large-scale person re-identification dataset. Through integration of this network we can increase robustness against misses and occlusions while keeping the system easy to implement, efficient, and applicable to online scenarios [14].

1.2 What is tracking?

Object tracking refers to the ability to estimate or predict the position of a target object in each consecutive frame in a video once the initial position of the target object is defined [8] [9].

1.3 Object detection

Object detection is a computer vision technique used for locating instances of objects within images or videos. Object detection algorithms typically leverage machine learning or deep learning to produce meaningful results. When humans look at images or video, we can recognize and locate objects of interest within a matter of seconds. The end goal of object detection is to replicate this intelligence using a computer or so called machine. Advanced driver assistance systems (ADAS), which enable vehicles to detect traffic lanes or carry out pedestrian detection to increase road safety, rely heavily on object detection as a crucial technology. Application areas like video surveillance or image retrieval systems benefit from object detection. Figure 1.1 shows the input image to an object detection model and the output image with the bounding boxes marking the detected objects within the input image. The image is taken from the official GitHub repository of TensorFlow and open source DeepMAC architecture.

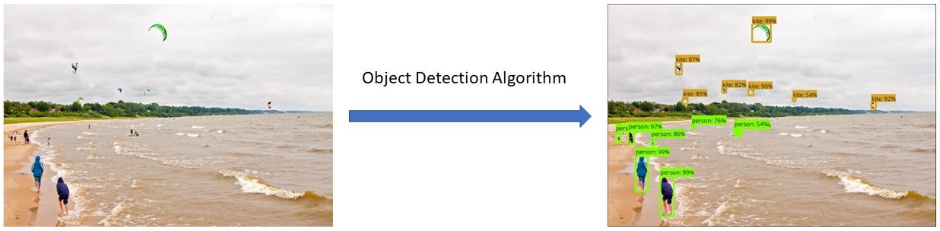


Figure 1.1: Using object detection algorithm to identify and locate objects

1.3.1 Object Detection Using Deep Learning

You can use a variety of techniques to perform object detection. Mobilenet SSD, R-CNN, and YOLO, three well known deep learning-based methods that use convolutional neural networks (CNNs), automatically learn to recognise objects in images. To use deep learning for object detection, we can choose between two main strategies:

- **Create and train a custom object detector:** To train a custom object detector from scratch, you need to design a network architecture to learn the features for the objects of interest. You also need to compile a very large set of labeled data to train the CNN if you choose a supervised approach. The results of a custom object detector can be remarkable. That said, you need to manually set up the layers and weights in the CNN, which requires a lot of time and training data.

- **Use a pretrained object detector:** Many object detection workflows using deep learning leverage transfer learning, an approach that enables you to start with a pretrained network and then fine-tune it for your application. This method can provide faster results because the object detectors have already been trained on thousands, or even millions, of images.

In this thesis, we use transfer learning technique.

1.4 Person Re-Identification (PRI)

Thousands of video frames are recorded every day as a result of the dramatic increase in the usage of security cameras and monitoring systems, creating enormous volumes of visual data. Strong, intelligent systems that can process images and videos are needed for handling this. One of the most difficult video processing problems is PRI, which aims to recognise a person in a camera on other, non-overlapping camera footage[15, 16]. In fact, the process of connecting people from many cameras at various locations and times is known as person re-identification (PRI)[17]. Each individual in the inquiry camera (probe) can be observed in other cameras in different orientations, illuminations, and viewpoints because these cameras are not overlapping. Since manual identification by operators is time-consuming and not accurate, automatic PRI methods are crucial. Strong PRI systems are used in domains including outdoor health care, urban control, criminal prevention, and video surveillance systems.

There are two primary types of PRI depending on the number of frames per person available: If there is only one frame per person in the probe set and gallery, PRI is a “single-shot”. To put it another way, in single-shot mode, one image or frame is available for the query person, and another one is available in the gallery set which is basically a collection of all pedestrians that have been seen by camera in previous frames and images. If there is more than one image, either in the probe or the gallery, the kind is known as “multi-shot” PRI [17]. The more photos that are available for a person, the more information is available for the PRI process.

There are two ways to handle the PRI problem. PRI is viewed as a ranking problem by many approaches, while it is viewed as a matching challenge by others. Re-identification of a person by ranking entails the system compiling a ranked list of images that are the most similar to the query person based on the query image(s) or frame(s) it receives. The top spots on the list go to the gallery member who is most likely to be the subject of the inquiry. This approach is similar to the image retrieval problem. Despite the ranking approach, the matching strategy tries to find an identical person in the gallery and probe folders. Simply put, the matching strategy is a type of binary classification in which two people are either identical or not [16, 17, 18].

Almost every PRI algorithm includes two main steps: first, feature extraction, and the second one is matching criteria or distance measuring. Feature extraction entails challenging and demanding image and video processing tasks. Despite the fact that feature extraction has been the subject of numerous studies, researchers continue to focus on it in order to improve picture representation due to the semantic gap between features and human concepts. Modern feature representation techniques are effective at classifying images and identifying shapes. The accuracy of PRI is still compromised by a number of issues. Similarity in subjects clothing color, making subjects biometric features to be distinguished in the PRI, long and varied subject distances from security cameras, poor camera footage, occlusion in crowded environments, and illumination and position changes on different cameras and times are some of the more significant difficulties [16, 17, 18].

The second stage of the PRI systems focuses on calculating the distance between the features that were extracted in the first stage . The resemblance between the subjects should be addressed in light of their characteristics, features, and representation; and to do that we can use similarity measurement tools, such as Euclidean distance [19]. Also, metric learning techniques have been favoured by many.[20, 21]. Similarity measurement is a crucial component of the PRI pipeline, just like feature extraction. Therefore, an effective feature extraction and precise similarity measurement should be used in a reliable PRI system. The objectives of the PRI systems are shown in Figure 1.2 .

Recognizing a person who has already been seen on a camera network is known as person re-identification. It is a difficult computer vision task that can offer practical tools for a variety of security applications of video surveillance, such as online tracking of individuals over various, non-overlapping cameras and off-line retrieval of the video sequences containing an individual of interest, whose image is given as a query. Since facial recognition is inefficient in video surveillance scenarios due to low image resolution and a diversity of positions, clothing appearance is the most frequently employed signal.

Modern re-identification techniques generate descriptors of clothing appearance based on the segmentation of the body into parts, and then capture a set of low-level features from each area of the body, such as SIFT points or small picture patches. Figure 1.3 shows a simpler representation of the a PRI system which tries to retrieve the person of interest across different cameras.

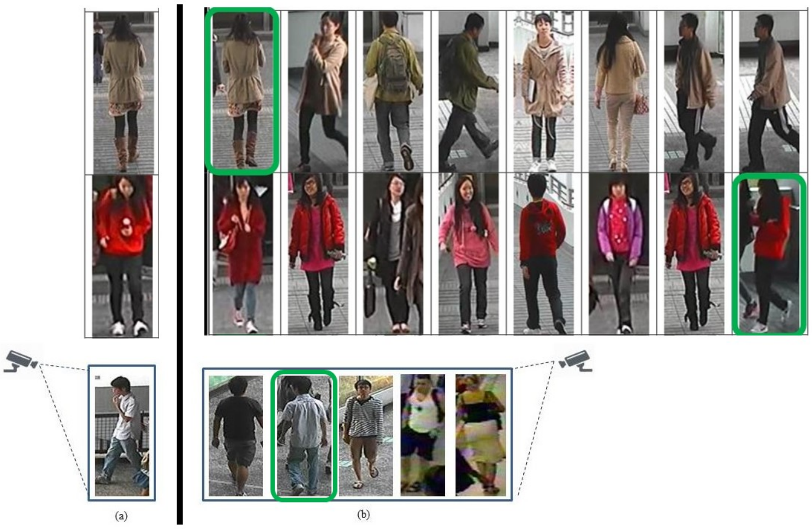


Figure 1.2: Aim of the PRI system: the probe person is seen via camera (a), and the same person between different persons of the gallery set should be identified correctly (b).

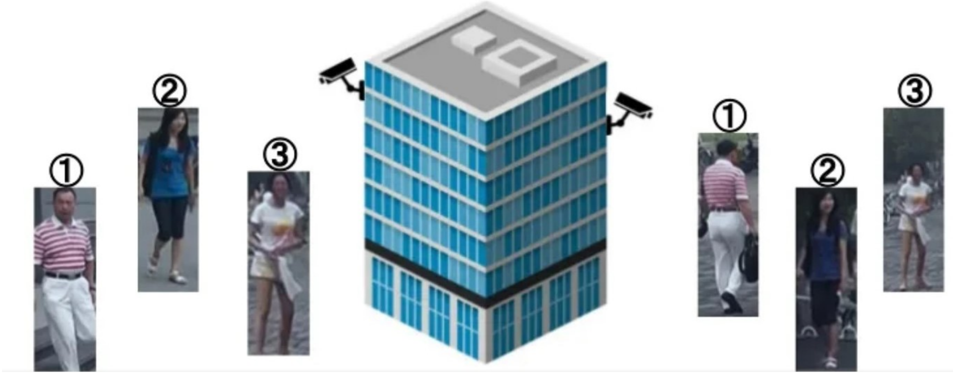


Figure 1.3: Aim of the PRI system: Retrieving the person of interest across different cameras [1]

1.4.1 A deeper look into PRI system

When a person is re-identified, it usually means that the person seen in one camera can be located and followed in another cameras that is not in direct line of sight with the original camera or simply put, the two cameras are no-overlapping. The number of individuals in the camera's field of vision is obviously very large in busy public spaces, so the subject must be picked out of all of these individuals. In other words, re-identifying a person entails contrasting the individual captured by the probe camera with a group of potential matches captured by cameras that do not overlap (non-overlapping cameras) with the probe camera [22].

Even though face recognition, smile detection, and other applications can now be implemented with high levels of accuracy thanks to new algorithms, there are still many difficulties in re-identifying people when using surveillance cameras because the distance between the subject and the camera is frequently variable and, in some cases, it is a very long distance. For example, in a surveillance system, a person leaving the frame of one camera will be seen in at least one other camera, which may have a different distance and angle compared to the first camera [17] and therefore the person's re-identification would be challenging. Therefore, the importance of re-identifying the individual is now clearer than before. PRI has many applications in industry. The most important of these are security and surveillance applications in the cities, airports, department stores, offices and other places that are controlled by surveillance cameras. Also, other important applications of PRI include its use in health and condition monitoring systems, long-distance tracking systems, identity acquisition using local image information, and crime prevention.

To categorize, In the most general form, person re-identification can be categorized into the following two groups according to the number of images (frames) related to the probe and gallery folders:

1. Single shot re-identification (single frame)
2. Multi shot re-identification (multi frame)

In single shot re-identification, there is only one photo of the requested person in the probe camera, and also in the image gallery, there is only one photo of each candidate. Therefore, the system is obliged to use a single photo as the input to find a photo in the gallery that belongs to the same person. But in multi-image recognition, there may be multiple images of each person in different modes, angles, and point of views. In this case, both, on the probe camera side and in the gallery, several frames per person can be found. However, the methods that have a single image on one side and several images of each person on the other side are also considered as multi-image methods. In video processing, most methods are multi-image. In Figure 1.4, you can see the difference between recognizing with a single image and multiple images.

As will be explained further, feature extraction from an image or frame plays an important role in re-identification of an individual. Due to the special conditions of the images captured by the surveillance cameras, an optimal system of individual re-identification should have some specific features, some of which will be mentioned below. As a first feature, the system must be illumination resistant.

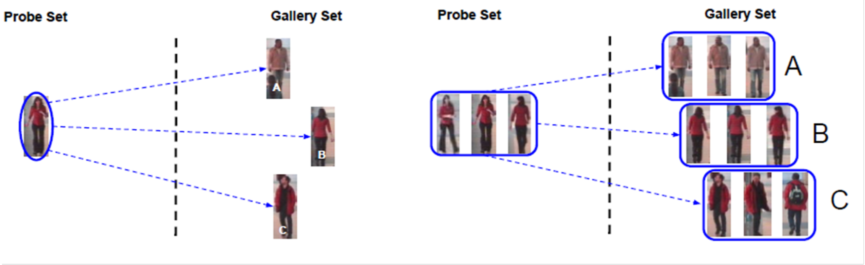


Figure 1.4: Two general modes of individual recognition, single-shot mode (left) and multi-shot mode (right).

In other words, the extracted features from the image should be such that the features do not change too much as the brightness of the image changes. Suppose a person is seen in the probe camera while in presence of enough light or in a lighter location. Now if the same person is in a darker location in other cameras (gallery), the system should recognize that this person is the same person seen in the probe camera. Another feature that is of particular importance is the resistance to the viewpoint. Naturally, each person is seen from different angles and viewpoints in different cameras installed at different angles and locations. Therefore, the characteristics that are extracted from each person must be resistant to changing the viewpoint so that each person can be recognized in any direction. Another necessary feature of the PRI system is its resistance to background cluttering. The background of the image taken from a person may be crowded and cluttered, making recognition more difficult. In addition, in crowded environments, blockages usually occur. In such a way that a part of the person's body is behind obstacles, and it is difficult to identify the person. But the system of individual recognition must also be resistant to the obstruction of the person and hiding behind obstacles, so that this obstruction does not prevent the correct identification.

Detection in low-resolution, low-quality images is also one of the tasks of an ideal recognition system. Because surveillance cameras in public places are usually of normal quality and the recorded images are usually of low resolution. Therefore, it is not possible to magnify the image until the biometric features of the person are clearly visible, and this is one of the main differences between person-re-identification and other areas of image processing. In Figure 1.5, a number of sample images related to the viper database can be seen. The low quality of the images and the changes in angle and other factors mentioned earlier are well evident in the images. This database will be explained in more detail in the following sections.

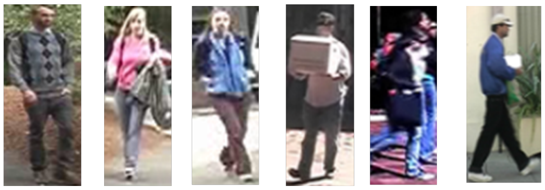


Figure 1.5: Some samples from Viper database [2]

In general, each PRI system consists of two main stages, the first stage is feature extraction and the second stage is dedicated to measuring the similarity between features. Although in some papers such as reference [17] in addition to the above steps, the creation of a suitable descriptor is recognized as a separate step in the re-identification of a person, but this step is generally not separate from the feature extraction step. Each of the above steps is of particular importance in the field of PRI. Therefore, in order to design a PRI system, in addition to ensuring the extraction of useful features, it is important to use an appropriate measurement criterion to identify similarities between individuals. Numerous studies have been conducted in the field of person re-identification, focusing on the extraction of appropriate characteristics. Also, successful research has been done in the field of similarity criterion design, which will be described in detail in the following sections and in the background of the research.

There are two different approaches to deal with the PRI problem. Many research have solved the problem from the perspective of ranking. In this case, upon receiving the image or frame of demand, the most similar to the person in demand are likely to be ranked in order of high to low, so that the less similarity is seen between the person seen in the probe image and the person in the gallery, then that person will be ranked lower. In the second method, the matching approach is used. The probe image is matched individually with each image in the gallery folder to see if the two are the same. Binary classification is usually used for this purpose. Although each of these approaches has advantages that are chosen according to the type of problem and its application, but naturally the ranking method is more comprehensive.

The main difference between PRI in video sequences and other areas of image recognition or image retrieval is that in image retrieval research in general, the number of classes or different classes is small and in each class, there are many samples which can be useful for the training purpose. But in PRI in the video, the number of classes is large (each person is a separate class) and the number of instances of each class is very small. Even in single-shot PRI mode, there is only one sample frame from each class (person). Therefore, this has created a major challenge in using training-based algorithms in this field.

1.5 Deep Learning Neural Networks

Machine learning algorithms and methods generate their own logic based on the input data. The algorithm learns by itself and code need not be written to solve every single problem. Typical example is categorizing emails into various bins such as input, spam, etc. Another important classification is for objects present in images. For example, an image contains the picture of an animal. The problem is to categorize the animal as deer, dog, cat, lion, etc. Large number of images with pictures of animals has to be fed as input to the algorithm from which it can learn to classify. If the images are already classified and used as input, it is supervised learning. If not, it is unsupervised learning. Machine learning algorithms and techniques basically classify based on patterns in the data. The data can be text, sound, image, etc. [3]

Neural networks have been found to be best suited for implementation of machine learning algorithms. Traditional neural networks have one input layer, one output layer and two or three hidden layers. Deep neural networks have one input layer, one output layer and hundreds of hidden layers, typically as shown in Figure 1.6. More number of hidden layers, deeper the network. The layers are interconnected, with the output of the previous layer being the input of the current layer. The inputs / outputs are weighted, and the weights determine the performance of the network. Training of the network involves acquiring the appropriate weights for the various layers. Deep networks need higher processing power, computing speed, large database and the appropriate software with parallel processing. [3]

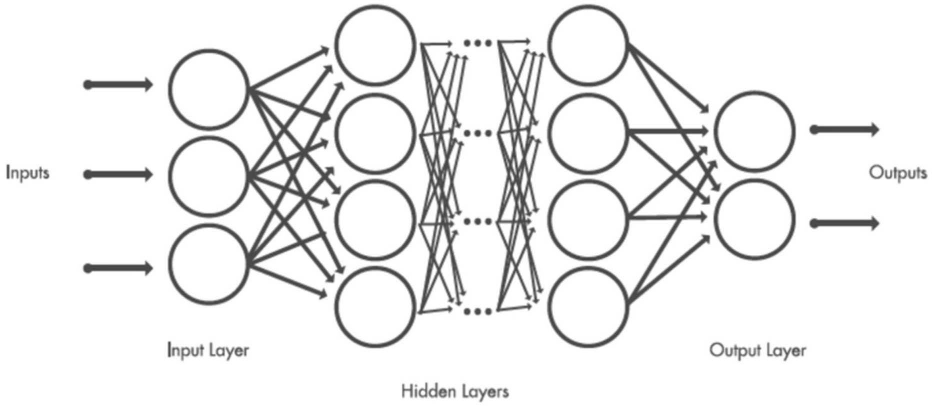


Figure 1.6: Typical Deep Neural Network [3]

Convolutional Neural Network (CNN) is a type of deep learning network that has become prominent for image classification. An example of CNN architectures are given in Figure 1.7. It consists of an input layer and hundreds of feature detection layers. For example, feature detection layers can carry out one of the following three operations: Convolution, Pooling, Rectified Linear Unit (ReLU). Convolution puts the image through convolution filters that detect certain features in the image. Pooling performs non-linear down sampling in order to reduce the number of data to be handled. Rectification Linear Unit maintains positive values and maps negative values to zero. The classification layer is the one before the output layer. It is a fully connected layer with N-dimensional output, N being the number of classes to be categorized. This layer output s a N-dimensional vector, each element of the vector is the probability that the input image belongs to one of the N classes. The final output layer uses a softmax function to give the classified output. Thousands or millions of photos must be sent into the network for accurate results. It requires higher computing power with several Graphics Processing Units (GPUs) operating in parallel. [3]

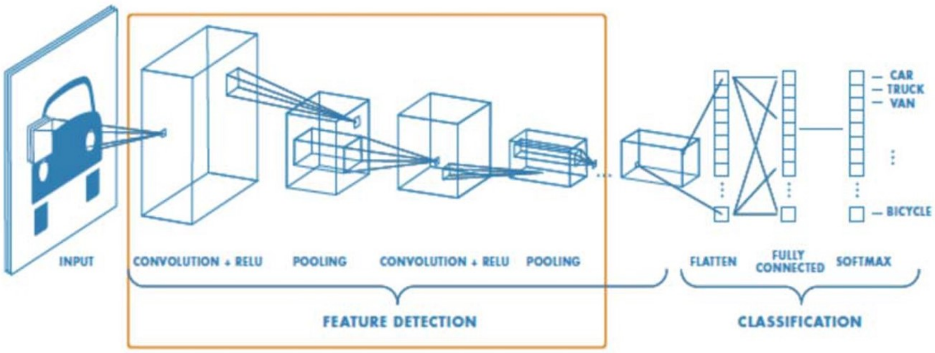


Figure 1.7: Convolutional Neural Network [3]

The deep learning network begins to recognise data features that can be utilised for categorization when millions of training photos are fed into it. Processing happens at each layer of the network and this is fed as input to the next consecutive layer. The biological structure of the visual cortex served as the basis for the design of CNNs. The simple and complex cells of visual cortex activate based on the subregions of a visual field, called receptive field. The neurons of a layer in CNN connect to the subregions of the previous layer instead of being fully connected. The neurons are not responsive to other subregions. The subregions are allowed to overlap and hence produce spatially correlated outcomes, unlike traditonal neural nets. This is the fundamental difference between CNN and other neural nets. The CNN reduces the number of parameters to be handled by reducing the number of connections, sharing the weights and by downsampling. [3]

1.5.1 BUILDING BLOCKS OF CNNs

In this section, we shall look at the basic building blocks of CNNs in general.

Max-Pooling

This operation can be thought of as a max filter, where each $n \times n$ region is replaced with its max value. This operation serves two purposes:

1. It picks out the highest activation in a local region, thereby providing a small degree of spatial invariance. This is analogous to the operation of complex cells.
2. It reduces the size of the activation for the next layer by a factor of n^2. With a smaller activation size, a smaller number of parameters need to be learned in the later layers.

Other types of pooling include average-pooling, winner-takes-all pooling, and stochastic pooling. However, these are not as commonly used as max-pooling. [23]

Fully-Connected Layers

A fully connected layer refers to a neural network in which each neuron applies a linear transformation to the input vector through a weights matrix. As a result, all possible connections layer-to-layer are present, meaning every input of the input vector influences every output of the output vector. In modern networks, using many fully connected layers are generally avoided as it uses an extremely large number of parameters [23].

Learning Algorithm

Learning is generally performed by minimizing a loss function which is dependent on the underlying task. Tasks based on classification use the softmax loss function or the sigmoid cross-entropy function, while those involving regression use the Euclidean error function [23].

Gradient-Based Optimization

Neural networks are generally trained using the backpropagation algorithm, which uses the chain rule to speed up the computation of the gradient for the gradient descent (GD) algorithm. However, for datasets with thousands (or more) of data points, using GD is impractical. In such cases, an approximation called the Stochastic Gradient Descent (SGD) is often used, where one computes gradients with respect to individual data points rather than the entire dataset. It has been found that training with SGD generalizes better than with GD [23].

Dropout

When training a network with a large number of parameters, an effective regularization mechanism is essential to combat overfitting. Dropout is a powerful regularization method that has been shown to improve generalization for large neural nets. In dropout, we randomly drop neurons with a probability p during training. As a result, only a random subset of neurons are trained in a single iteration of SGD. At test time, we use all neurons, however, we simply multiply the activation of each neuron with p to account for the scaling. Hinton et al. [23] showed that this procedure is equivalent to training a large ensemble of neural nets with shared parameters, and then using their geometric mean to obtain a single prediction [23].

1.6 CNNs usage

Deep learning techniques have been the focus of many studies in this field since 2012, when their use in artificial intelligence research began to dramatically increase. The PRI problem is no exception to this norm, and most recent successful PRI investigations and research have included these deep learning techniques in some way. The original idea for deep learning dates back many years. But due to the lack of strong hardware, it was not possible or difficult to implement. Recently, deep learning in artificial intelligence research has been widely used and welcomed. The main reasons for the deep learning boom in recent years are:

1. Significant increase in chip processing capability (such as the creation of GPUs)
2. Reduce the cost of computing hardware
3. Significant improvements in machine learning algorithms

Because deep networks can extract good features from input data, recent research has shown that they are successful in machine vision tasks. The widespread usage of deep algorithms by researchers goes back to the success of these algorithms in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition. This challenge is organized by ImageNet every year and in it, a large number of images are provided to the participants to introduce an algorithm for accurate classification of this large volume of images. And the algorithm which has the least error in classification and matching the images on the given database, is selected as the superior and winner algorithm. Since 2012, the best results of this challenge have been achieved by deep algorithms, and this has led to the widespread use of these algorithms.

The convolutional neural network (CNN) is one of the most important deep learning methods in which multiple layers are trained in new and robust ways. These networks are a subset of multi-layer perceptron (MLPs) and are designed for 2D data such as image and video [24]. The main difference between a convolutional neural network and a multi layer neural network is that in multi layer neural networks, there are a number of weights and biases in each layer to model the nonlinear behavior of the system so that the output of each layer is multiplied by a weight and is summed with bias. These weights and bias values will be adjusted as best as possible during the training phase to minimize error. But in convolutional neural network, in each layer, filters are applied to the image or output of the previous layers, and during the training phase, it is the filter coefficients that are optimized to minimize error.

There are several convolution layers in every CNN architecture. In these layers, the input image is convolved with filters whose weights are trainable. These filters are moved on the image. The depth of the filter is equal to the depth of the image. A feature page is created for each individual

Table 1.1: Comparison of convolutional algorithms in recent years

Network	Year & Rank	Architecture	Advantage
AlexNet	1^{st}in 2021	5convolutional layers + 3 fully connected layers	An important architecture that attracted the attention of researchers
Clarifai	1^{st}in 2013	5convolutional layers + 3 fully connected layers	Ability to view what happens inside the layers
SPP	3^{rd}in 2014	5convolutional layers + 3 fully connected layers	Remove the limit on the size of the input image
VGG	2^{nd}in 2014	5convolutional layers + 3 fully connected layers	Increase accuracy byincreasing network depth
GoogleNet	1^{st}in 2014	13-15convolutional layers + 3 fully connected layers	Increase depth and increase accuracy without increasing computations

filter. If we use n filters, n feature pages will be created. The weights of each filter, called w, are trainable and are constantly updated during network training. After the convolution operation, the result is summed with a bias number and stored in the feature page.

To this date, there are different architectures of CNN networks, available. Due to the importance of the diversity of architectures and their application in the field of PRI, a brief explanation of the types of architectures would be a great idea. In the 1990s, Yann Lecun and his colleagues [25] first used the convolution network to retrieve handwritten numbers. Their network was called LeNet-5. Their network consisted of seven layers and the input image to the network was 32×32. In recent years, ImageNet has made a collection of images available and is holding a challenge to test the algorithms developed by the researchers on this database, and the best algorithm is selected as the winner of the challenge each year. In 2012, the algorithm developed by Alex Krizhevsky and his colleagues [26] won the challenge. This algorithm, later known as the AlexNet-algorithm, turned all attention to deep learning in the field of computer vision. In this architecture, five layers of convolution and three fully connected layers are used. This architecture requires an input image with a fixed size of $3 \times 224 \times 224$. This architecture has two major drawbacks: one is that the size of the input image is fixed and the other is that there is still no understanding of why this architecture works very well. In 2013, a method was introduced by which the activities performed within the layers could be seen [27]. Thus, their proposed model, known as Clarifai, was able to win the award this year. But in all these algorithms, the size of the input image was fixed, and they had to change the dimensions of the image before entering the network.

K. He and his colleagues proposed a new integration strategy [28] called spatial pyramid integration, in which the size limit of the input image was removed, thus improving system performance. The VGGnet model was later introduced by karen Simonyan and his colleagues In 2014 [29]. In this model, by adding layers of convolution, they increased the depth of the network and improved the performance compared to previous methods. The GoogleNet architecture then topped the network in 2014 with a further increase in network depth [30]. The AlexNet architecture is a versatile architecture and the basis of most other architectures, but its depth is shallow, so the GoogLeNet architecture with 21 convolve layers performs better. If we do not want to change the size of the input images, we can use the spp architecture. Table 1.1 shows a summary of the types of architectures and their differences.

Chapter 2

State-of-the-art

2.1 CNN usage in PRI

The first method of PRI using deep learning is introduced in [4] according to the authors of the paper. In this study, convolutional neural networks have been used to automatically learn the appropriate features of images. Also, the similarity comparison stage is trained and regulated by the same networks. In previous studies, the feature extraction stage and the similarity comparison stage were performed simultaneously. Therefore, the presence of errors in any part of the system, caused a defect in the performance of the entire system. For example, if an error occurs in the similarity comparison stage, the useful features will no longer work. But in this research, all stages are trained and regulated separately. The input of this model is two frames, and the output determines whether these two frames are the same person or not. In other words, a binary classification is performed at the output. In order to evaluate this method, the authors have also used the CUHK03 database, which will be introduced in future sections as one of the important tools for evaluating PRI systems.

In another study in 2015, researchers proposed a method for simultaneously training the features and evaluation metrics in a system using convolutional networks [31]. In this method, similar to the previous method, the system, by receiving two input images, determines whether two people are the same or not, with the difference of defining a layer that calculates the local differences of reciprocal inputs. The architecture of this method is such that the input image pair first passes through two layers of convolution with maximum integration and is filtered by the filters of these two layers. Filtered output images are referred to as feature pages, so they include two-dimensional vectors that contain high-level features of input images. Then, in order to compare and find the differences between these feature pages related to the two input frames, a layer for calculating the local differences of the reciprocal inputs is placed, in which the difference between the values of the properties in Two different angles around the feature location are examined. The architecture of this method can be seen in Figure 2.1.

After the local differences of the reciprocal inputs are measured, these differences are summarized in the next layer. A high-level representation of these differences is then calculated, and finally the fully connected layers are placed in order to train the similarity metrics. Although this method performed better than the previous methods, but because the pair of images had to be applied separately to the system, in databases with a large number of gallery images, this evaluation takes time. Similar to this approach, in another study in 2017, the authors designed a network to re-identify the individual from the image and video sequences [32]. This network is called P2Snet and consists of three cost functions, three k-nearest neighbors (KNN-triplet), and a deep convolutional network that simultaneously extracts features from images and video, and determines the evaluation metric.

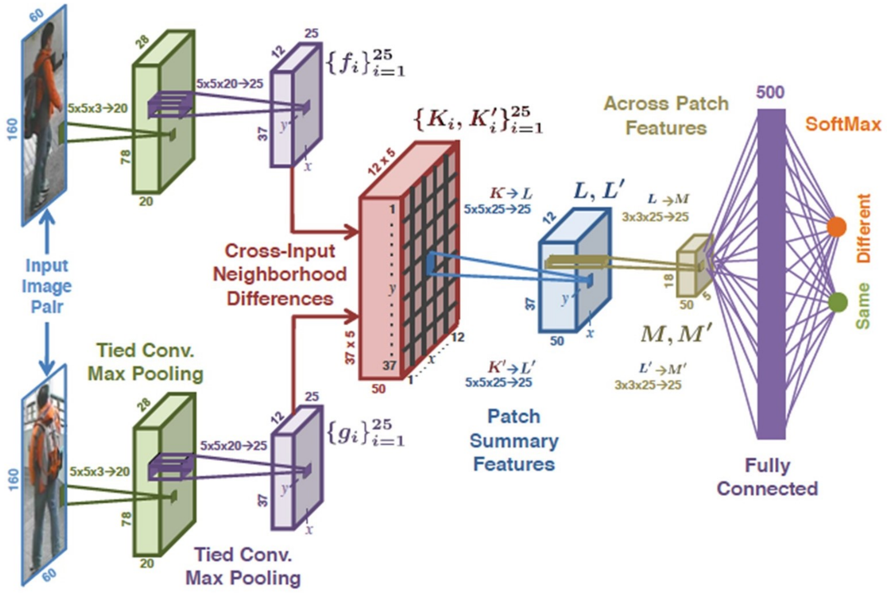


Figure 2.1: Simultaneous training of the model, reference evaluation metrics [4].

In 2018, a method was introduced in [5] in which, using Laplace structural diagrams and deep convolution neural networks, a system for the PRI was presented. This method extracts deeply distinguishing features from images of individuals using interpersonal cluttering and intrapersonal compression. The block diagram of this method can be seen in Figure 2.2. Although this method has achieved higher accuracy in identifying the individual than other previous methods, but its computational complexity is higher.

In another study conducted in 2019, the authors provided a comprehensive review of PRI methods using a variety of methods for extracting low-level features, deep features, and various similarity metrics training methods [22]. In this study, the authors have thoroughly evaluated and compared different combinations of 11 feature extraction methods and 22 similarity metrics training methods on 16 PRI databases. According to this study, the best performance overall is the IDE-ResNet method. This method uses ResNet deep convolution neural network to extract features from the last layers and compare these features, and has achieved high accuracy in PRI on most databases. After that, the IDE-VGGNet and IDE-CaffeNet methods have had the best performance. Among the methods other than deep learning, the best performance among the manual feature extraction methods is GOG feature.

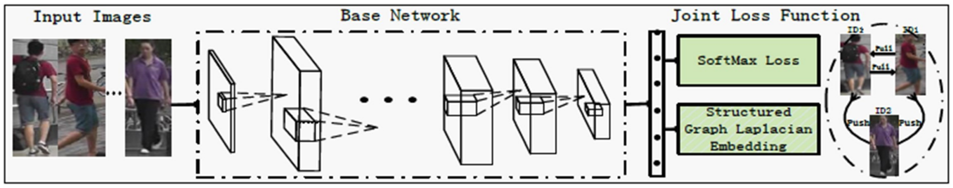


Figure 2.2: Method of using deep convolution neural network and Laplace structure diagram [5].

Based on our previous discussions and the findings in [22] which state that the best performance is related to IDE-ResNet and GOG methods, we have decided to expand these techniques and incorporate their benefits into our proposed technique.

2.2 GOG, LOMO, and other Related Work

In 2016, researchers presented a descriptor based on the hierarchical distribution of pixel properties for PRI [6]. The descriptor introduced in this method is in fact, an improved version of a hierarchical descriptor of covariance, which has been used very well in image classification applications. However, as mentioned in this reference, when using the hierarchical descriptor of covariance, the information about the mean of the pixels is deleted, while in identifying the person, the mean of the pixels plays an important role. The importance of the role of the mean of the pixels is well illustrated in Figure 2.3.

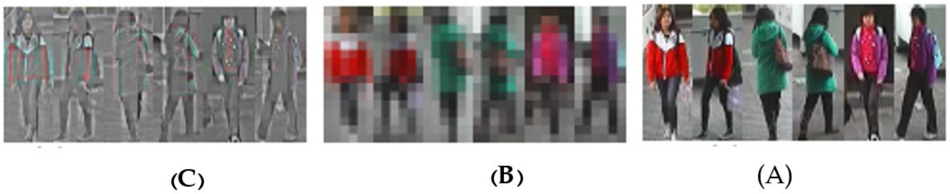


Figure 2.3: Shows the importance of the average pixels. (A) The original form. (B) the average of the pixels locally 10 * 10; and (C) the original image after deleting the mean. (It can be seen that it is much easier to re-identify people in the original image than in an image in which the average pixels are removed [6].)

It is challenging to recognise people when the mean of the pixels is removed, as demonstrated in Figure 2.3. Therefore, in the method described in the reference [6], a descriptor is used in order to extract the feature from different areas of the image, which uses a hierarchical Gaussian distribution. In this manner, the feature vector will incorporate both the pixel covariance and mean information. In this method, because the Gaussian distribution is used twice, the descriptor is called Gaussian of Gaussian (GOG), in which the color and texture properties of the images are extracted simultaneously. To do this, each image is first divided into g regions, which are horizontal bands. Then each division, squares with dimensions of $k \times k$ pixels are considered, each of which contains p pixels. Then for each pixel inside this square, a feature vector f_i of length 8 is defined as follows:

$$f_i = [y, m_{0}^{\circ}, m_{90}^{\circ}, m_{180}^{\circ}, m_{270}^{\circ}, R, G, B]^T \tag{2.1}$$

where *y* represents the location of the pixels in relation to the image's top, m is the magnitude of the pixel brightness gradient in the four directions 0 to 270 degrees, and R, G, and B are the pixel values in the red, green, and blue channels, respectively. After calculating this vector for each pixel, its values are normalized between 0 and 1. The Gaussian distribution including mean and covariance is now used to represent these feature vectors for each square. In the continuation of this method, the features of all the squares that are inside one area are transformed and represented again with another Gaussian distribution. Thus, for each region, we will ultimately have a Gaussian distribution, which is made up of the mean and covariance features of the pixels. Figure 2.4 summarises the procedures to use this technique.

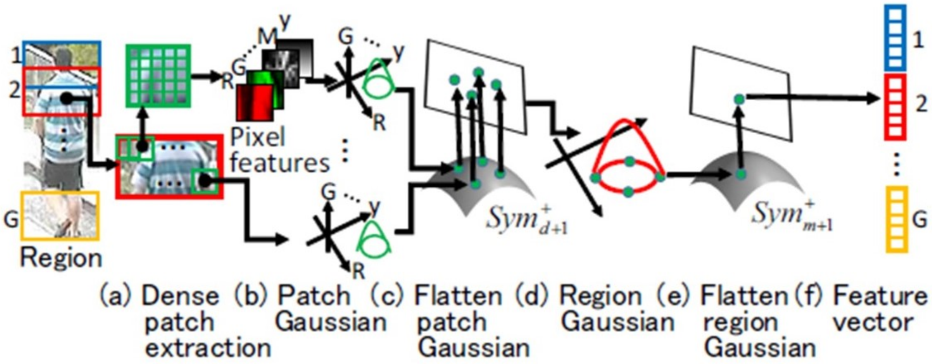


Figure 2.4: Steps of feature extraction from image in GOG method [6]

The size or amount of the brightness gradient of the pixels in various directions serves as an image texture feature in the feature vector represented by the GOG approach. In other word, it plays the role of image texture features. The method is therefore preferable to methods that employ only color features since it uses texture features in addition to color features and keeps average pixels in place.

Numerous researchers have already expressed interest in the subject of human re-identification, and their findings have enhanced feature extraction or metric measurement, which has improved the re-identification system. Additionally, most studies have focused on both metrics and features. In general, features can be taught and trained or handcrafted. Cai et al. in [33] studied color descriptors for PRI, suggesting Global Color Context (GCC) and calculating similarity between individuals and their neighbors as opposed to calculating distance between two individuals. Another study by Yang et al. in [34] used color names for PRI. This method takes advantage of salient colour names, representing RGB colours with a probability distribution , and uses this distribution to determine how similar two people are. In order to improve the re-id system's performance against viewpoint changes, Varior et al. [35] used a model to extract color patterns between individuals in different cameras. An unique signature made up of texture and colour features was proposed by Munaro et al. in [36]. These features were retrieved from skeletal joints using SIFT, SURF, and color values. Also, as explained before, hierarchical distributions of pixels were employed by Matsukava et al. in [6, 37] to offer a novel descriptor for human photos. They used both mean and covariance of pixels in a hierarchical Gaussian distribution and offered the Gaussian of Gaussian (GOG) descriptor for the PR problem.

Metric learning strategies have been the main subject of several articles as well. For instance, Kostinger et al. introduced an effective metric in [38] that learnt similarity based on equivalent or corresponding constraints. To determine if two photos were similar or dissimilar, they utilised a metric called ”KISSME.” By examining cross-view quadratic discriminant, Liao et al. in [7] presented a metric learning algorithm and gave it the name XQDA. They also presented the LOMO (Local Maximal Occurrence Representation) Feature Representation, a brand-new feature representation that is resistant to changes in lighting and angle of view. LOMO analyzes the horizontal occurrence of local features and maximizes the occurrence to make a stable representation against viewpoint changes.

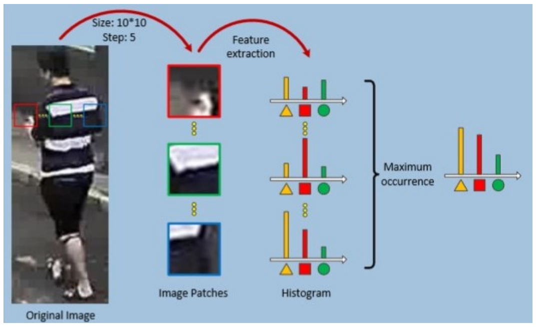


Figure 2.5: Illustration of the LOMO feature extraction method [7]

As a result of the successful development of deep learning applications in image processing, many researchers have been convinced to use this approach for carrying out image and video vision tasks [26, 39]. Deep learning has helped PRI as well, particularly in recent years. A popular subfield of machine learning, deep learning has experienced a strong growth curve in recent years [39].One of the most widely used deep learning techniques for computer vision problems, where filters in hierarchical layers learn to extract high-level characteristics from a lot of input data, is CNN. Following the popularity of AlexNet [26], other CNN architectures have been developed, like VGGNet [29] and ResNet [40], to name a few. In contrast to handcrafted features, the CNN has demonstrated a substantial ability for image and video classification, segmentation, and recognition; as a result, it is employed in the majority of contemporary Computer vision studies [41, 42, 43]. The first research on PRI that applied deep learning was done by Li et al. [4]. A convolutional neural networks (CNN)-based system that simultaneously learns features and metrics was introduced by them. By feeding pairs of pictures to their model, they can train it to determine whether or not each set of images represents one person. Additionally, Varior et al. [44] suggested a CNN-based network based on the Siamese network that learns features and metrics at the same time [45]. In addition to using a gate function to identify local similarities between pairs of images, they also employed a novel loss function for CNN. CNNs based on the Siamese network have also been used in numerous research investigations [31, 46]. For instance, training CNNs for classification tasks can extract superior representation for PRI, as shown by, Zheng et al. in [47]. They used ID-discriminative Embedding (IDE) to fine-tune CaffeNet on the re-id dataset and recovered features from prior CNN layers.

In conclusion, numerous studies have investigated PRI in relation to either feature representation or metric learning. Metric learning improves accuracy, but it takes time to train the metrics and prepare the pairs of images. This work aims to improve accuracy by proposing a new PRI system and to benefit from the advantages of both handcrafted and deep features, for which it integrates GOG and handcrafted features within an end-to-end process. In this study, we train a resilient CNN against changes in lighting and viewpoint.

Chapter 3

Methodology, Implementation and Results

3.1 Deep Learning-Transfer learning approach for PRI

Like each PRI system, the proposed model consists of two main steps: feature extraction and similarity measuring. Using a CNN and preparing a dictionary require an offline step by which the network's different layers learn to extract appropriate features. In this thesis, like IDE networks [47], each person is considered a different class to train a CNN. In this case, many individuals are available with few frames or images; hence we have a large number of classes with few samples in each class. Although many different CNNs have been successfully trained on different classification and recognition tasks, few samples of each class among the massive number of classes prevent CNN's correct training. So, we use the transfer learning technique to tackle this problem.

Here we introduce our proposed method based on Deep learning, CNN, and transfer learning for extracting the features representing pedestrians and we call it *deep features*.

As discussed before, PRI is greatly impacted by low-quality surveillance footage, crowded environments, lighting, viewpoint variation, occlusion, and feature representation. To address these challenges, our proposed method focuses on deep features extracted from CNN network. We also propose combining deep features extracted from our proposed models, and hand-crafted features which resulted in better accuracy and performance compared to only using the deep features extracted from our proposed models.

We use deep features in our work. For extracting deep features we propose ResNet-PRI and VGG-PRI models. Details are shown in Figures 3.1 and 3.2.

A CNN automatically extracts features, so hand-crafting features have become unnecessary for most applications. CNN learns what features to extract via backpropagation.

The approach that we use in this thesis is based on a deep learning paradigm that enables the creation of complex networks for solving the problem of feature extraction, usually tackled by means of CNNs, where deep layers in these complex networks act as a set of feature extractors that are often quite generic and, to some extent, independent of any specific classification task. This means that deep learning obtains a set of features learned directly from observations of the input images. The idea behind this approach is to discover multiple levels of representation so that higher-level features can represent the semantics of the data, which in turn can provide greater robustness to intra-class variability and other challenges that we face in PRI.

In the history of PRI methods, the method introduced in [4] in 2014 is the first time deep learning approach was used. In this study, convolutional neural networks have been used to automatically learn the appropriate features of images. Also, the similarity comparison stage is trained and regulated by the same networks.

Coming to our proposed models for extracting deep features, we tested many combinations of putting the layers together and the first structure that we propose is what we call ResNet-PRI, and is shown in Figures 3.1. As you can see in the figure the feature maps are extracted from the base model which in our case is ResNet50 [48]. We freezed all the layers of base model. So the extracted feature maps are connected to a GlobalAveragePooling2D layer, a Dropout(0.5) layer, a Dense layer with 'relu' activation function, and a Dense layer with 'softmax' activation. The number of output for the last layer should be defined by the number of classes per training and testing set. The last layer is just for monitoring the accuracy, loss and other parameters during the training phase and in practice we use the output of the layer before that as our feature vector. So far we have defined the structure of our ResNet-PRI model. Now, we compile it with the "categorical-crossentropy" loss function, and RMSprop optimizer with the learnnig rate of $(1e-3)$. The next step would be training the model. So, we fit the model with setting the batch size to 32, and setting epochs to 20. We tested our proposed model on CUHK01, CUHK03, and GRID datasets and the results for training is shown in figures 3.9, 3.10, and 3.11 respectively.

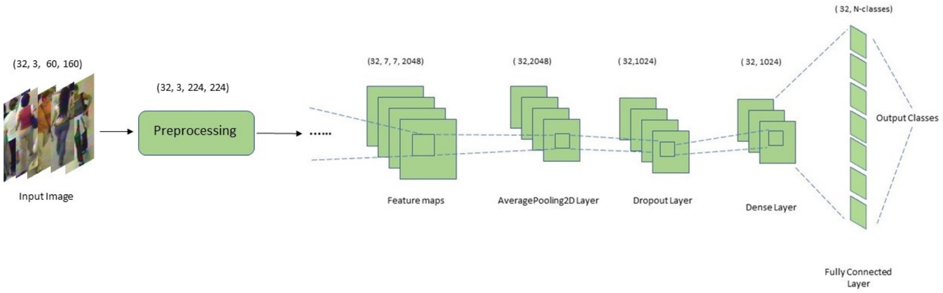


Figure 3.1: The proposed ResNet-PRI model schema

As shown in Figure 3.1, the training set's input photos are prepossessed in batches of 32. The image is then prepared and resized in accordance with the model's expectations. Then, from the basis model, feature maps having a size of $32*7*7*2048$ are extracted as there are 32 photos in each batch. We currently have 2048 feature maps each having size of $7*7$. Next, we move to the AveragePooling2D layer, which is the following layer. Thus, we calculate a value by averaging across $7*7$ maps. The output is 2048 values. Then, in the following layer, we remove half of the extracted values to make the model's classification task more challenging and ensure that the model can generalize and forecast more accurately. The collected features are then sent to a fully connected layer for classification, whose shape should be modified according to the number of classes present in the training set.

Coming to our second proposed model for extracting deep features, again we tested many combinations of putting the layers together and the best structure that we propose and we call it VGG-PRI is shown in figure 3.2. As you can see in the figure the feature maps are extracted from the base model which in our case is VGG16 [49]. We freezed all the layers of base model and did not trained them. So the extracted feature maps are connected to a MaxPooling2D layer, a GlobalAveragePooling2D layer, a Dense layer with 'relu' activation function, and a Dense layer with 'softmax' activation. The number of output for the last layer should be defined by the number of classes per training and testing set. The last layer is just for monitoring the accuracy, loss and other parameters during the training phase and in practice we use the output of the layer before that as our feature vector. So far we have defined the structure of our VGG-PRI model. Now, we compile it with the ”categorical-crossentropy” loss function, and RMSprop optimizer with the learnnig rate of $(1e-3)$. The next step would be training the model. So, we fit the model with setting the batch size to 32, and setting epochs to 35. We tested our proposed model on CUHK01, CUHK03, and GRID datasets and the results for training is shown in Figures 3.9, 3.10, and 3.11 respectively.

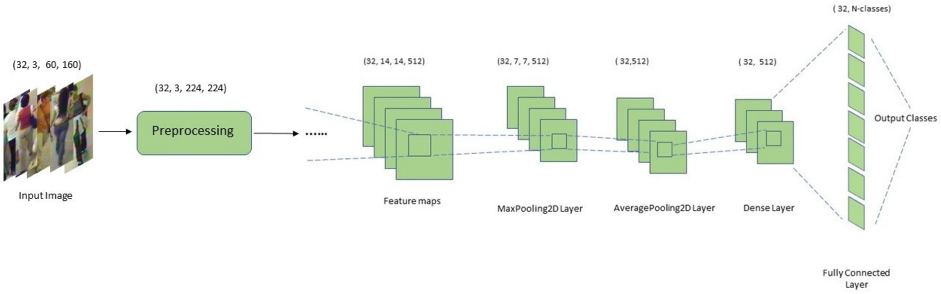


Figure 3.2: The proposed VGG-PRI model schema

So far, we have proposed our models for extracting deep features.

3.2 An end-to-end method for PRI based on Deep features, Hand-Crafted features, and Cosine similarity criterion

As mentioned in the related work part, in previous studies, the feature extraction stage and the similarity comparison stage are performed simultaneously [4]. Deep learning and convolutional neural networks were used to automatically learn the appropriate features of images, but the same networks were utilised to train and regulate the similarity comparison step. Therefore, the presence of errors in any part of the system caused a defect in the performance of the entire system. For example, if an error occurs in the similarity comparison stage, the useful features will no longer work. But in this research, all stages are trained and regulated separately. The input of our models described in previous section, is two images, and the output would be the feature vectors. Then the system determines whether these two frames are the same person or not. In other words, a binary classification is performed at the output. Deep features are used to and represent each person's features.

The feature extraction process includes extraction of deep features from the last Fully Connected (FC) layers of our proposed models shown in Figures 3.1 and 3.2, providing the in-depth features of each person. The main pipeline for this method is shown in Figure 3.3

Going deeper on CNN, the dimension of features is reducing. Also, deeper layers indicate more complex details of input image. Therefore, the last FC layer is considered for extracting features. In the architectures used here, the extracted deep features will be a vector of size 1×1024.

Also, in this thesis, we show that the combination of deep features (the features extracted by our proposed model), and so called “hand crafted” features lead to higher accuracy of the PRI system. So, for better accuracy we concatenate both these features in the final feature vector and by doing

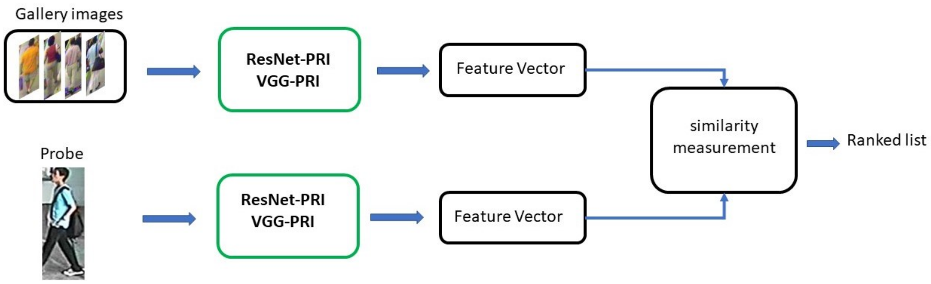


Figure 3.3: Main pipeline of the proposed method.

so we extend the advantages of our approach by adding the upsides that the either methods can offer. As shown in Table 3.1 we can see that GOG-Lab is doing better between the hand-crafted features and also ResNet-PRI has a sightly better performance compared to VGG-PRI; so for the combination of both features we use the combination of GOG-Lab and ResNet-PRI features and the results are shown in table 3.1.

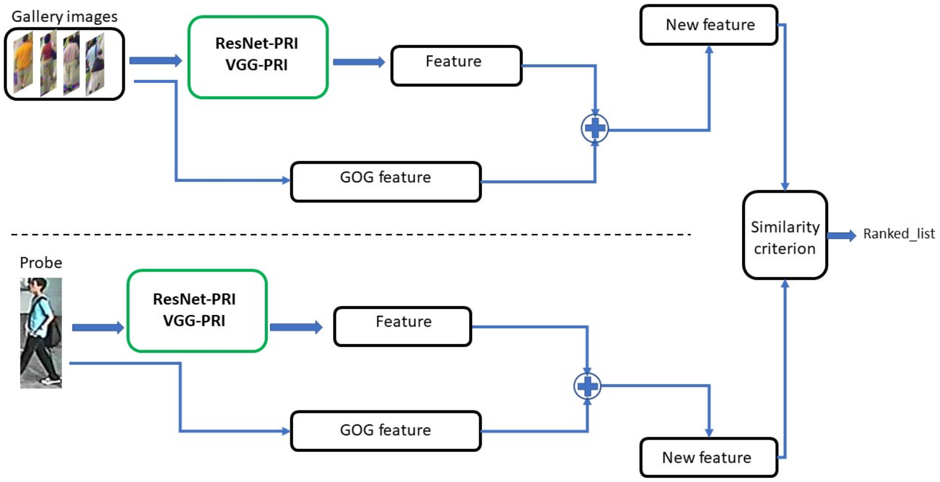


Figure 3.4: Main pipeline of the proposed method.

The GOG descriptor of each person is calculated in order to obtain the handcrafted features, and the final feature vector is formed by fusing deep and GOG features and forming a new feature vector.

As mentioned in [6], GOG features are computed using two hierarchical Gaussian distributions. To this end, each image is divided by N horizontal stripes, each of which contains region patches of $k \times k$ pixels. For each pixel of a patch, a feature vector is defined which contains pixel values, pixel location and intensity gradient in four directions. Using these features, each patch is modeled with a Gaussian distribution. So, we have some patch of image each of which has specific average and covariance. Again all patches are modeled with another Gaussian distribution. By concatenating all averages and covariances, each region will be represented by a mean vector and covarience matrix.

For an image of a person, there are G regions of patches. So we will have G mean vectors and space covariance matrices which concatenating them is considered as GOG feature vector of image. the dimension of this feature vector depends on hyper-parameters, G and K.

Therefore, the dictionary matrix is constructed by computing all training image feature vectors. This dictionary will be used in the online step described below. The schematic of the proposed method is given in Fig. 3.4.

The dictionary for similarity representation is constructed using gallery images. Each column of dictionary indicates feature vector of one person which is constructed from deep and GOG features.

As soon as the query frame is given, it will be resized to 224×224 to fit the trained and proposed CNN. After applying the feedforward neural network, the feature vector is extracted from the last FC layer and the GOG image features are extracted as well. By concatenating these features, the conclusive feature vector of the probe person is computed as G which stands for gallery images.

Instead of using the time-consuming similarity learning, we can define the PRI problem as a similarity problem. By considering the conclusive feature vector for probe person as P, we will compute the Cosine Similarity of G and P.

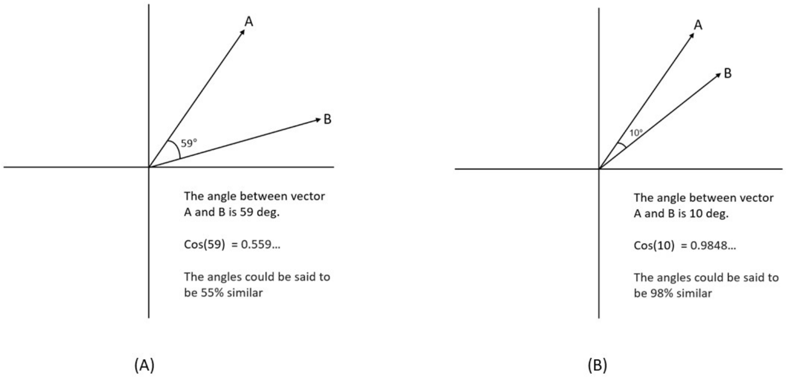
The more relevant person in the gallery yields a large corresponding number in the output *similarity* output vector. Thus, to employ a ranking strategy, the output vector is sorted from high to low. Thus, indices of the relevant individuals are found.

3.2.1 Understanding Cosine Similarity And Its Application

A similarity measurement called Cosine Similarity is used for this project. The main advantages of using this approach are the simplicity of this implementation, and more importantly, it is calculated in a separate step. Cosine Similarity is a calculation that expresses how similar two or more vectors are. The cosine similarity is described mathematically as below:

$$similarity = cos(\Theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.1}$$

Cosine Similarity is a value that is bound by a constrained range of 0 and 1. Suppose the angle between the two vectors is 90 degrees. In that case, the cosine similarity will have a value of 0; this means that the two vectors are orthogonal or perpendicular to each other. As the cosine similarity measurement gets closer to 1, then the angle between the two vectors A and B is smaller. The images below depict this more clearly.



The angle between vector A and B is 59 deg.

Cos(59) = 0.559...

The angles could be said to be 55% similar

The angle between vector A and B is 10 deg.

Cos(10) = 0.9848...

The angles could be said to be 98% similar

Figure 3.5: (A): Two vectors with 98% similarity based on the cosine of the angle between the vectors. (B): Two vectors with 55% similarity based on the cosine of the angle between the vectors

In nutshell, Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.

3.3 Experimental Results

The proposed method is assessed using three datasets, and renown metrics are used to compare the results with the most recent approaches. The datasets, evaluation measures, and processes are detailed in the lines that follow before the numerical findings are shown.

3.3.1 Datasets

For the evaluation, three datasets are used:

CUHK01

The first one, CUHK01 [50], features 971 pedestrians in an outdoor setting. Four frames from two separate cameras with various viewpoints are provided for each individual in the database. Thus, a total of 3884 photos are accessible. 971 individuals are included in this dataset from two separate camera angles.

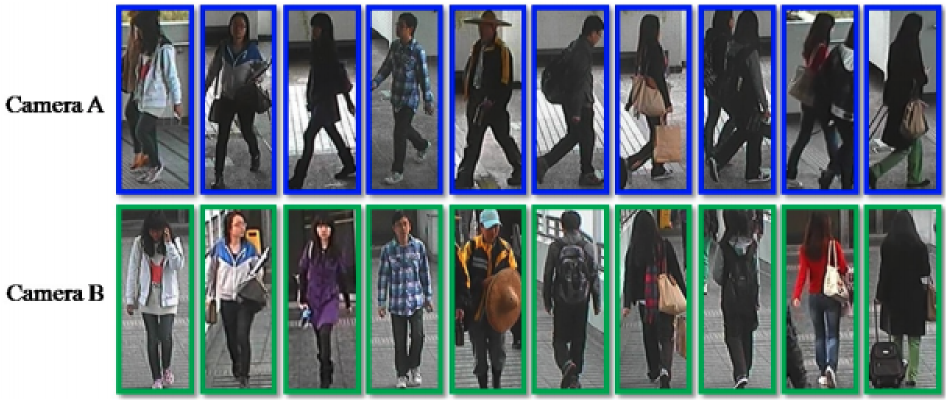


Figure 3.6: Sample frames of PRI CUHK01 dataset

CUHK03

Since the deep models require a substantial quantity of data to be trained, a large database of 13164 images known as CUHK03 [4] is produced. This database contains 1360 pedestrians who are photographed by six outside cameras with two distinct (non-overlapped) viewpoints.

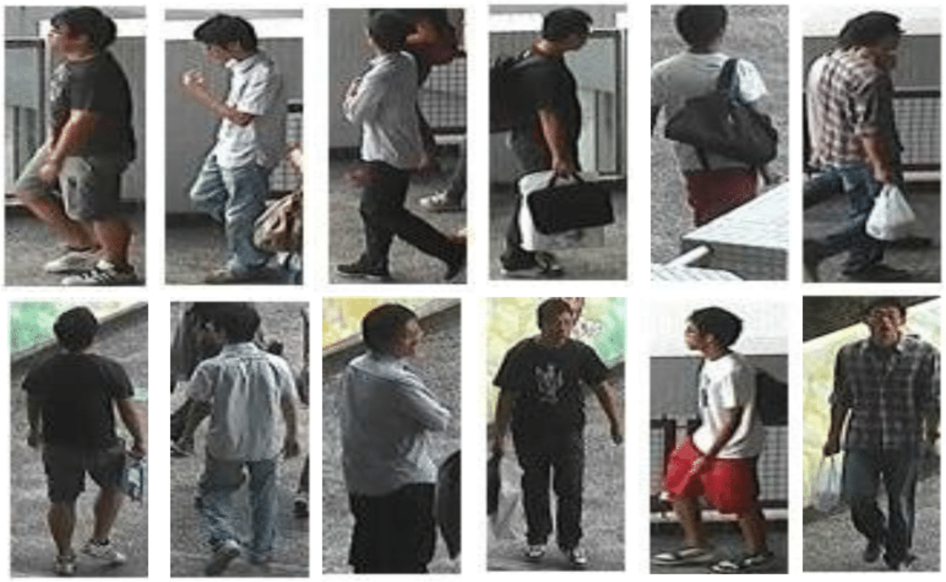


Figure 3.7: Sample frames of PRI CUHK03 dataset.

GRID

Another challenging database that is collected in a busy underground medium is GRID [51]. There are 250 pedestrians in total, each carrying two frames.The photos in this collection are taken by eight separate cameras in a variety of lighting conditions with cluttered backgrounds, low resolutions, and occlusion. A few representative frames from these datasets are shown in Fig. 3.8.

There are 250 pedestrian image pairs in the underGround Re-IDentification (GRID) dataset. Each pair includes two pictures of the same person taken from various angles by the camera. All photographs are taken using 8 separate camera viewpoints that are installed in a crowded subway stop. Eight disjoint cameras captured the images in this dataset under different illuminations, low resolutions, and crowded backgrounds with the presence of occlusion.

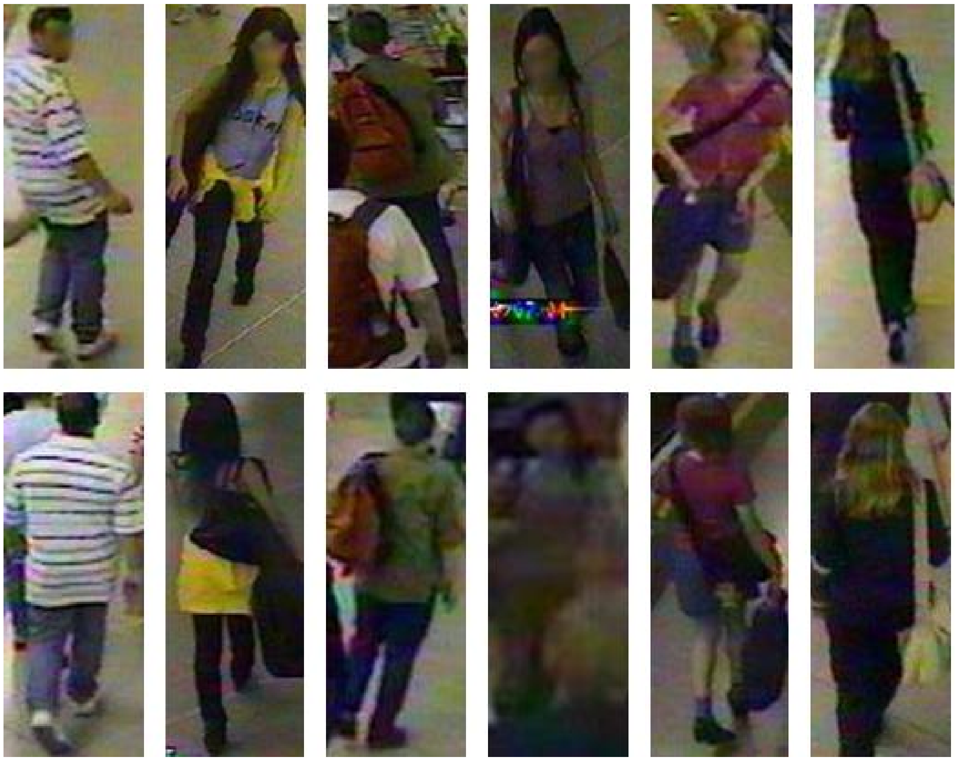


Figure 3.8: Sample frames of GRID dataset.

3.3.2 Evaluation metrics

Because the suggested approach is defined in accordance with a ranking protocol, the initial metric for assessment is Rank-1. This statistic seeks to determine the proportion of individuals in the probe sets who are accurately referred and matched to their gallery pictures. So, if the first re-identified person is successfully identified after the creation of a ranked list for a probe person, rank-1 is equivalent to 1 for that particular person. The average rank-1 of all test images is the rank-1 for a dataset. The chance of a relevant individual appearing in the k-th position of a ranked list is indicated by Rank-k, another metric that has been utilised in the literature. The Cumulative Match Characteristics curve (CMC) is a different popular statistic for assessing overall performance. The CMC is determined using [52]:

$$CMC(r) = \frac{1}{N} \sum_{k=1}^{N} E(P_k \leq r), \tag{3.2}$$

$$E(P_k \leq r) = \begin{cases} 1 & P_k \leq r \\ 0 & otherwise \end{cases}, \tag{3.3}$$

where P_k indicates the position of the $k-th$ probe person in the ranked list, and N denotes the total number of probe persons in the database. A widely utilised statistic in PRI research studies is the CMC [22, 51]. The re-identification accuracy of the overall top k matches is reported by this curve. Better performance is represented by the CMC curve. Because it shows how models perform across all ranks rather than just one in particular.

3.4 Implementation

Due to the high quantity of data that our model requires in case we want to build a model and train this model from scratch, transfer learning is used to keep the CNN model from overfitting. In addition, cross validation technique is used to account the loss and accuracy of our model in each step of the training. The results are reported in Figures, 3.9 , 3.10, and 3.11.

In the training phase, we extend VGGNet [29] and ResNet50 [40], two popular convolutional neural networks that, according to an analysis of the literature, have produced better results [22] as the base model and constructed a new model structure shown in Figures 3.1 and 3.2.

The subjects used for system training will not be used for tests, and in the test, every subject is brand-new to the network because each dataset is split into two non-overlapping subsets for training and testing. We used Tensorflow and Keras for training. Fig. 3.9 shows the accuracy and loss function during VGG-PRI and ResNet-PRI's training on CUHK01 PRI dataset, indicating that no substantial overfitting occurred during training. Also, Figure 3.10 and Figure 3.11 represent the same info for the CUUHK03 and GRID datasets, respectively. And again confirm that no substantial overfitting occurred during training. Besides, Fig. 3.12 shows trained CNNs visualization for some layers of ResNet-PRI. IT can be observed, that captured features by our proposed network share a lot of similarities and can extract more specific features of the person in deeper layers. This demonstrates how effective and (semantically) accurate the proposed strategy is.

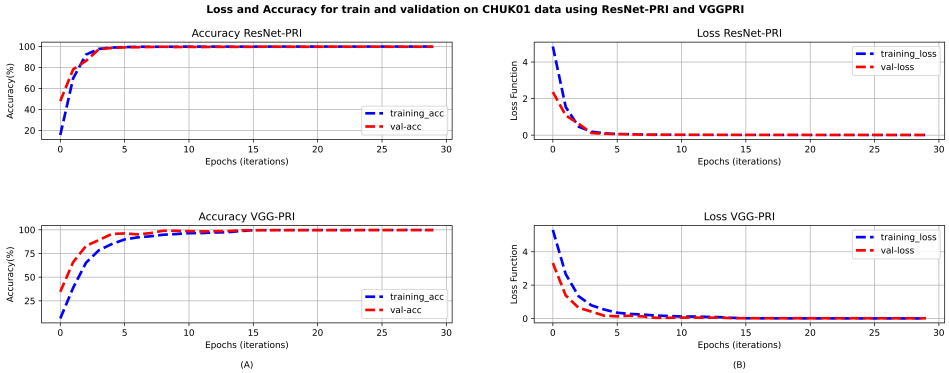


Figure 3.9: The accuracy (A) and loss function (B) of training ResNet-PRI and VGG-PRI on re-id database, CUHK01. The performance on validation data is following the training data in both accuracy and loss function.

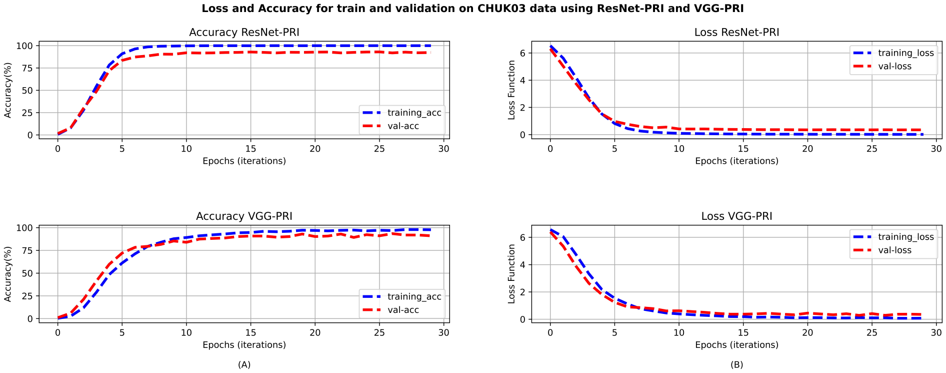


Figure 3.10: The accuracy (A) and loss function (B) of training ResNet-PRI and VGG-PRI on re-id database, CUHK03. The performance on validation data is following the training data in both accuracy and loss function.

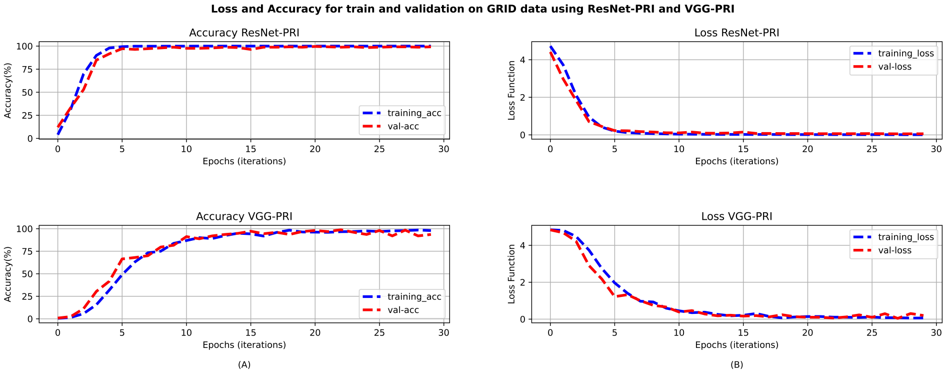


Figure 3.11: The accuracy (A) and loss function (B) of training ResNet-PRI and VGG-PRI on re-id database, GRID. The performance on validation data is following the training data in both accuracy and loss function.

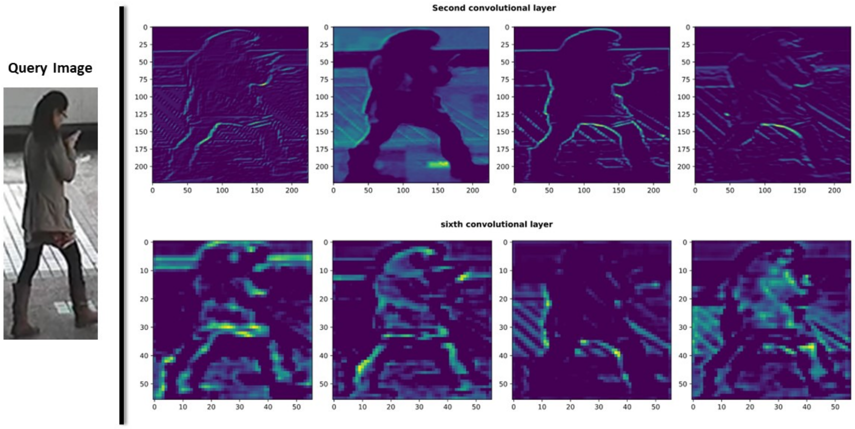


Figure 3.12: Visualization of some filters of ResNet-PRI in the second convolutional and fifth convolutional layers. It can be seen that deeper layers learn to extract more specific features of the person.

3.5 Results

Five high performing handcrafted feature representations are taken into consideration For evaluation: GOG with RGB,RnB, LAB, HSV color spaces[37], and LOMO [52]. ResNet-PRI and VGGNet-PRI are utilized for extracting deep features. Among the handcrafted features in the CUHK01 database, GOG-Lab [37] performed better in Rank-1= 0.52, and between deep networks, ResNet-PRI performed better by achieving rank-1= 0.944. Thus, the combination of GOG-Lab and ResNet-PRI is considered for creating the new feature vector. The proposed method of combining deep features with GoG features for PRI system is compared with the state-of-the-art methods in Tables 3.2, 3.3, and 3.4.

Fig. 3.13 illustrates the CMC curves of the implemented methods on CUHK01. It is shown that for this dataset, Deep features outperform handcrafted features in a substantial way. Additionally, the CMC curve for the proposed approaches are higher than other curves.

The GOG has been the best handcrafted feature for image retrieval, according to the evaluation of various feature extraction approaches in [22].

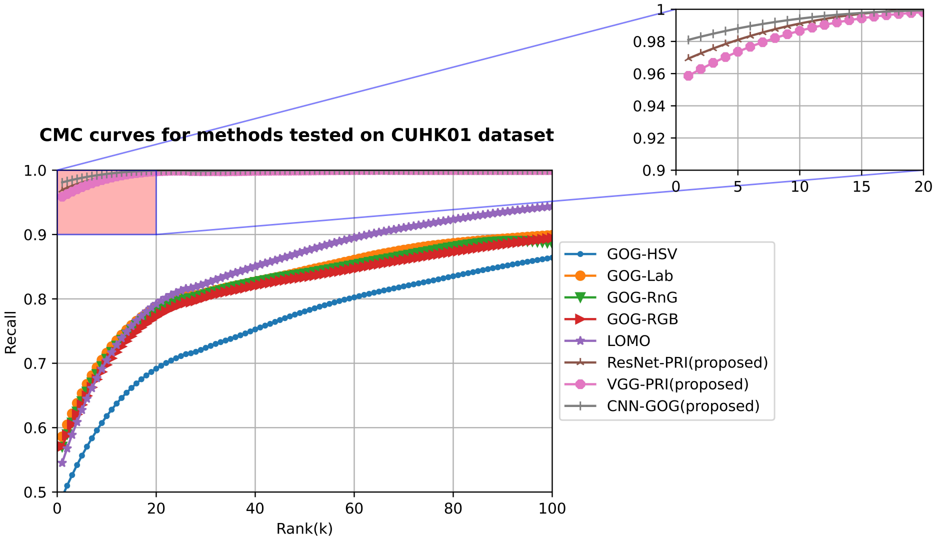


Figure 3.13: The CMC curves of the implemented methods on the CUHK01 database. The top left corner of the CMC diagram is magnified at the right side of the figure.

Fig. 3.14 illustrates the CMC curves for the CUHK03 database. Since CUHK03 has more frames for each person, the overall performance of methods is better. CMC of the proposed method in Rank-5 is equal to 0.99.

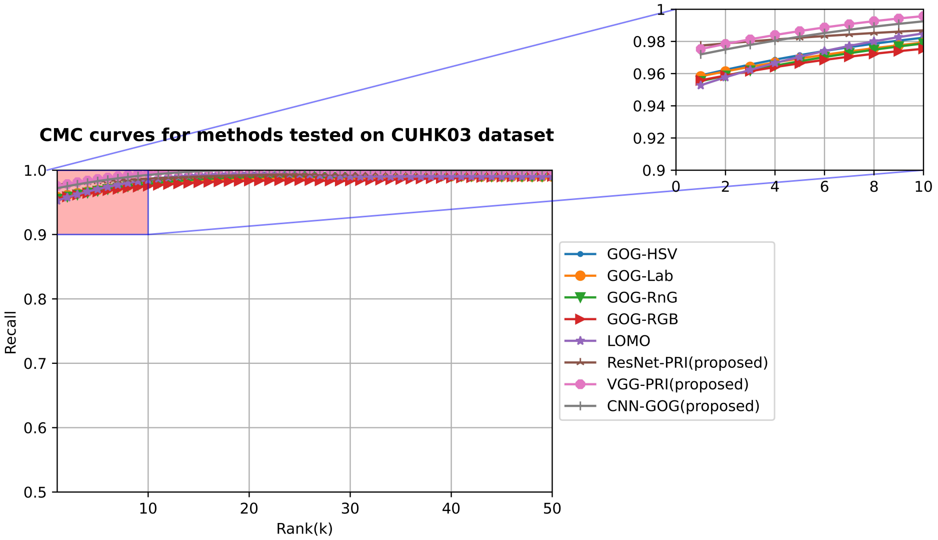


Figure 3.14: The CMC curves of the implemented methods on the CUHK03 database. The top left corner of the CMC diagram is magnified at the right side of the figure.

Fig. 3.15 shows the output evaluations on the GRID database. The precision of approaches is far from ideal because of the low-quality frames and significant number of occlusions of person's images in this dataset. In contrast to other strategies that have been used, our proposed methods get the best rank-1 on the GRID database with a higher CMC. For the ResNet-PRI suggested method on the GRID, rank-1 is equal to 0.32. All of the numerical findings across all databases are shown in Table 3.1. For better comparison, it includes both the proposed methods and some of the most recent ones.

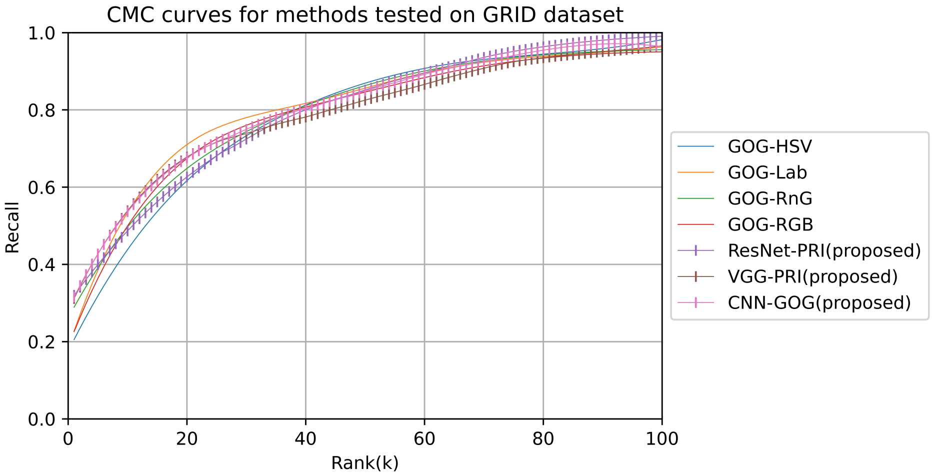


Figure 3.15: The CMC curves of the implemented methods on the GRID database.

Outputs in Tables 3.1 show the rank-1, rank-5, and rank-20 of our proposed methods and the state-of-the-art methods on the three datasets. We call our proposed model that combines *deep* features with *hand-crafted* features) as CNN-GOG.

For CUHK01, among the different techniques, the ResNet-PRI and GOG handcrafted features combination (ResNet-PRI+GOG-Lab) called CNN-GOG produced a rank-1 accuracy of 95.9%. The proposed method showed better performance than all the methods in the CUHK01 dataset with 95.9% in rank-1 and 99.8% in rank-20.

Regarding the CUHK03 dataset, our 3 proposed methods achieved 96%, 99%, and 100% accuracy in rank-1, 5, and 20, respectively. See table 3.1

For the single-shot, low-resolution, and highly occluded GRID dataset, most effective was CNN-GOG and ResNet-PRI, reaching 31.5% and 32% in rank-1, respectively. The CNN-GOG proposed strategy almost outperformed other methods in this dataset with 31.5%, 45% in rank-1, 5 respectively and achieved 68% in rank-20 which is comparable to the closest methods, GOG-lab, with 71% in rank-20. Be aware that the suggested approach does not employ any metric learning techniques in its effort to re-identify subjects. The results of the evaluations suggest that the proposed method performed better than state-of-the-art methods when more than one frame is available for each sample , i.e., in the multi-shot tasks.

Since the CNN-GOG is performing better between the proposed methods in this thesis, the performance of the CNN-GOG strategy is compared to the state-of-the-art models in Tables 3.2, 3.3, and 3.4 .

Table 3.2 shows rank-1, rank-5, rank-10, and rank-20 for the CUHK01 database. However, compared to other techniques, Parsing and Saliency Detection (PSD) [53] achieved 83.2% in rank-1, the proposed method outperforms it by 14.7% in rank-1. On the other hand, the Eliminating Background-bias (EBb) [54] works better than the PSD in all the other ranks. However, the proposed

Table 3.1: Numerical results and comparison between the investigated methods and other state of the art methods on the three datasets.

Dataset	CUHK01			CUHK03			GRID		
RANK **Method**	1	5	20	1	5	20	1	5	20
GOG-RGB	52	65.6	76.3	93	98	98	22.5	36.3	67.4
GOG-RnG	50.5	66.2	77.5	91	98	99	28.9	39.1	64.8
GOG-Lab	52.2	68.5	77.9	94	98	99	22.7	38.5	71
GOG-HSV	44.1	57.7	68.9	93	98	99	20.5	31.8	61
LOMO	48.9	65.2	78.8	92	98	99	-	-	-
ResNet-PRI (proposed)	94.4	98.9	99.8	96	99	99	32	40	63
VGG-PRI (proposed)	94.2	97.7	99.4	96	99	100	31	42.6	67
CNN-GOG (proposed)	**95.9**	**99.6**	**99.8**	**96**	**99**	**100**	**31.5**	**45**	**68**

method surpasses EBb in all ranks. Achieving 95.9% in rank-1 indicates that our proposed method outperforms the state-of-the-art models and performs well on the CUHK01 dataset. Also, the rank-20 in the proposed method is 99.8% and surpassed PSD by 1%.

Table 3.3 depicts the indexing results in the CUHK03 dataset. Amongst the other methods, EBb performs better in rank-1 and rank-5, achieving 92.5% and 98.4%, respectively. In rank-10 and rank-20, PSD outperforms others, achieving 99.1% and 99.6%. The proposed method surpasses all by 3.5%, 0.6%, and 0.4% in rank-1, rank-5, and rank-20, respectively.

Table 3.4 presents comparison outcomes on the GRID dataset. Although Supervised Smoothed Manifold (SSM) [55] performs better than the relevant models by achieving 26.5% in rank-1, the proposed CNN-GOG model is the best and achieves 31.5% in rank-1. It can be observed from Table 3.4 that a fusion of GOG features and XQDA metric learning, called GOG-fusion+XQDA [37], achieves 24.7% in rank-1, while our model surpasses it without using metric learning algorithms. Utilizing a hybrid of handcrafted features and deep learning, GOG+WHOS+FTCNN [56] achieves 24.5% in rank-1, however the proposed strategy outperforms in Rank-1 and is comparable in other ranks.

Table 3.2: Comparison results on the CUHK01 dataset

Rank **Method**	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
SGL-DGDnet [5]	70.9	89.8	93.5	95.9
FCDF [57]	48.1	-	81.7	92.5
PSD [53]	83.2	-	97.1	98.8
EBb [54]	82.5	96.1	98.2	99
SpindleNet[58]	79.9	94.4	97.1	98.6
HGD-ResNet[37]	80.3	-	97	98.7
CNN-GOG (proposed)	**95.9**	**99.6**	**99.8**	**99.8**

Table 3.3: Comparison results on the CUHK03 dataset

Rank Method	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
SGL-DGDnet [5]	84.7	97.4	98.9	99.5
DeepList [59]	50.7	80.5	89.3	93.4
PSD [53]	91.8	-	99.1	99.6
EBb [54]	92.5	98.4	98.9	99.5
SpindleNet [58]	88.5	97.8	98.6	99.2
HGD-ResNet [37]	88.5	-	98.4	99.2
CNN-GOG (proposed)	**96**	**99**	**99**	**100**

Table 3.4: Comparison results on the GRID dataset

Rank Method	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
GOG-fusion+XQDA[37]	24.7	47	58.4	69
SCSP [60]	24.6	44.3	55.2	65.8
DR-KISSMe [61]	20.6	39.3	51.4	62.2
GOG+WHOS+FTCNN [56]	24.5	47.3	57	68.2
SSM [55]	26.5	46.3	56.1	66.8
CNN-GOG (proposed)	**31.5**	**45**	**55.4**	**68**

Fig. 3.16 shows the visually ranked lists of certain queries in which the left column displays the query samples, and the ranked list of the persons who are re-identified, on the right side using the proposed method. As can be observed, the re-identified person in the first rank for the first query sample is accurate and matched to the inquiry. The other frames re-identified for the initial query share a lot of similarities: They all wear coat and are dressed in the same color. Concentrating on the example query in the last row, we can see that all re-identified individuals were carrying bags. This demonstrates how effective and (semantically) accurate the proposed strategy is.

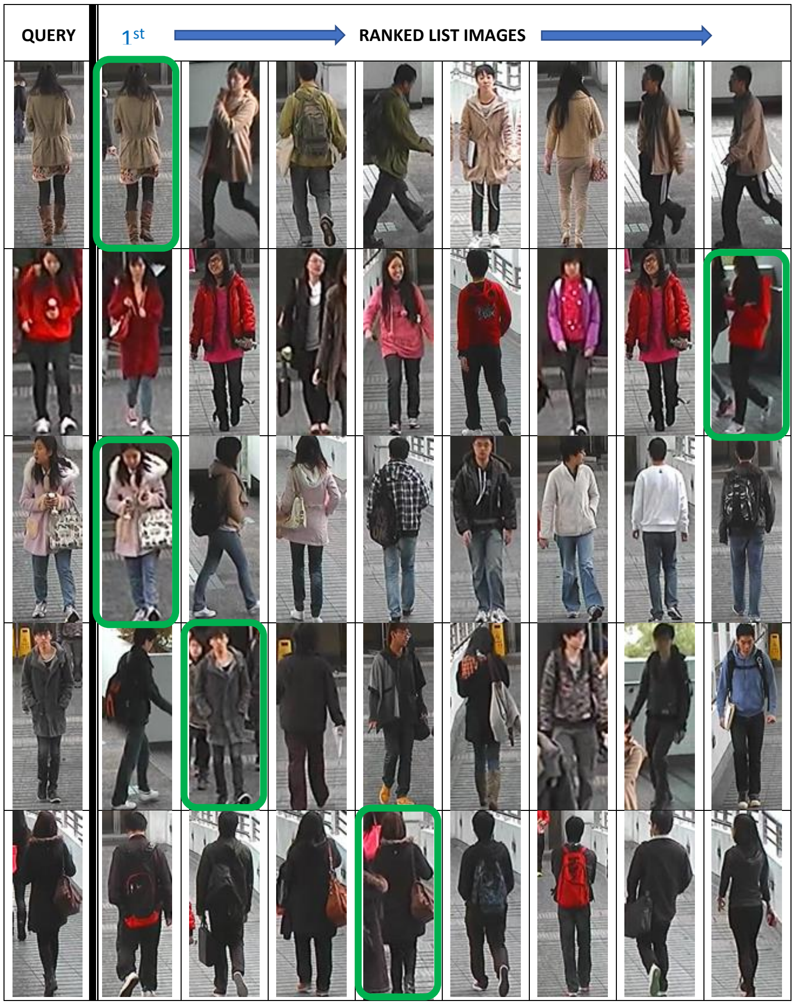


Figure 3.16: Visual results of some sample queries (left column) and their re-identified ranked list (sorted from left to right) using the proposed method.

Chapter 4

Conclusion and future works

Person tracking is one of the most important tasks in computer vision. It has a multitude of real-life applications, including use cases such as traffic monitoring, robotics, medical imaging, autonomous vehicle tracking, and more.

We realized that either we want to track the pedestrians in video feeds and extract their trajectories or we want to track people within non-overlapping cameras; first we should detect our target and then by focusing on the detected person, we will be able to track that person by assigning a unique ID to her. We defined models and methods that focus on the visual appearances of the pedestrians. In this thesis, we addressed the problem of using PRI algorithm for pedestrian trajectory extraction. We proposed a re-identification approach that extracts deep features. PRI is greatly impacted by low-quality surveillance footage, crowded environments, lighting, viewpoint variation, and occlusion. Also, for trajectory extraction we need to have a model that is robust toward the aforementioned challenges. As a result, the proposed method focuses on deep features extracted from CNN network. In addition, we showed by combining our deep features extracted from our proposed models and hand-crafted features, we get a higher accuracy. The study compared several effective descriptors and their combinations, and found the combination of GOG-Lab and ResNet-PRI (CNN-GOG) to be the most effective one. The PRI problem was transformed into a similarity model instead of utilizing metric learning, and concatenated features were used to generate a new feature vector. When the similarity verification step is solved, people in various cameras, or within different frames could be identified. The proposed method is tested on three datasets, and experimental findings show that our method outperforms other new methods in the majority of ranks.

In order to aid in the training of feature extraction networks, particularly for one-shot databases, we will test generative networks to create additional samples of a human image. Additionally, a few techniques can be used to select the best subset of features from a large collection of handcrafted features. In order to improve the accuracy of finding the best matches, further approaches of similarity measurement will be investigated. Also a weighted combination of the proposed method with different motion-based tracking algorithms will be tested to make sure that we will be able to take advantage of both approaches.

Bibliography

[1] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. https://www.slideshare.net/NaverEngineering/learning-disentangled-representation-for-robust-person-reidentification, 2020 (accessed October 12, 2022).

[2] Gray et al. Viper (viewpoint invariant pedestrian recognition). https://paperswithcode.com/dataset/viper, (accessed October 19, 2022).

[3] eds Hemanth DJ, Estrela VV. Deep neural networks for image classification. In *Deep learning for image processing applications*, pages 27–50. IOS Press, Amsterdam, Netherlands, 2017.

[4] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159. IEEE, 2014.

[5] D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann, and N. Zheng. Deep feature learning via structured graph laplacian embedding for person re-identification. *Pattern Recognition*, 82:94–104, 2018.

[6] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372. IEEE, 2016.

[7] H. LiaoShengcai, Zh. Xiangyu, et al. Person re identification by local maximal occurrence representation and metric learning. In *Proc of the 33rd IEEE Conf on ComputerVisionandPatternRecognition. Piscataway, NJ: IEEE*, pages 2197–2206. IEEE, 2015.

[8] P.L. Mazzeo, S. Ramakrishnan, and P. Spagnolo. Visual object tracking with deep neural networks. IntechOpen, 2019.

[9] Darko Musicki Subhash Challa, Mark R. Morelande and Robin J. Evans. An introduction to deep convolutional neural nets for computer vision. In *Fundamentals of Object Tracking*, pages 1–370. Cambridge University Press, Cambridge, UK, 2011.

[10] Dawei Zhang, Zhonglong Zheng, Tianxiang Wang, and Yiran He. Hrom: Learning high-resolution representation and object-aware masks for visual object tracking. *Sensors*, 20(17), 2020.

[11] Forrest Anderson. Real time, video image centroid tracker. In *Acquisition, Tracking, and Pointing IV*, volume 1304, page 82. SPIE, 1990.

[12] Hitesh A Patel and Darshak G Thakore. Moving object tracking using kalman filter. *IJCSMC Journa*, 02(4):326–332, 2013.

[13] Greg Welch and Gary Bishop. An introduction to the kalman filter. 1997.

[14] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. 2017.

[15] B. Hadjkacem, W. Ayedi, M. Abid, and H. Snoussi. Multi-shot human re-identification using a fast multi-scale video covariance descriptor. *Engineering Applications of Artificial Intelligence*, 65(2):60–67, 2017.

[16] T. Wang, Sh. Gong, X. Zhu, and Sh. Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016.

[17] C. C. Loy S. Gong, M. Cristani and T. M. Hospedales. The re-identification challenges. In G. Shaogang, C. Marco, and L. Shuicheng, Y.and Chen Change, editors, *Person Re-Identification*, pages 1–20. Springer, London, 2014.

[18] acopo Masi, Giuseppe Lisanti, Federico Bartoli, and Alberto Del Bimbo. Person re-identification: Theory and best practice. http://www.micc.unifi.it/reid-tutorial/slides/, (accessed October 19, 2022).

[19] Sh. Wang, Sh. Wu, L. Duan, Ch. Yu, Y. Sun, and J. Dong. Person re-identification with deep features and transfer learning. In *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, volume 42, pages 704–707. IEEE, 2017.

[20] X. Yang, M. Wang, and D. Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017.

[21] J. Wang, Zh. Wang, Ch. Liang, Ch. Gao, and N. Sang. Equidistance constrained metric learning for person re-identification. *Pattern Recognition*, 74:38–51, 2018.

[22] M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):523–536, 2018.

[23] Greenspan Hayit Zhou, Kevin and Dinggang Shen. An introduction to deep convolutional neural nets for computer vision. In *Deep learning for medical image analysis*, pages 26–43. Elsevier Science Technology, San Diego, 2017.

[24] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, 2017.

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the Institute of Radio Engineers*, 86(11):2278–2323, 1998.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[27] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

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

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[31] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916. IEEE, 2015.

[32] Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. P2snet: Can an image match a video for person re-identification in an end-to-end way? *IEEE Trans. Cir. and Sys. for Video Technol.*, 28(10):2777â2787, oct 2018.

[33] Y. Cai and M. Pietikäinen. Person re-identification based on global color context. In *Asian Conference on Computer Vision*, pages 205–215. Springer, 2010.

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

[35] R. R. Varior, G. Wang, J. Lu, and T. Liu. Learning invariant color features for person reidentification. *IEEE Transactions on Image Processing*, 25(7):3395–3410, 2016.

[36] M. Munaro, S. Ghidoni, D. Tartaro Dizmen, and E. Menegatti. A feature-based approach to people re-identification using skeleton keypoints. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 5644–5651. IEEE, 2014.

[37] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptors with application to person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2179–2194, 2019.

[38] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012.

[39] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

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

[41] Ch. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu. Attention-guided cnn for image denoising. *Neural Networks*, 124:117–129, 2020.

[42] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. A full-image full-resolution end-to-end-trainable cnn framework for image forgery detection. *IEEE Access*, 8:133488–133502, 2020.

[43] Ch. Tian, R. Zhuge, Zh. Wu, Y. Xu, W. Zuo, Ch. Chen, and Ch. Lin. Lightweight image super-resolution with enhanced cnn. *Knowledge-Based Systems*, 205:106235, 2020.

[44] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016.

[45] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a âsiameseâ time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993.

[46] A. Subramaniam, M. Chatterjee, and A. Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in neural information processing systems*, pages 2667–2675. IEEE, 2016.

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

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 09 2014.

[50] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian conference on computer vision*, pages 31–44. Springer, 2012.

[51] M. Iacopo, L. Giuseppe, B. Federico, and B. Alberto Del. Person re-identification: Theory and best practice, 2015. share `http://www.micc.unifi.it/reid-tutorial/`.

[52] Sh. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206. IEEE, 2015.

[53] Y. Liu, Y. Zhang, S. Coleman, B. Bhanu, and Sh. Liu. A new patch selection method based on parsing and saliency detection for person re-identification. *Neurocomputing*, 374:86–99, 2020.

[54] M. Tian, Sh. Yi, H. Li, Sh. Li, X. Zhang, J. Shi, J. Yan, and X. Wang. Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5794–5803. IEEE, 2018.

[55] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2530–2539. IEEE, 2017.

[56] N. Perwaiz, M. M. Fraz, and M. Shahzad. Person re-identification using hybrid representation reinforced by metric learning. *IEEE Access*, 6:77334–77349, 2018.

[57] M. Fayyaz, M. Yasmin, M. Sharif, J. H. Shah, M. Raza, and T. Iqbal. Person re-identification with features-based clustering and deep features. *Neural Computing and Applications*, 32(14):10519–10540, 2020.

[58] H. Zhao, M. Tian, Sh. Sun, J. Shao, J. Yan, Sh. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1077–1085. IEEE, 2017.

[59] J. Wang, Zh. Wang, Ch. Gao, N. Sang, and R. Huang. Deeplist: Learning deep features with adaptive listwise constraint for person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):513–524, 2016.

[60] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1268–1277. IEEE, 2016.

[61] D. Tao, Y. Guo, M. Song, Y. Li, Zh. Yu, and Y. Y. Tang. Person re-identification by dual-regularized kiss metric learning. *IEEE Transactions on Image Processing*, 25(6):2726–2738, 2016.