Person Reidentification in Multi-Camera Network Using CNNs

Author

Jawad Qammar
18F-MS-CP-05

Supervisor

Dr. Waqar Ahmad
Assistant Professor

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF TELECOMMUNICATION AND INFORMATION ENGINEERING
UNIVERSITY OF ENGINEERING AND TECHNOLOGY
TAXILA

September 2021

Person Reidentification in Multi-Camera Network Using CNNs

Author

Jawad Qammar

18F-MS-CP-05

A thesis submitted in partial fulfillment of the requirements for the degree of

M.Sc. Computer Engineering

Thesis Supervisor:

Dr. Waqar Ahmad

Assistant Professor, Computer Engineering Department

External Examiner Signature:_____

Thesis Supervisor Signature: _____

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF TELECOMMUNICATION AND INFORMATION ENGINEERING
UNIVERSITY OF ENGINEERING AND TECHNOLOGY
TAXILA

September 2021

ABSTRACT

Person Reidentification in Multi-Camera Network Using CNNs

Jawad Qammar

18F-MS-CP-05

Thesis Supervisor: Dr. Waqar Ahmad

Assistant Professor, CPED, UET Taxila

The advancement of deep learning has facilitated rapid progress in person re-identification (re-id) task. Its applications in intelligent video surveillance made it a key component of today's smart cities infrastructure. Person re-id task is aimed to identify the person in distributed camera setup with non-overlapping views. The feature extraction process is an important part of person re-id technique. The present state of art methods mostly used ResNet as a backbone for feature extraction, which results in low geometric transformation modeling and low-resolution representation learning. We addressed these two major challenges by integrating deformable convolution module to enhance the transformation modeling capability and replaced the traditional ResNet backbone for person re-id with the novel feature extraction network named as HRNet, which is based on high-resolution representation learning without any additional supervision. The verification of our approach performance is done by conducting an experiment on a person re-id dataset named Market-1501. We achieved 90.57% Rank-1 accuracy and 75.43% mAP, outperforming the ResNet baseline results, which confirmed the effectiveness of our approach and will have a promising future in person re-id.

Keywords: Person re-identification, intelligent monitoring, convolution neural networks, deep learning.

UNDERTAKING

I certify that research work titled "*Person Reidentification in Multi-Camera Network Using CNNs*" is my own work. The work has not been presented elsewhere for assessment. Where material has been used from other sources it has been properly acknowledged / referred.

Jawad Qammar

18F-MS-CP-05

ACKNOWLEDGEMENTS

First, I would like to thank Almighty Allah, who gave me the courage, skills, wisdom, motivation, dedication, and everything that I required in completing this thesis. His blessings provide me the ability to make an important contribution to the repository of knowledge and experience that is currently available.

I am grateful to Dr. Waqar Ahmad, my honorable supervisor, for his encouragement, support, and direction throughout my thesis work. He aided me in grasping the core concepts underlying my research domain in order to accomplish my objectives.

I would like to express my gratitude to my parents for their unwavering support and encouragement during this thesis's completion. Additionally, I am indebted to my friends, university faculty, coworkers, juniors, and seniors for their unwavering support and assistance, which I required most of the times.

The author would also like to thank every researcher whose work playing its role in the improvement of the Person Re-ID task.

Thank you very much!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The emergence of smart cities has created a large amount of data that can be utilized to obtain meaningful insights out of it. Among many other components of smart cities, the smart camera's network is of prime importance as it has many useful applications in intelligent transportation, security, forensic, etc.

1.1 Camera Network

1.1.1 Background

Video surveillance systems monitor the behavior, activities, or other changing information of people by means of electronic equipment. Today a huge amount of video surveillance or closed-circuit television (CCTV) cameras are installed throughout the world. Based on recent shipment figures and expected product lifespans, it is estimated that the total number of video surveillance cameras worldwide will be 1 billion by 2021. These cameras occur in various domains ranging from rather small home surveillance applications in private areas, to medium-sized and large installations for monitoring public areas, e.g., shopping centers, airports, train stations, public transportation, sports centers and so on.

The system was very quickly popularized in the world. In Pakistan, more than 8 thousand surveillance cameras are installed in Lahore safe city project only and the demand is on the rise.

1.1.2 Importance of Video Surveillance

The reason for this enormous increase is probably twofold: firstly, cameras have become relatively cheap and secondly, during both day and night, they help to protect both people

and property. The stated goal of the installation of CCTV cameras is to reduce crime and enhance the safety of the public. The CCTV cameras get the images and send them to a video surveillance center where all scenes are viewed (real-time mode) or stored (posteriori mode) as shown in Fig. 1.1.

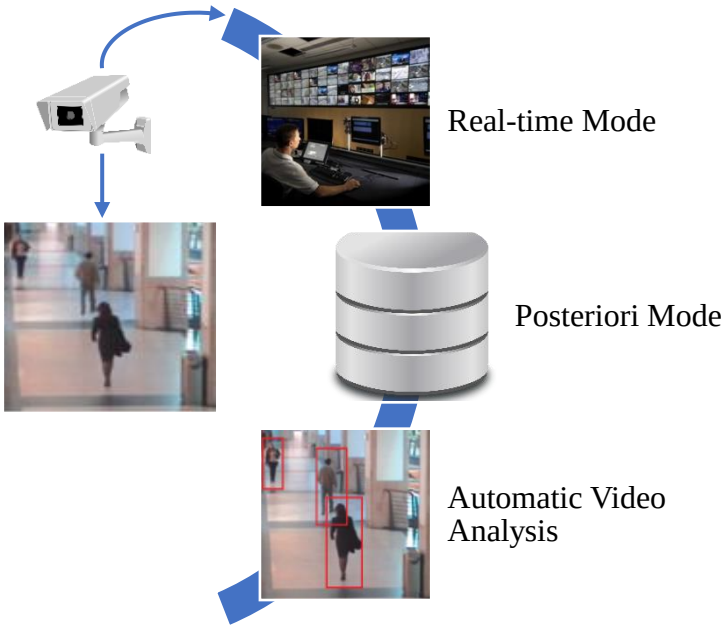

Fig. 1.1: Different operating mode of video surveillance system.

It is an effective method of crime prevention. Surveillance cameras have been found to have a deterrent impact on potential offenders, which leads to a drop in overall crime rates. In the event of a real-time scenario, it can be used to prevent crimes. These security installations assist in watching the people who enter a location and to keep an eye on their activities in case of any form of threat is detected. Policing or security agents could rapidly intervene ahead of time to prevent incidents by using real-time behavioral surveillance. Additionally, it is efficient for locating evidence following an accident or an attack in the a posteriori mode. Because of this recorded footage, the people or vehicles in concern can

be identified, as well as the timeline of the incident or a crime. Additionally, it can be utilized for traffic monitoring, industrial processes, and other purposes.

1.1.3 Video Analysis Methods

Most present video surveillance systems merely provide the infrastructure for video capture, storage, and distribution, leaving the work of danger detection entirely to human operators. It is a tremendously labor-intensive process to monitor surveillance video of humans. A large amount of visual attention is required to observe a video transmission. Specifically, being attentive, being able to focus, and being quick to react to unlikely events can be challenging because slips in attention can lead to errors. Moreover, millions of hours of video data generated by a large number of cameras require more and more operators for the task. It’s almost infeasible to achieve real-time prevention due to the high cost, thereby reduces the effectiveness of surveillance severely. With the proliferation of digital cameras and the advent of powerful computing resources, automatic video analysis has become possible and more and more common in video surveillance applications, thus reducing considerably this cost.

For the real-time mode, the purpose of automatic video analysis for surveillance and security is to notice occurrences and circumstances that require security personnel's attention. Automated analysis of vast amounts of video footage not only speeds up data processing but also improves the capacity to predict problems dramatically. Automatic processing augments security personnel, increasing their efficiency and efficacy.

For the posteriori mode, given person of interest is searched in thousands of hours of recorded videos provided by many cameras, requires assigning a large number of enforcement officers to this task, and requires a lot of time to be performed. Automated

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

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

1.2 Person Re-identification

1.2.1 Definition

Person re-id is the task of finding the query image of a person in the image gallery, in a multi-camera network having a non-overlapping field of views. The task is different from the classic identification and detection tasks. The identification consists in determining the person identity in an image and the detection consists in discriminating people from the background without knowing the identity. Re-identification identify whether a given image is of the same person as a query image. The identification task helps us to know who it is, and the detection task indicates whether it is a person. But the re-identification tells when and where this person appeared with respect to a given camera and, using several cameras, potentially allows for the estimation of his/her trajectory over a short period of time.

1.2.2 Re-id in Multi-camera Network

The emergence of smart cities has created a large amount of data that can be utilized to obtain meaningful insights out of it. Among many other components of smart cities, the smart camera's network is of prime importance as it has many useful applications in

intelligent transportation, video surveillance, security, forensic, etc. Person re-identification (re-id) is the main key role in all these applications. Person re-id is the task of finding the query image of a person in the image gallery, in a multi-camera network having a non-overlapping field of views. The query image can be from a different camera or the same came on different occasions and does not overlap with gallery images. In person re-id, the query image is searched in the image gallery to find its best match by considering the features contained in his/her whole body (e.g., clothes, color, height) instead of only facial features which works when the person is near to the camera and faces it. But this is not typically the case in CCTV video, and one may also hide the face from the camera.

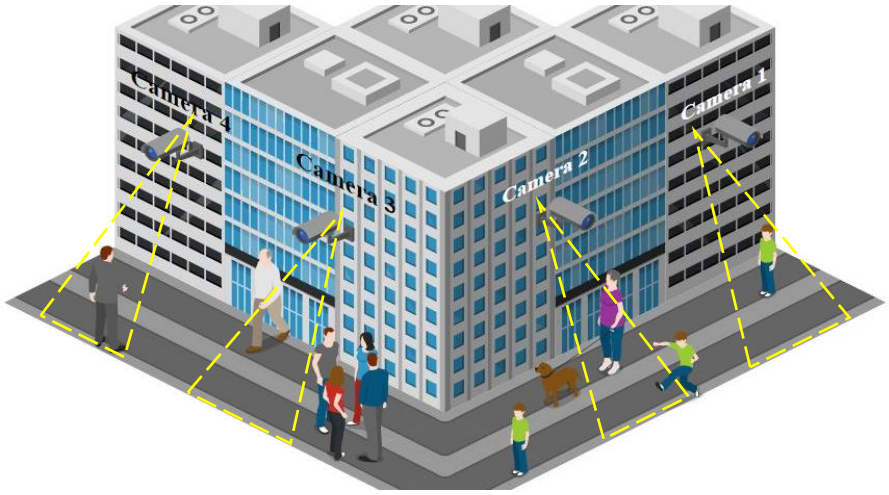

Fig. 1.2: Person re-id illustration using multi-camera surveillance network.

An example of a multi-camera network is shown in Fig. 1.2. The figure depicts the side view of a public place where the different persons are passing through, four cameras are placed in a non-overlapping field of view settings. Camera 2 will re-identify the person which is moving from camera 1 as shown in Fig. 1.2.

1.2.3 End to End Practical Person Re-Id System

Re-ID is an easy task for human by exploiting person descriptors based on clothing, walking pattern, weight, height, hairs, face, etc., but it is an incredibly challenging task for a computer to solve, due to the existence of various viewpoints [1], low-resolution images [2], illumination [3], occlusion [4], misalignment, color, etc. In recent years, several deeply learned methods have been proposed, which greatly boosted the person Re-id task, due to the superior feature learning capability of deeply learned methods [5] as compared to handcrafted methods [6].

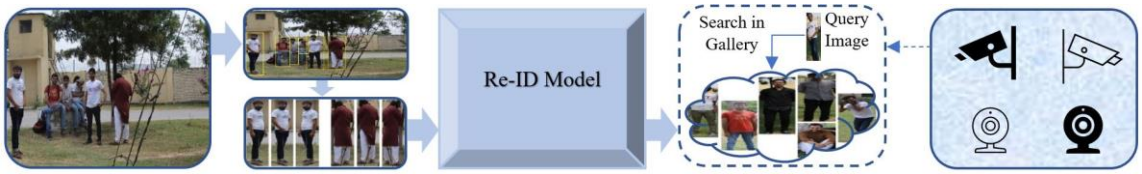

Fig. 1.3: Demonstration of end-to-end practical person re-id system from the collection of Raw image data, Person detection, Person tracking, Feature extraction & Descriptor generation using re-id model and to Matching & Ranking of the query image.

An end-to-end practical re-id System consists of seven key steps named as Raw Image Data, Person Detection, Person tracking, Feature extraction, Descriptor generation, Matching, Ranking. The overall structure of the re-id system is illustrated in Fig 1.3 [7]. Firstly, the raw data is collected from cameras, which contain a lot of background clutter. Then person detection or tracking algorithms [8] are used as it is practically not possible in large-scale data to manually crop all the person images. The feature extraction and descriptor generation are done using CNN [9], which is trained on person image data. To overcome the person re-id challenges, numerous models have been developed based on distance metric learning [10], features representation learning [11], or by combining both. The unknown Ids obtained using descriptor generation are then compared with known Ids from the gallery using similarity analysis techniques [9], to obtain a retrieved ranking list.

The first three steps are generally considered to be independent tasks in computer vision, therefore most of the re-id research is focused on the remaining steps [7].

Fig. 1.4: General pipeline of person re-identification model.

The general pipeline of person re-identification model is shown in Fig. 1.4. The person reidentification is based on the pedestrian detection task. In the first step, we form a pedestrian gallery set by collecting the cropped pedestrian images or extracted pedestrian image signature from each camera scene in the network. The CNN-based approaches for re-identification of individuals rely solely on the CNN model, which is integrated with various networks and matching algorithms. Then, we measure the similarity or distance to the query image. Finally, we show the best matched images according to the measured similarity. An assumption that is generally made is that individuals keep the same clothing in different scenes. This is most effective in a circumstance where the timeframe is limited to a comparable appearance, ensuring the constraint of a visual presentation. Re-identification cannot be used to determine similarities between persons after many days due to likely visual appearance changes. The reason for this hypothesis is that more accurate biometrics like faces are not always available in videos in "far-sight" surveillance settings, which are the most common in practice. In that case, the re-identification algorithms rely mainly on the overall appearance of the individual.

Most surveillance cameras have overlapping fields of vision to cover a vast geographic region. If a single person left the area without revealing their identity, they would have to be discovered and identified somewhere else within the area over time. They would be

distinguishable from multiple people in the area who seem similar but are different. When photos of people are recorded without a sufficient temporal or spatial continuity, data association can be accomplished through re-identification. This may be utilized for a long-term study of people's activities and behavioral patterns and for a wide range of software applications.

1.2.4 Relationship with Classification and Retrieval

The specific nature of a person's re-identification rests in the spot between image classification and instance retrieval in the context of training and testing classes. Training images are available for each class, and testing images fall within these predetermined classes, labelled in Table 1.1 as previously "seen." For example, in retrieval, there is typically no training data because the query's content is unknown in advance and the gallery may contain a variety of different types of objects. Thus, training classes are marked as "not available," whereas testing classes (queries) are marked as "unseen."

In comparison to image classification, person re-ID is comparable in that training classes with images of various identities are accessible. Also, persons' re-ID is the same as instance retrieval in that they have no overlap with the training IDs; this excludes the training images but includes the testing images, which depict pedestrians.

Table 1.1: Comparison of various computer vision algorithm.

Task	**Train Class**	**Test Class**	**Advantage**
Classification	Available	Seen	Discr. Learning
Retrieval	Not Available	Unseen	Efficiency
Person re-ID	Available	Unseen	Discr. + Efficiency

As a result, person re-id might be structured to benefit from both classification and retrieval. On the one side, feature embeddings [12] or discriminative distance metrics [13] can be trained in the person space using training classes. On the other side, efficient indexing systems [14] and hashing algorithms [15] can aid with re-ID in a huge gallery.

1.2.5 Applications of Person Re-identification

A huge potential for practical applications in several areas, spanning from security and surveillance to retail and health care, exists in re-identification approaches.

- **Cross-camera person tracking:** Computer vision requires the capacity to monitor people across numerous cameras, analyze crowd movement, and distinguish activities. Person re-identification is used to build connectivity between unconnected tracks to monitor a person across numerous cameras. This permit retracing a person's path across a scene.
- **Tracking by detection:** Person reidentification could be useful even with single camera monitoring. Tracking many persons is difficult in crowded situations with frequent occlusions and interactions. The aim is to model data from movies that involve long-term occlusion and varying numbers of people. Tracking-by-detection uses human detection techniques to track. People are detected, estimated mobility patterns are estimated, and detections are associated in distinct frames. Data association is a form of re-identification.
- **Person retrieval:** In this scenario, re-identification involves recognition. A target person's specific query is searched in a vast database. Re-identification is used for image retrieval and gives ranked lists, comparable items, etc.

- **Human-machine interaction:** In this, Re-id can be considered as "non-cooperative target recognition" in robotics, where the interlocutor's identity is retained, allowing the robot to be always aware of its surroundings.
- **Long-term human behavior and activity analysis:** For example, analyzing customer shopping trends by observing them touching, surveying, and trying products in stores under different surveillance cameras. Another example, geriatric health care analysis explores the elder people’s long-term behavior to assist doctors to make more accurate diagnoses.
- **Surveillance in security:** To detect if a person is present at a specific area at a specific moment, and then use the detected location to estimate where the person is travelling by tracking their path.
- **Retail or shopping:** In order to give helpful data for improving client service, such as individualized product recommendations, as well as data for managing retail spaces.
- **Healthcare**: For the purposes of caring and monitoring, such as tracking patients as they move about a hospital.
- **Personalization of services in smart home:** To track a person's location within a home, to activate a personalized television or radio station when a person enters a living room, and to provide coffee making service when enters a kitchen.
- **Personalization of services in public spaces:** To provide visitors with personalized services such as personal guides that deliver narratives about the show and a personal virtual collection, in order to create more engaging encounters.

- **Speech, gestures, and facial expressions:** Additional application areas for human-robot interaction include education, home, entertainment, field robotics, and companion robots, hospitality, rehabilitation and elder care, and robot assisted therapy.

1.2.6 Challenges

Solving the person re-id problem is inherently challenging. To match a person across different scenes, it has to deal with intra-class variation, for example the same person under different views may undergo large appearance changes, and to avoid inter-class confusion, i.e., distinct people can appear similar across camera viewpoints. Some challenging examples are shown in Fig. 1.5. The challenging factors and their effects are explained in the following.

Fig. 1.5: Person re-id challenges examples. Except (g) all have the same person. (a) viewpoint variation, (b) pose variations, (c) illumination changes, (d) partial occlusion, (e) inaccurate pedestrian detection, (f) accessory change (the person has a back bag in the first image, but not in the second), (g) low resolution, (h) different people with similar clothing [17].

- **Illumination variation:** Illumination conditions can vary in different camera scenes or during the day. The same person observed under different lighting conditions can have a color difference on the appearance. This increases the intra-class variation.
- **Camera viewpoint variation:** Since the camera's height, the distance between the person and the camera and the direction in which the people are facing are varying, different shapes or sizes of pedestrians can be observed under different viewing angles. A person cannot be viewed from 360 degrees in a single image. Each view contains in fact partial information about the person's appearance. Some parts are not visible in one viewpoint but could be observed in another. In terms of shape, person images from different people from the same viewpoint may look more similar than two images from the same person from distinct viewpoint. The viewpoint variation is one of the most challenging problems which increases at the same time the intra-class variation and the inter-class confusion.
- **Pose variation:** The human body's articulation results in deformations in the appearances of the same individual. A learned model on standing pose will probably fail to detect a running, crouching or a sitting person. Pose variations imply that the body part localization and visibility changes within a given bounding box. And is difficult to predict in the resulting images, which are most often of relatively low resolution and quality.
- **Low resolution:** In most realistic settings, the cost of the required number of cameras in all zones could be very high, so that the coverage is rather sparse, leaving "blind gaps". So, cameras are usually installed in high places on walls, and pedestrians are

thus usually far away from the camera. Even for high resolution cameras, for a given person, the image could still be of relative low resolution.

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

- **Large number of candidates in gallery set:** A camera network may cover a large public space, like a train station or a campus. Thus, there can be a huge amount of candidate for a given re-identification query, and the number of candidates increase over time. The computation for matching with a large gallery set becomes expensive. Temporal reasoning and camera positioning can be employed to help with this issue.

- **Similar clothing:** In a large gallery set, there is a high probability that people have similar clothing. It is standard practice for people in public settings to wear dark clothing in the cold, as a large number of people dress in practically identical blue jeans. As a result, the matching process becomes more ambiguous and unpredictable. In this case, it's more difficult to find the discriminative signature from the visual appearance of different people.

- **Partial occlusion**: Persons are sometimes partially or fully obscured by overlaps with other people or by environmental architecture. If some important or discriminative

parts are not visible, the matching fails probably. It happens when people are walking in group or in a public space which is crowded.

- **Real-time constraint:** In some emergency situations, we should find the location of a suspect immediately. It is critical to have a real-time, low-latency system capable of processing many input video streams and quickly providing query results. With several prospective candidates to be discriminated, the search space for person matching can be exceedingly wide. Thus, the searching time is crucial.

- **Clothing or accessories change:** In realistic setting, the appearance constancy in the person re-identification problem could be easily violated. The larger the time and space difference between views, the more likely it is that persons will appear in separate camera perspectives wearing different outfits or carrying different goods. For example, taking the bag pack from back in hand or taking off a hat.

- **Camera setting:** This object appears to have color differences as it is acquired by different cameras. The same person with the same clothes can be rendered in different ways. There may also be some geometric differences. For example, the shape of a person may be observed with varying aspect ratios.

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

- **Data labeling:** This is a common difficulty in the computer vision field. Training a good model that is resistant to all variants in a supervised way couldn't be done without

a sufficient amount of annotated data. Manually collecting and annotating of data for a large camera network, is very expensive.

1.3 Problem Statement

Person re-id is the task of finding the query image of a person in the image gallery, in a multi-camera network having a non-overlapping field of views. The query image can be from a different camera or the same came on different occasions and does not overlap with gallery images. In person re-id, the query image is searched in the image gallery to find its best match by considering the features contained in his/her whole body (e.g., clothes, color, height) instead of only facial features which works when the person is near to the camera and faces it. The present multi-camera networks deploy cameras that are of low resolution due to outdated cameras or the resource constraints such as processing power, bandwidth, cost, etc. However high resolutions are needed for position-sensitive tasks. The person detection is done using deformable part model [16] in large scale dataset such as Market-1501 [17], results in misalignment problem in which we may have an excessive background or missing body parts.

1.4 Proposed Framework

The feature extraction process is an important part of person re-id technique. The present state of art methods mostly used ResNet as a backbone for feature extraction, which results in low geometric transformation modeling and low-resolution representation learning. In this latter, we address both these problems by replacing the traditional ResNet baseline with High-Resolution Network (HRNet) [18] and adding a deformable convolution module [19] to it. We integrate deformable convolution module to enhance the transformation modeling capability and replaced the traditional ResNet backbone for person re-id with the

novel feature extraction network named as HRNet, which is based on high-resolution representation learning without any additional supervision.

1.4.1 Aims and Objectives

The goal of this research is to design and analyze CNN based method by incorporation deformable convolution module into it. The aims and objectives of the study are set as follows:

- To improve the person reidentification in a multi camera network using CNN.
- To integrate the deformable convolution module in traditional baseline for enhancing the geometric transformation modeling capability.
- To replace the traditional ResNet backbone for person re-id with the novel feature extraction network named as HRNet for high-resolution representation learning.
- To design a person re-ID technique based on HRNet and deformable convolution module, which will address the low-resolution representation learning and the misalignment problem for person re-ID architecture.
- To evaluate and validate proposed design against well-established performance criteria.

1.5 Thesis Organization

The remaining part of this thesis is organized as follows: in chapter II, the brief description of related work is elaborated. Chapter III provides the description of proposed Person Re-ID technique in detail, which includes the briefing about the dataset and evaluation metric. The implementation of the experiment at different stages are also provided in detail. Chapter IV contains information regarding experimental results and performance analysis.

In chapter V, a brief explanation about the conclusion and future work for the proposed person re-ID technique are briefly described.

1.6 Publications

As part of outcomes of this research, the following conference paper have been produced.

1.6.1 Conference Paper

- J. Qammar and W. Ahmad, "Resolution Representation Based Person Re-Identification for Smart Cities Using Deep Neural Networks (DNNs)," 2021 International Conference on Information Technology (ICIT), 2021, pp. 533-537, doi: 10.1109/ICIT52682.2021.9491740.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we present the literature related to person re-id. The first section introduces the deep learning techniques. We start from introducing the classical neural networks, then we present Convolutional Neural Networks (CNN), and some of its complex variants as well as the Siamese Neural Networks which are often applied in person re-id. The second section concerns the state-of-the-art in person re-id.

Intelligent video surveillance systems for urban roads have advanced significantly in recent years. The cameras that have been placed as sensing devices are linked together to make a sensing network. Intelligent video surveillance systems are a critical component of smart city infrastructure. The cameras installed in metropolitan cities monitors the traffic and person movements which can be utilized as IoT service to ensure the safety of citizens using multi-camera network.

The data created by digital surveillance is now so huge in today's society, with CCTV cameras installed at every step, that it is impractical for a person to make sense of it. Machine vision techniques (together with developments in hardware that can support parallel processing) that can analyze the large amounts of data and deliver useful information have provided some solutions.

The vision technique we are going to address in this paper is slightly different than the commonly used techniques in cameras and social media like face detection and face recognition, which uses facial features provided the person facing the camera and near to it. In person re-identification, the query image is searched in the image gallery to find its

best match by considering the features contained in his/her whole body (e.g., clothes, color, height) instead of only facial features.

2.1 Deep Learning Introduction

In recent years, neural network-based deep learning algorithms become a popular branch of machine learning. Deep learning methods make use of several processing layers with complicated structures or are otherwise built of multiple non-linear transformations to attempt to represent high-level abstractions in data. In the field of computer vision, natural language processing, robotics, just to name a few, they have proven to perform better than state-of-the-art methods.

2.1.1 Neural Networks

As their name indicates, neural networks are inspired from biology and the human brain. The nervous system, which is made up of nerve cells or neurons, regulates human behavior. In the human brain, there are around 85 billion neurons. As shown in Fig. 2.1, a neuron reacts with other neurons to pass the information. When a neuron's dendrites receive excitatory input, the neuron's membrane potential steadily increases. If the voltage of membrane hits a predetermined threshold, an action potential is started and propagated via the axon of the presynaptic neuron to postsynaptic neurons.

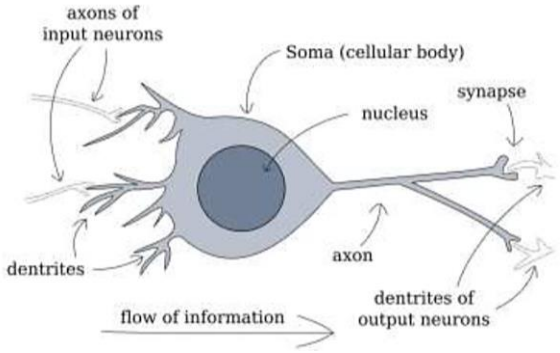

Fig. 2.1: Structure of a biological neuron [20].

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

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

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

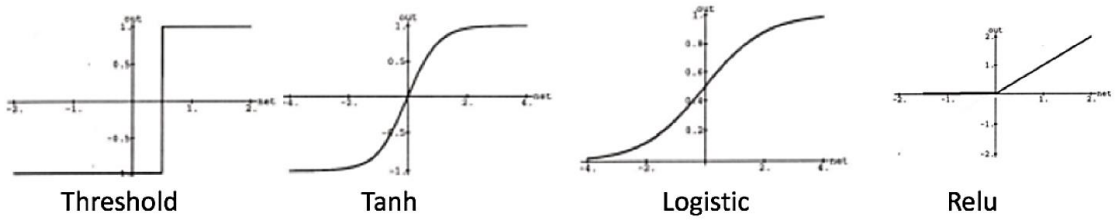

Fig. 2.2: Different neural activation functions.

Artificial neurons are capable of doing just rudimentary computations and acting as a weak classifier on their own. They may, however, be ordered to create sophisticated operations through neural networks. While neurons can be ordered randomly in theory, they are frequently placed in an acyclic graph in practice, represents that even indirectly their input does not depend on their output. Because activations can propagate forward in the network, neural networks with this architecture are referred to as feed-forward neural networks. Recurrent neural networks, on the other hand, can have cyclic connections. While recurrent neural networks may be better at representing dynamical systems, the presence of cycles significantly complicates training.

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

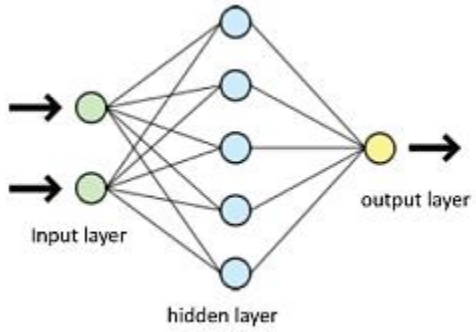

Fig. 2.3: Illustration of a multi-layer perceptron.

The basic idea of neural network is to combine the elementary simple function to fit whichever function by toning the parameters of the network. According to the universal approximation theorem [20], assuming minimal assumptions on the activation function, feed-forward neural networks with a single hidden layer may approximate any continuous function on compact subsets of R'.

2.1.2 Backpropagation

Backpropagation [21] is one of the most commonly used training algorithms for feed-forward neural networks and may be the most ubiquitous. It is an application of the gradient descent optimization algorithm in case of neural network. There are two phases of Backpropagation algorithm: (1) a forward phase between the input and output layers for the purpose of computing neural activations (2) To compute the errors, forward propagate the gradients and update the weights, there a backward phase from the output layer to the

input layer. For a batch of training samples, a cost function E is generated between the computed outputs and their desired targets.

2.1.3 Motivations and Problems of Deep Neural Networks

By definition, a deep network comprises of multiple layers, but a feed-forward network has a hidden layer is already a universal approximator. So, there could be 2 main reasons to form a deep neural network. Firstly, the deep network is more efficient. Some mathematic results [22] show that certain function, which is representable by a neural network of depth d, need an exponential number of parameters with a network of depth d. However, these results are rather theoretical, and it is not clear to what extend this holds in practical applications

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

Although, deep networks extract rich and high-level features and reduce the need of one of the most time-consuming components of machine learning is feature engineering, increasing the depth does not necessarily improve their performance. Deep learning algorithms also have mainly two disadvantages. Firstly, they easily suffer from the over-

fitting problem, that means the model does not generalize well to real-world cases although it fits the training data well. Also, the deeper the network, the more difficult to train and the more training data we need for convergence.

The second disadvantage is the vanishing gradient problem [23]. Due to the fact that the gradient is backpropagated to prior layers, repeated multiplication may result in an infinitely tiny gradient. The error gradients vanish quickly in an exponential fashion with respect to the depth of the network. As a result, as the network becomes more established, its performance becomes saturated or even degrades significantly. It turns impossible to train a deep network. To improve these two problems, a number of solutions have been proposed. We will introduce some of them, especially for vision tasks, in the following sections.

2.1.4 Convolutional Neural Network

It is a feedforward neural network with a complex structure that includes convolutional calculation. Three distinct types of neural network layers comprise a convolutional neural network: the convolutional layer, the pooling layer, and the fully connected layer as shown in the Fig. 2.4.

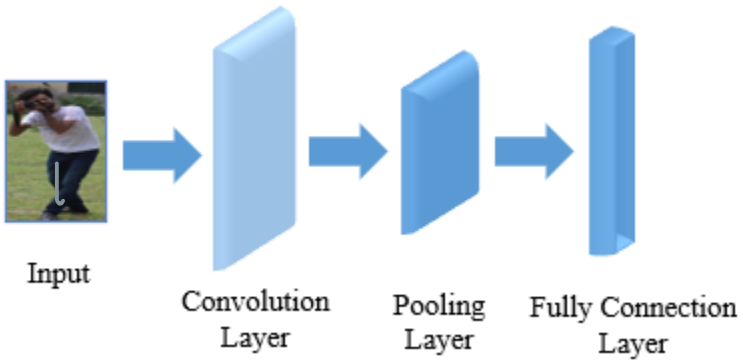

Fig. 2.4: Basic structure diagram of convolutional neural network.

2.1.4.1 Convolutional Layer

The feature representation of input data is mainly learned via the convolution layer. Convolution operations are applied to the input image, using multiple convolution kernels, to compute various feature maps.

The input data is an RGB image, as illustrated in Fig. 2.5. If the color image is 6*6*3, the three indicates the presence of three-color channels, then the convolution operation is performed using a 3*3*3 convolution kernel to represent the red, green, and blue channels. Multiply each of the 27 integers by the values in the appropriate red, green, and blue channels, and then add them all together to obtain the first value in the feature graph's output.

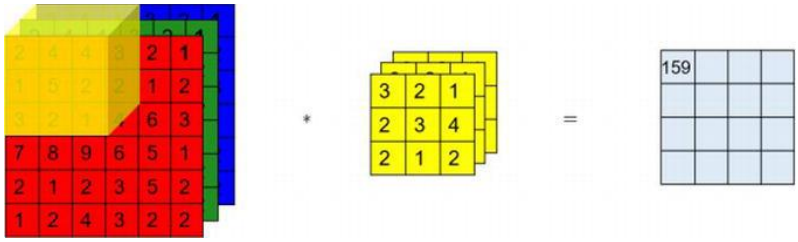

Fig. 2.5: Convolution diagram.

The following equation represents the principle of convolution layer

$$x_j^l = f\left(\sum_i x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{2.1}$$

where f(∗) represents the activation function; x_j^l represents the j[th] feature map of output layer l, x_j^l represents the layer l i[th] feature map, k_{ij}^l represents the current input layer i[th] feature graph of the convolution kernel and the output layer j[th] feature graph on the layer l, b_j^l is the bias term of layer l j[th] feature graph.

2.1.4.2 Pooling Layer

Pooling layer is frequently employed in the convolutional network for the aforementioned reasons: to minimize model size, boost processing performance, and enhance extracted feature robustness. When two or more objects are being manipulated together, the result is translation, rotation, and scale invariance. Routine procedures performed on the pooling layer include averaging and pooling. In Fig. 2.6, the maximum pooling operation is depicted. Divided into separate zones, the input of 4*4 consists of four different sections. The result is the maximum value for each color zone.

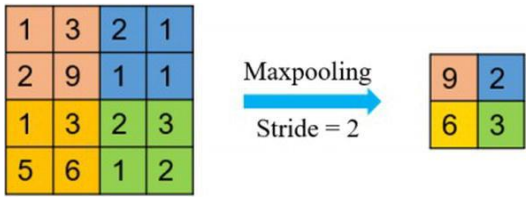

Fig. 2.6: Max pooling diagram.

Reducing the number of parameters in the input tensor is the fundamental goal of a pooling layer thus overfitting is decreased, useful features are extracted from the input tensor, and computational costs are thus lowered. The Pooling layer is receiving tensor as an input. Two kinds of pooling exist:

1) Average Pooling

2) Max Pooling

In Average Pooling, a kernel of size n*n is moved across the matrix and using all the values the average is computed for each position and placed in the corresponding location of the output matrix.

In Max Pooling, a kernel of size n*n is moved across the matrix and from all the values the maximum is taken for each position and placed in the corresponding location of the output matrix. An example is shown in the Fig. 2.6.

Every channel is assigned an independent value in the input tensor. As a result, we receive the tensor's output. However, in contrast to Pooling down samples, the number of channels (depth) is not affected, even though the image's height and breadth are pooled.

2.1.4.3 Fully Connection Layer

The features that were retrieved from the preceding layer are interconnected to every node of the fully connected layer. Additionally, the general completely connected layer also has the most parameters because of its fully connected nature. In the distributed feature representation learning phase, the full join layer acts as a mapping of the learnt "distributed feature representation" into the sample tag space. A linear transformation takes place from one eigenvector space to another. It is also possible to treat every dimension of the target space with respect to all dimensions of the source space. This style of CNN architecture can be found in the last few layers of the network and is used to compute a weighted sum of the preceding layers' features. This whole connecting layer's schematic architecture is seen in Fig. 2.7.

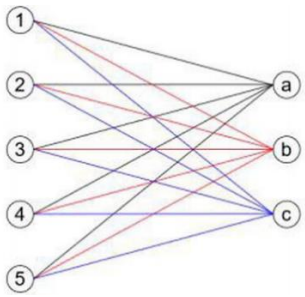

Fig. 2.7: Fully connection layer.

Fully Connected Layer form the last few layers in the network is simply a feed forward neural network. After flattening the output from the final Pooling or Convolutional Layer, it is fed as input into the fully connected layer. Instead of ReLU, the SoftMax activation function is utilized to produce probability of the input being in a given class after passing through the fully connected layers (classification).

Thus, we now have the probability that the object in the image belongs to one of the various classes. Thus, the Convolutional Neural Network operates. And the input photographs are labelled.

2.1.5 CNN Architectures

In this section, some important CNN architectures will be introduced from the literature and explain how the basic architecture was improved to achieve state-of-the-art performance in image classification.

LeNet [24] is the most classic CNN which has been developed by Yann LeCun in the early 1990s for handwritten character recognition as given in Fig. 2.8. The CNN contains 3 types of layers. First, a convolutional layer performs convolution operation on the input image. As the convolution coefficients are learnt automatically by backpropagation, which means that there is no “manual” extraction of features. Then, in a local region the pooling layer aggregates the information, and the next hidden layer will obtain the summarized statistics. It can be the maximum or average value, for examples. This permits to reduce the number of parameters and to be invariant to the small translations in the Image. Another convolution and pooling layer repeat this operation to extract higher-level features. At the end, the completely connected layers perform the classification as a feed-forward neural network (like an MLP).

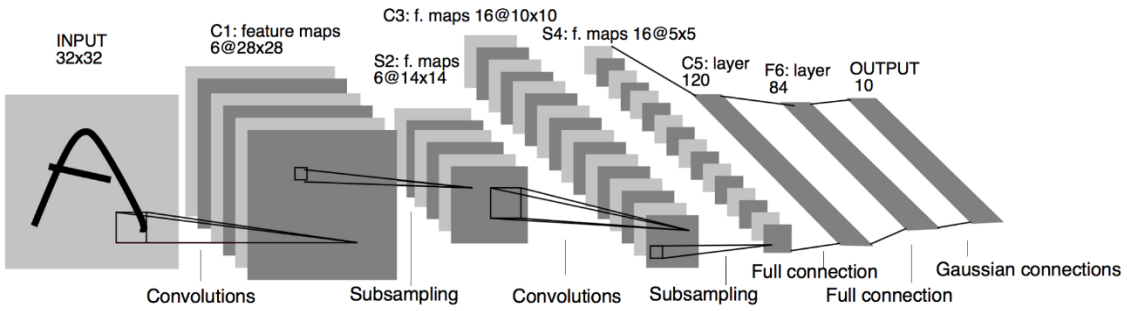

Fig. 2.8: Diagram of LeNet-5 [24].

AlexNet [25] is the winning model in 2012 in ImageNet Large Scale Visual Recognition Competition (ILSVRC). ImageNet is a collection of over 15 million high-resolution photos that have been categorized and classified into around 22,000 categories. ILSVRC makes use of a subset of ImageNet that has approximately 1000 photos for each of 1000 categories. AlexNet improved the ImageNet classification accuracy significantly in comparison to traditional approaches. The overall architecture is made up of five convolutional layers, followed by three fully connected layers. In AlexNet, the authors proposed solutions of vanishing gradient problem and over-fitting problem.

A new activation function named ReLU (Rectified Linear Unit) is used in the network for the non-linear transformation. The advantage of the ReLu is that it needs light computation and, more importantly, it alleviates the vanishing gradient problem. Since one reason for this problem is that in the saturating area, the derivative of sigmoid is absolutely minimal. But the derivative of ReLU is equal to 1 when x is greater than zero, but otherwise it is 0 (see Fig. 2.2). So, the advantage of ReLU is that when its derivative is backpropagated there will be less degradation of the error signal. The over-fitting problem is also reduced by the dropout technique, which randomly drops some units after every fully connected layer from the network during training. Dropout technique has a probability p, and the

values are individually applied to every neuron in the feature map. With a probability of p, it randomly deactivates activation. One motivation of dropout is to make units learn meaningful features independent from others and to avoid co-adaptations among them. Another view of dropout is related to model ensembles. In fact, different activated subsets of neurons represent different architectures, and training of these architectures are done in parallel, each architecture receiving weight values based on the number of architectures in use. The weighted sum of these random architectures thus actually corresponds to an ensemble of different neural networks.

VGG [26] net replaced large kernel-sized filters in Alexnet 11 in the first and 5 in the second convolutional layer, with multiple 3→3 kernel-sized filters. Multiple convolutional layers with smaller kernel size can perceive the field of the same size as a larger size kernel with fewer parameters. Additional non-linear layers add depth to the network, making it capable of learning more complicated features.

ResNet [27] or deep residual network has been proposed by He et al. in 2015. The authors noticed a phenomenon with training a deep network: A degradation problem has been shown as deep networks converge. Accuracy gets saturated as the network depth increases, and subsequently declines rapidly. The authors of Resnet point out that a deeper network should at least get the same accuracy as a shallow network, because the early levels of a deeper model can be substituted with a shallow network and the later layers can simply serve as an identity function. To overcome this degradation problem, the authors proposed to learn a residual function instead of learning a direct mapping function H(x) (A few stacked non-linear layers with x being the input of the layers). The residual function is defined as F(x)=H(x)-x, where x represents the identity function. So, we can reform the

equation as H(x)=F(x)+x. According to the author, it is easier to optimize the residual mapping function F(x) than to optimize the original, unreferenced mapping Fl(x). Intuitively, it is easier to learn a function like F(x)=0 rather than F(x)=x using stack of non-linear convolutional layers as function. To implement this idea in a CNN, the authors add some so-called skip-connections, shown in Fig. 2.9, to add the input of one layer to the output after one or more layers. This essentially drives the new layer to learn something different from what the input has already encoded. It also alleviates the vanishing gradients problem by allowing the gradient to flow without any changes from the top layers to the bottom by means of identity connections. It leads to the fact that very, very deep networks can be trained. The winning residual network of the ImageNet Challenge 2015 has 152 layers.

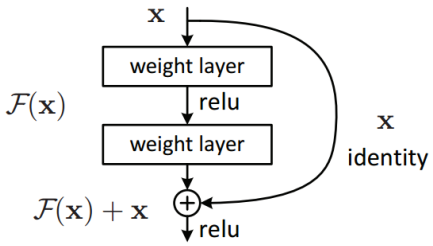

Fig. 2.9: A residual learning block used in deep residual neural networks [27].

With these proposed robust CNN models, the classification accuracy on ImageNet has been continuously improved. In practice, deep convolutional neural networks, on the other hand, have a large number of parameters, frequently in the millions. Training such a q network on a limited dataset has a significant impact on the model's generalizability, frequently resulting in overfitting. Therefore, one practical solution is fine-tuning an existing neural network model. There are some networks that are trained on a large dataset such as ImageNet can be used as initialization for the network. And we replace the last layer and

continue to train on the available smaller dataset on another task. Usually, a small learning rate should be used. It is possible to train a model on a dataset that is radically different from the original dataset, such as ImageNet, and the model will learn features that are relevant to the new objective.

2.1.6 Siamese and triplet neural networks

Siamese Neural Networks are neural architectures that receive a pair of examples at the input and produce an output vector. This metric, known as non-linear similarity, is measured by continuously presenting pairs of instances to be classified, for example, ones that belong to the same class and those that do not. The concept is to train the neural network to map the input vectors into a non-linear subspace, where a simple distance in this subspace, for example, the Euclidean distance, is a good approximation of the "semantic" distance in the input space. A short distance should be seen between two images that represent the same category, and a significant distance should be observed between two images that represent a different category.

Although Time Delay Neural Networks (TDNN) were first used in [28] to introduce Siamese Neural Networks, this time delay technology has now been employed to solving a different problem, namely signature verification, which involves verifying the legitimacy of signatures. After this approach was successfully implemented by Chopra et al., [29] who used Siamese CNNs, it was then applied by the group of researchers of Siamese CNNs (also known as "Siamese CNNs") in the context of face verification. A more specific definition of the system's capability is: The system is provided with two face images, and it must determine if they belong to the same person or not.

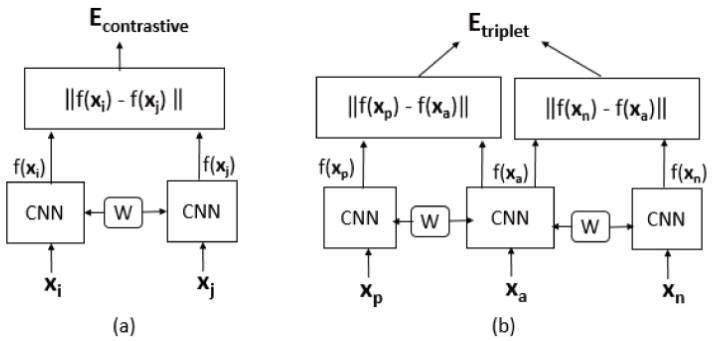

Fig. 2.10: Diagram of (a) Siamese neural network and (b) Triplet neural network [29].

The most common loss function used by the Siamese architecture is the contrastive loss. The contrastive loss function was calculated based on the Euclidean distance. The goal is to keep similar items at similar distances from each other and to keep any two data sets from being too far apart. The triplet neural architecture is a variant of the Siamese architecture. Instead of having two sub-networks with pairs of images as input, the triplet neural network is composed of three sub-networks (see Fig. 2.10). An anchor example, a positive image from the same person as the reference, and a negative image from a different person are shown to the network as a triplet of images. The weights of the network for the three input images are shared like Siamese network. The triplet loss function is based on a relative distance rather than an absolute distance in contrastive loss. Compared to the standard neural network, the Siamese or Triplet architecture has several advantages:

- The Siamese or triplet neural network is able to learn on data pairs or triplets instead of fully labeled data. An alternate explanation is that the Siamese/Triplet neural network is suitable for weakly supervised situations when we have no access to training instance labels. The Siamese/Triplet neural network has a greater generalization ability. For the case where there is no common class between training and test set, the learned model can be easily applied to unseen classes in the training set.

- When there is a very large number of classes like tens of thousands or there are very few images per class, it is hard to train a good classifier and the Siamese/triplet neural network might be a better option.
- It is more practical to update continuously the Siamese/triplet models. When there are new classes added in the training set, we can update the Siamese/triplet model directly with new class images, which is impossible for the model trained with classification loss.

2.2 Person Re-identification

2.2.1 Datasets

In the literature, numerous of datasets for person re-identification have been released over the last years. We present the most used state-of-the-art datasets for person reidentification, categorizing them into tree classes and highlighting their challenging aspects. A summary of the common datasets is given in the Table 2.1.

Table 2.1: Person re-identification dataset overview.

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

2.2.1.1 Single-shot datasets

Datasets are qualified “single-shot” if, in gallery set in the test set of the dataset, each identity has one image for query and one true match image. The most commonly used single-shot datasets are VIPeR, GRID, PRID2011(S), CUHK01. They are relatively small, containing hundreds of identities. VIPeR, CUHK01 was captured in university campus scenes. PRID2011(S) was captured in street scenes. GRID was collected in a subway station. In VIPeR, CUHK01 and PRID2011(S), viewpoint changes are the main challenge since the images are captured under very different angles. PRID2011(s) and VIPeR have illumination changes between two cameras. GRID and PRID2011(S) have a number of distractor images (additional images that do not belong to any of the probes) in the gallery set.

2.2.1.2 Multi-shot datasets

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

2.2.1.3 Video datasets

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

2.2.2 State of the Arts

The advancement in deep learning and the immense practical applications of person re-id attracted several researchers. The development in person re-id before 2014 is mainly focused on hand-crafted methods [6]. Poriki [30] in 2003, proposed a non-parametric model based on correlation coefficient matrix, and the cross-view target matching is achieved across different cameras by obtaining the color distribution of the target. Zajdel et al. [31] in 2005, proposed for the first time of re-id in a multi-camera network. Gheissari et al. [32], in 2006 proposed for the first time of person Re-ID concept, using salient edge and color histograms. Farenzena et al. [33] in 2010, the person re-id article was published for the first time in the top tier conference named Computer Vision and Pattern Recognition (CVPR). In the year 2014, for the first time, deep learning [9] was used in the domain of person re-id. After which it is greatly boosted with the use of Convolution Neural Networks (CNNs) and other deep learning methods in designing person re-id techniques [34].

The person re-id community usually used two different kinds of CNN models. The first is the classification model used in object detection [35] and image classification [36]. The other is the Siamse model in which triplets [37] or image pairs [25] are used as input. The

detailed features of the person can be extracted using CNN due to its advantages in position, rotation, and scale invariance. A typical re-id technique can be split into two basic components, the first involved obtaining a unique person descriptor, and the other involved the comparison to decide either a match or not. The CNN model is combined with matching analysis methods for designing person re-id techniques.

In CNN-based re-id systems, the appearance of the person is influenced by many factors which include pose, illumination, resolution, etc. The method proposed in [38], where the regions are matched with the use of two filters by introducing a horizontal region matching layer. The method proposed in [39] uses a triplet loss function in which the inputs are three images. For every single image, the body parts that are overlapping after the first convolution was partitioned, then a fully connected layer is used to fuse it with the global one. After segmentation, the size of the local area becomes smaller due to the data coming from surveillance which is of low resolution. The misalignment problem is addressed in [40], using affine transformation for feature wrapping which is expensive and difficult. The inverse STN method [41] replaces this with deformable convolution but it does not have a weighted summation step to generate new feature maps and is difficult to integrate into CNN architecture.

Fig. 2.11: Network architecture milestone (2012 to present).

The recently developed popular classification networks are depicted in Fig. 2.11. Most of them such as AlexNet [25], GoogleNet [42], VGGNet [26], ResNet [27], and DenseNet

[43] are based on the design rule of LeNet [24], which results in low-resolution representation. The state of arts Re-ID techniques that are recently developed are based on the ResNet classification network such as PCB [44], Auto-ReID [45], Hi-CMD [46]. As ResNet is the most preferred classification network used by the person re-id community, so we take it as a comparison to HRNet for the person re-id technique.

In this latter, we concentrate on a unique target of seeking person Re-ID technique which consists of high-resolution representation learning and deformable convolution without any additional supervision. The proposed technique shares the same high-level spirit in comparison to the state-of-art techniques in the person re-id field and experimental results show the effectiveness of our technique. Our approach shows 1.7 % Rank-1 and 2.8 % mAP improvement in performance. The HRNet used as a backbone in our technique can replace the traditional Resnet used as the backbone in state-of-art techniques, which results in improved performance.

2.3 Deformable Convolution Module

The deformable convolution network is proposed in [19]. The author argues that the geometric structures in CNN building modules are fixed due to which they are able to model only geometric transformations. Therefore, to enhance this capability of CNN to model non-rigid objects, a deformable convolution is proposed.

2.3.1 Architecture

The deformable convolution consists of standard convolution with the addition of 2D offset to the sampling grid locations, which enabled the free form deformation in it. The preceding feature maps are used to learn offsets through additional convolutional layers.

Therefore, using the input features the deformation is trained in a dense, local, and adaptive manner, which is very useful for modeling non-rigid objects such as a person.

In offset learning, it required fewer parameters and less computation so it is a light-weight module which can easily substitute their plain counterparts in deep CNNs and using standard backpropagation it can be easily trained. The detailed implementation can be found in [19].

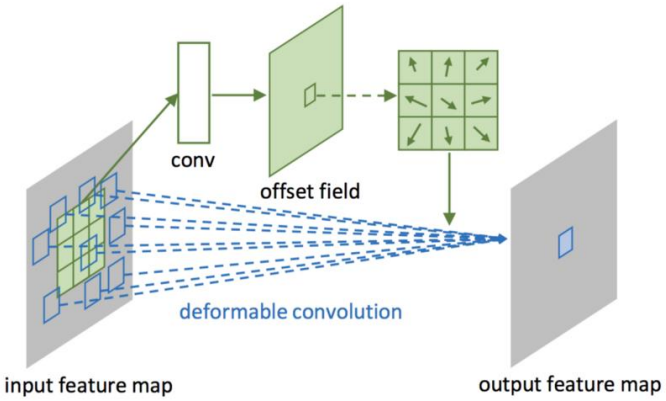

Fig. 2.12: Illustration of 3 × 3 deformable convolutions [19].

In Fig. 2.12, we use a convolutional layer over the same input feature map to retrieve the offsets. Spatial resolution of the output feature map is the same as the input feature map. The 2N-dimensional channel dimension represents N 2D offsets. As the training proceeds, both the output feature and offset learning occurs simultaneously. Bilinear operations are used to compute the gradients, and then those gradients are used to propagate the gradients backwards through the computational graph.

Deformable convolutional is made up of two parts: one convolutional layer with parameters to learn 2D offsets for each input, and another convolutional layer to be trained with offsets from the previous layer. Instead of feeding the regular convolutional layer into the green squares, the image is fed into the blue squares as shown in the diagram. In the

input feature map, you will see that the blue squares are adjusted using the arrow in the offset field.

The authors demonstrate empirically that deformable convolution is capable of "expanding" the object's receptive field. To compute the mean dilation between each offset, they use the formula “effective dilation” (i.e., the blue squares in the Fig. 2.10). This deformable filter was shown to have a bigger “receptive field” around larger item.

2.3.2 Understanding Deformable Convolution

This is based on using an offset in the convolution to enhance the spatial sampling locations while learning the offsets in relation to task requirements.

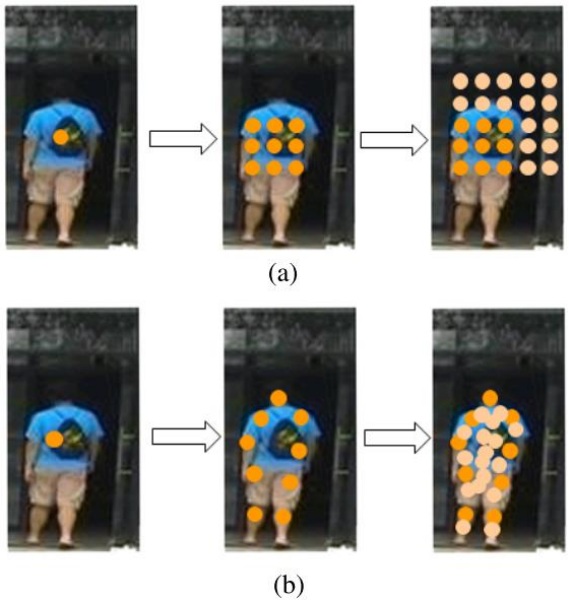

Fig. 2.13: Illustration of (a) standard convolution having the fixed receptive field and (b) deformable convolution having the adaptive receptive field.

The effect of the composited deformation is substantial when the deformable convolution module is applied. Fig. 2.13 illustrates this. The receptive field and sampling locations are fixed across the top feature map in normal convolution (up). In deformable convolution,

the offsets are adaptively modified according to the scale and shape of the objects (down). Adaptive deformation is strengthened for nonrigid objects, such as pedestrians.

CHAPTER 3

PROPOSED WORK

The experiment was carried out on a large-scale public dataset Market-1501 [17]. The performance of re-id techniques is measured using two evaluation metrics: Cumulated Matching Characteristics (CMC) and mean Average Precision (mAP). The "ID-discriminating Embedding" (IDE) [47] is used as a baseline identification model, which is effectively trained on ResNet-50 and HRNet-W44-C CNN model separately. Both basic CNN models are pre-trained on ImageNet and fined tuned for the prediction of person identities by generating a 2048d vector for every image. We follow the PyTorch implementation of both networks.

3.1 High-Resolution Network - HRNet

The High-Resolution Network (HRNet) is proposed in [18]. The proposed novel architecture HRNet keeps the high-resolution representation throughout the network. It begins from a high-resolution convolution stream then a high to low-resolution convolution stream is gradually added one after the other and the multi-resolution streams are connected in parallel.

3.1.1 HRNet Structure

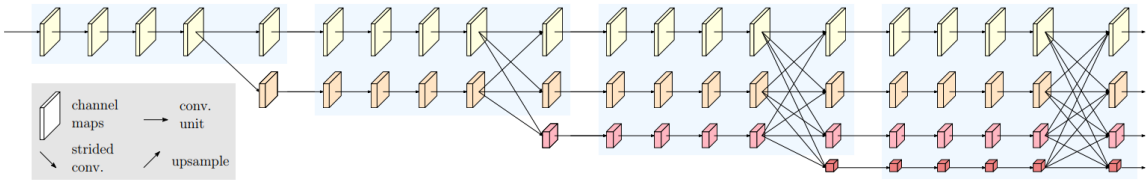

Fig. 3.1: Illustration of only the main body of a high-resolution network and the stem (two stride-2 3 × 3 convolutions) is not included [18].

It consists of a four-stage network, in which semantically strong and spatially precise high-resolution representations are learned from HRNet, due to high to low-resolution convolution streams which are connected in parallel instead of series, and the high-resolution representations are boosted by repeating the multi-resolution fusion using low-resolution representation and vice versa. The first stage consists of high-resolution convolution and the second, third and fourth repeats two, three and four resolution blocks respectively. The four-stage network structure consists of four parallel streams which can be logically described as:

$$\begin{aligned} B_{11} \rightarrow B_{21} \rightarrow B_{31} \rightarrow B_{41} \\ \searrow B_{22} \rightarrow B_{32} \rightarrow B_{42} \\ \searrow B_{33} \rightarrow B_{43} \\ \searrow B_{44} \end{aligned} \tag{3.1}$$

Where B_{sr} represents s^{th} stage sub-stream and 'r' represents resolution index of branch B. The 'r' for 1st stream is one and for others it is $\frac{1}{2^{r-1}}$ of the 1st stream.

The main body, depicted in Fig. 3.1 and described further below, is made up of many components: parallel multi-resolution convolutions, repetitive multi-resolution fusions, and the representation head depicted in Fig. 3.2.

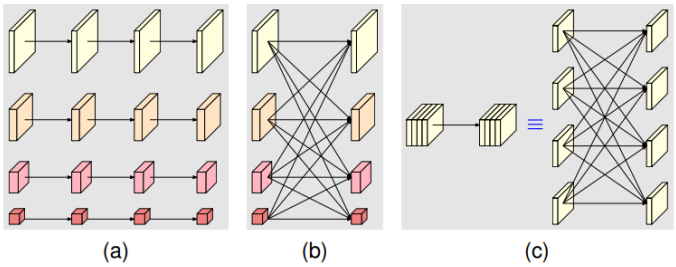

Fig. 3.2: (a) Multi-resolution parallel convolution, (b) multi-resolution fusion. (c) A normal convolution (left) is equivalent to fully connected multi-branch convolutions (right) [18].

The modularized block has two components: multi-resolution parallel convolutions as given in Fig. 3.2a and multi-resolution fusion as given in Fig. 3.2b. The group convolution is similar to multi-resolution parallel convolution. It separates the input channels into numerous subgroups of channels and executes a normal convolution over each subset at different spatial resolutions individually, whereas the resolutions are the same in the group convolution. This relationship indicates that multi-resolution parallel convolution benefits from group convolution in some way. The multi-resolution fusion unit is similar to the normal convolution's multibranch full-connection version, as seen in Figure 3.2c.

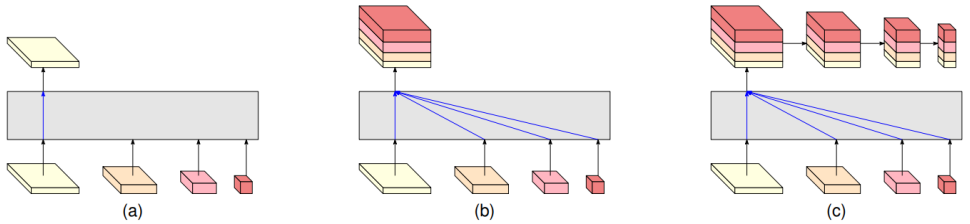

Fig. 3.3: The network in Fig 3.1 output the four-resolution representations is shown at the bottom of each sub-figure in the bottom, and the gray box represent how we use them as input to obtain the output. (a) HRNetV1 (b) HRNetV2 (c) HRNetV2p [18].

Representation Head

There are three various kinds of representation heads as depicted in Fig. 3.3 named as HRNetV1, HRNetV2, and HRNetV1p.

HRNetV1 is depicted in Fig. 3.3a. The output is a representation of the high-resolution stream using only the representation from the high-resolution stream. The remaining three representations are disregarded.

HRNetV2 is depicted in Fig. 3.3b. Rather than downscaling the low-resolution images, we use bilinear up sampling without changing the number of channels to generate higher-

resolution images, and then concatenate the four images to create a single image with the combined output. This is followed by a one-dimensional convolution to blend the four images.

HRNetV2p is depicted in Fig. 3.3c. HRNetV2 up samples the output to several levels, while we go in the opposite direction by down sampling the output from HRNetV2.

3.1.2 Classification Head

The classification head in HRNet consists of three steps as shown in Fig. 3.4. In the first step, feature maps of four resolutions pass through a bottleneck and then increased the output channel numbers to 128, 256, 512, and 1024, from C, 2C, 4C, and 8C, respectively. The C represents the number of channels. In the second step, a 2-strided 3×3 convolution is used for the high-resolution representations down sampling results in 256 channels which are added to the second high-resolution representations, and to obtain 1024 channels over the small resolution, the whole process is repeated twice. In the third and last step, a 1×1 convolution is used in the transformation of 1024 channels to 2048 channels and then a global average pooling is applied. In the end, the 2048-dimensional representation output is fed into the classifier. The detailed implementation can be found in [18].

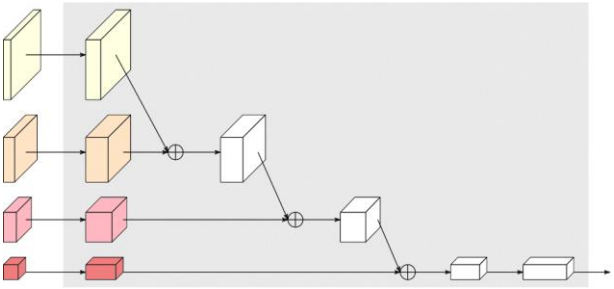

Fig. 3.4: Illustration of ImageNet classification. The representations of four resolutions are used as input of the box [18].

3.2 Dataset

The performance of our technique is evaluated on Market-1501, which is a large-scale benchmark person re-id dataset mostly used by the person re-id community. It comprises images collected by using six cameras at Tsinghua University supermarket. The DPM detector [16] is used for the automatic detection of the images. The person re-id dataset is divided into training images, verification images, query, and gallery. The dataset has 32,668 detected person rectangles with 1501 identities in total. Furthermore, the dataset has 12,936 training images with 751 identities and 19,732 testing images with 750 identities without overlapping. Detailed statistics are given in Table 3.1.

Table 3.1: Statistics of Market-1501 dataset.

Market-1501 Dataset	
Release Year	2015
Total Number of Identities	1501
Total Number of Images	32668
Number of Camera	6
Scene	Outdoor
Label Method	Hand/DPM
Evaluation Metrics	CMC/mAP
Number of Annotated Box	25259
Number of Box per ID	19.9
Number of Gallery Box	19732
Number of Distractors	2793+500k
Number of Identities (Training-Testing)	751-750
Number of Images (Training-Testing)	12936-19732
URL	www.liangzheng.org/Project/project_reid.html

3.3 Evaluation Metrics

The person re-id techniques performance is usually evaluated on two evaluation metrics CMC and mAP.

The CMC accuracy reported in this paper is Rank-1 by taking re-id as a ranking problem. The Rank-1 accuracy is represented by the probability of correct images retrieved in top-1 results. CMC curve is used to show the performance. If there are many ground truths exist in the gallery, CMC will consider only the first match. Therefore, the CMC is accurate when there is only one ground truth present.

$$cmc(N) = \sum_{n \epsilon N} r(n) \tag{3.2}$$

Where r(n) represents rank-n.

The mAP represents the performance precision and recall rate. It is accurate when there are many ground truths exist. The perfect Re-ID system is that which should return all the matches that are true. Therefore, it is used in addition to CMC for the Market-1501 dataset in which for each query many ground truths exist from different cameras.

$$AP = \frac{1}{R} \sum_{r \epsilon R} P(r) \tag{3.3}$$

$$mAP = \frac{1}{M} \sum_{m \epsilon M} AP(m) \tag{3.4}$$

Where M represents identities number in the test set, R represents recall and P represents precision.

3.4 Implementation Details

The implementation of our experiment is done in four phases, which consists of ResNet or HRNet as a backbone network and with or without deformable convolution module, respectively. The brief description of every phase is provided from section 3.4.1 onwards. The experiment is carried out using baseline identification model IDE. As there are 751 training identities in Market 1501, so in the last FC layer we set the output dimension equal to 751 and all other settings are set to default. We train the model for 60 epochs and initialized the base learning rate at 0.001 and after 40 epochs it decayed to 0.0001. In training, SoftMax loss is used. In testing, pool5 or FC layer is used to extract the feature map to learn the global descriptor for Market-1501.

3.4.1 ResNet IDE

In the first phase, we implement the IDE given in [44] following similar settings with some modification as described above. Our implemented IDE achieves comparable results with the implementation in [44]. The experiment is carried out using baseline identification model IDE consist of pre-trained ResNet-50 as a backbone network without the addition of deformable convolution layer. The overall structure can be seen in Fig. 3.5.

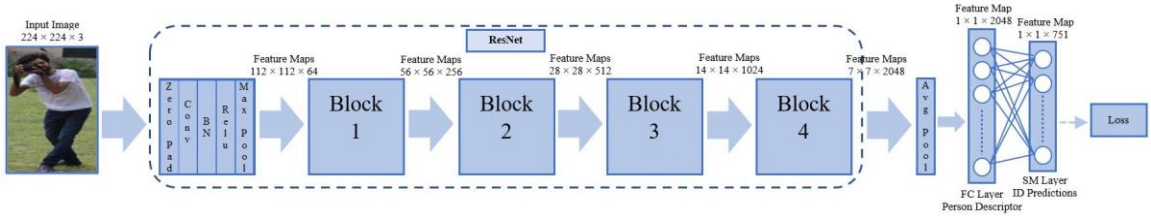

Fig. 3.5: Person re-id model using ResNet as backbone network without the addition of deformable convolution layer.

3.4.2 ResNet IDE with Deformable Convolution

In the second phase, the experiment is done with similar settings in phase one except the deformable convolution module is added between the ResNet block 5 and the average pooling layer. The deformable convolution module can be easily added to existing CNN because its input and output are similar to their plain version. The brief structure of the proposed approach is illustrated in Fig. 3.6.

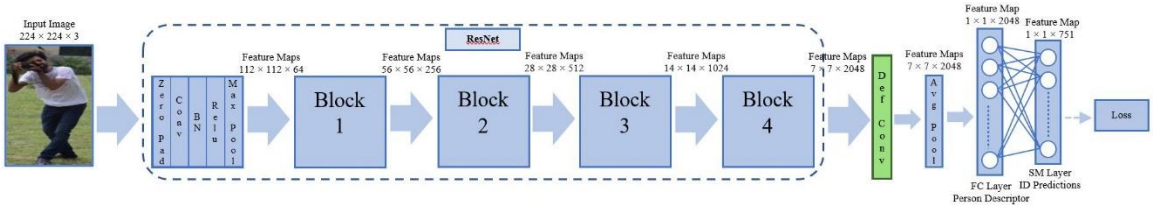

Fig. 3.6: Proposed person re-id model using ResNet as backbone network with the addition of deformable convolution layer.

3.4.3 HRNet IDE

In the third phase, the experiment is repeated following the similar setting in phase one except the backbone network in IDE is changed with a pre-trained HRNet model. The brief structure of the proposed approach is illustrated in Fig. 3.7.

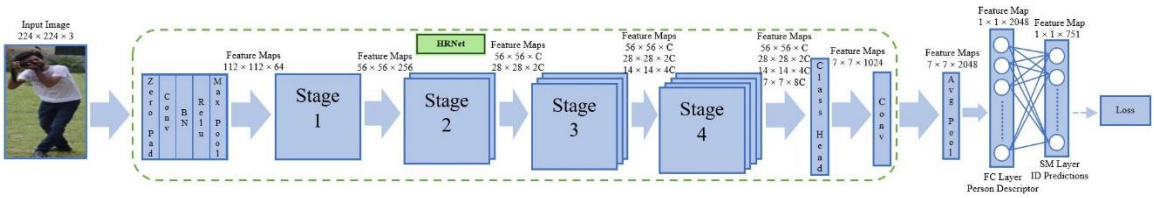

Fig. 3.7: Person re-id model using HRNet as backbone network without the addition of deformable convolution layer.

3.4.4 HRNet IDE with Deformable Convolution

In the fourth phase, the experiment is repeated following the similar setting in phase two except the backbone network in IDE is changed with a pre-trained HRNet model. The brief structure of the proposed approach is illustrated in Fig. 3.8.

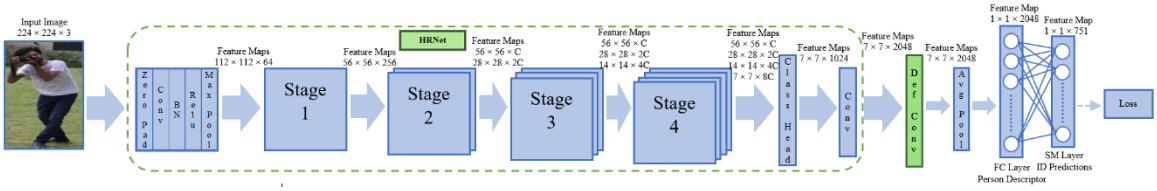

Fig. 3.8: Our proposed person re-id model using HRNet as backbone network with the addition of deformable convolution between the average pooling and the last convolution of HRNet.

The results for all four phases are reported for comparison.

CHAPTER 4

EXPERIMENTS RESULTS AND DISCUSSION

The experimental results were obtained by implementing the base and the proposed frameworks as explained in the section 3.3. The performance of the all the frameworks is evaluated using two evaluation metrics CMC and mAP. The four frameworks discussed are further divided into three categories each, which are briefly describe in following section.

4.1 Experiments on Market-1501

The IDE feature is trained on the backbone network for the evaluation of our approach on Market-1501. We consider the single query setting in all our experiments. The Euclidean distance among query image and person image in the gallery, is used to rank the images. The effectiveness of our technique is also verified on two different distance metrics, XQDA (Cross view Quadratic Discriminate Analysis) [48] and KISSME [49]. The results of our approach for four phases on the above metrics are given in Table 4.1. As we can see the Rank-1 and mAP results in table considerably improved with the addition of deformable convolution module and incorporating HRNet in baseline IDE instead of ResNet on different distance metrics. The best results are achieved when we implement the IDE with both of them in phase 4 of the experiment as shown in the Table 4.1.

The results on the above metrics show the superior performance of our proposed approach for person re-id using HRNet over the ResNet. The KISSME metric is proved to be a good partner with the IDE baseline for person re-id. We did not consider the re-ranking algorithm for simplicity, which may further enhance the results.

Table 4.1: Comparison of our approach.

CNN based Person Re-id		
Re-id Technique	***Rank-1***	***mAP***
IDE ResNet [34]	87.06%	69.13%
IDE ResNet + XQDA	87.82%	70.91%
IDE ResNet + KISSME	88.87%	72.63%
IDE ResNet + Def. Conv. [a]	88.13%	70.12%
IDE ResNet + Def. Conv. + XQDA	88.63%	72.01%
IDE ResNet + Def. Conv. + KISSME	89.96%	73.78%
IDE HRNet	88.14%	70.37%
IDE HRNet + XQDA	88.73%	71.11%
IDE HRNet + KISSME	89.72%	73.09%
IDE HRNet + Def. Conv.	88.61%	70.83%
IDE HRNet + Def. Conv. + XQDA	89.15%	72.98%
IDE HRNet + Def. Conv. + KISSME	90.57%	75.43%

[a.] Represents Deformable Convolution

4.2 Discussion

The experimental results are reported in the Table 4.1 shows the re-id technique performed well when it is implemented using HRNet, with the addition of deformable convolution module and KISSME metric. We achieved 90.57% Rank-1 accuracy and 75.43% mAP, outperforming the ResNet baseline results, which confirmed the effectiveness of our approach and will have a promising future in person re-id. The addition of deformable convolution module addresses the misalignment problem arises due to DPM in Market-1501 dataset and the use of state-of-the-art HRNet architecture address the low-resolution problem in images which considerably improved the person re-id results.

The result of three images given as a probe on the Market 1501-Dataset is shown in Fig. 4.1. The image in yellow boundary is the probe as represented in the first column. According to the similarity score the retrieved images are sorted from left to right. The

image in green boundary is correct match and the image in the red boundary is false match as represented in the Fig.4.1.

Probe Retrieval Results

Figure 4.1: Result of three images given as a probe on the Market 1501-Dataset.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this paper, a state-of-art novel architecture HRNet is utilized to design a person re-id system using the traditional baseline. We also add a deformable convolution module with HRNet for better transformation modeling capability along with high-resolution representation learning. In the experiment, we thoroughly compared our technique with the most preferred ResNet architecture used in state-of-the-arts approaches on one of the prominent and challenging re-id datasets. The results show the improved performance and confirm the effectiveness of our approach, it suggests that the HRNet would serve as a stronger backbone for the person re-ID problems.

The recent development in computer vision, playing a vital role in designing smart city applications, especially intelligent surveillance monitoring and intelligent transportation due to the popularity of the Internet of Things (IoT). These applications usually have low-resolution representation due to limited constraints. In the future, the proposed technique can be incorporated in the state-of-art person re-id model to investigate the performance of our approach and can also address some other person re-id challenges such as viewpoint variation, illumination changes, pose variations, occlusion, change of person accessory, similar clothing issue, etc.

REFERENCES

[1] Karanam, S., Yang Li and R. Radke. “Person Re-Identification with Discriminatively Trained Viewpoint Invariant Dictionaries.” 2015 IEEE International Conference on Computer Vision (ICCV) (2015): 4516-4524.

[2] X Li, X., W. Zheng, Xiaojuan Wang, Tao Xiang and S. Gong. “Multi-Scale Learning for Low-Resolution Person Re-Identification.” 2015 IEEE International Conference on Computer Vision (ICCV) (2015): 3765-3773.

[3] Huang, Yukun, Zhengjun Zha, Xueyang Fu and Wei Zhang. “Illumination-Invariant Person Re-Identification.” Proceedings of the 27th ACM International Conference on Multimedia (2019): n. pag.

[4] Huang, Houjing, Dangwei Li, Z. Zhang, Xiaotang Chen and K. Huang. “Adversarially Occluded Samples for Person Re-identification.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 5098-5107.

[5] Zhou, Sanping, F. Wang, Zeyi Huang, and Jinjun Wang. “Discriminative Feature Learning with Consistent Attention Regularization for Person Re-Identification.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 8039-8048.

[6] Yang, Fan, Ke Yan, S. Lu, Huizhu Jia, X. Xie and W. Gao. “Attention Driven Person Re-identification.” Pattern Recognit. 86 (2019): 143-155.

[7] Zheng, L., Y. Yang and A. Hauptmann. “Person Re-identification: Past, Present and Future.” ArXiv abs/1610.02984 (2016): n. pag.

[8] Ristani, Ergys and Carlo Tomasi. “Features for Multi-target Multi-camera Tracking and Re-identification.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 6036-6046.

[9] Wang, K., H. Wang, Meichen Liu, Xianglei Xing and Tian Han. “Survey on person re-identification based on deep learning.” CAAI Trans. Intell. Technol. 3 (2018): 219-227.

[10] Yi, Dong, Zhen Lei, and S. Li. “Deep Metric Learning for Practical Person Re-Identification.” ArXiv abs/1407.4979 (2014): n. pag.

[11] Xiao, Tonglin, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. “Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 1249-1258.

[12] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, “Mars: A video benchmark for large-scale person re-identification,” in European Conference on Computer Vision, 2016.

[13] S. 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, 2015, pp. 2197–2206.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[15] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, “A survey on learning to hash,” arXiv:1606.00185, 2016.

[16] Felzenszwalb, Pedro F., Ross B. Girshick, David A. McAllester and D. Ramanan. “Object Detection with Discriminatively Trained Part Based Models.” IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2009): 1627-1645.

[17] Zheng, L., L. Shen, L. Tian, S. Wang, Jingdong Wang and Qi Tian. “Scalable Person Re-identification: A Benchmark.” 2015 IEEE International Conference on Computer Vision (ICCV) (2015): 1116-1124.

[18] Wang, Jingdong, K. Sun, Tianheng Cheng, Borui Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, Mingkui Tan, Xinggang Wang, W. Liu and Bin Xiao. “Deep High-Resolution Representation Learning for Visual Recognition.” IEEE transactions on pattern analysis and machine intelligence (2020): n. pag.

[19] Dai, Jifeng, Haozhi Qi, Y. Xiong, Y. Li, Guodong Zhang, H. Hu, and Y. Wei. “Deformable Convolutional Networks.” 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 764-773.

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

[21] W. Huang, R. Hu, C. Liang, Y. Yu, Z. Wang, X. Zhong, and C. Zhang, “Camera network-based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations,” in International Conference on Multimedia Modeling. Springer, 2016, pp. 174–186.

[22] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, “Person re-identification with correspondence structure learning,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3200–3208.

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

[24] LeCun, Y., L. Bottou, Yoshua Bengio and P. Haffner. “Gradient-based learning applied to document recognition.” (1998).

[25] Radenović, Filip, Giorgos Tolias and O. Chum. “CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples.” ECCV (2016).

[26] Simonyan, K. and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” CoRR abs/1409.1556 (2015): n. pag.

[27] He, Kaiming, X. Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 770-778.

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

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

[30] Porikli, F. “Inter-camera color calibration by correlation model function.” Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429) 2 (2003): II-133.

[31] Zajdel, W., Z. Zivkovic and B. Kröse. “Keeping Track of Humans: Have I Seen This Person Before?” Proceedings of the 2005 IEEE International Conference on Robotics and Automation (2005): 2081-2086.

[32] Gheissari, N., T. Sebastian, and R. Hartley. “Person Reidentification Using Spatiotemporal Appearance.” 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) 2 (2006): 1528-1535.

[33] Farenzena, M., Loris Bazzani, A. Perina, Vittorio Murino and M. Cristani. “Person re-identification by symmetry-driven accumulation of local features.” 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010): 2360-2367.

[34] Goodfellow, Ian J., Yoshua Bengio and Aaron C. Courville. “Deep Learning.” Nature 521 (2015): 436-444.

[35] Girshick, Ross B., J. Donahue, Trevor Darrell, and J. Malik. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014): 580-587.

[36] Krizhevsky, A., Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet classification with deep convolutional neural networks.” Communications of the ACM 60 (2012): 84 - 90.

[37] Schroff, Florian, D. Kalenichenko and James Philbin. “FaceNet: A unified embedding for face recognition and clustering.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 815-823.

[38] Li, W., R. Zhao, Tonglin Xiao, and Xiaogang Wang. “DeepReID: Deep Filter Pairing Neural Network for Person Re-identification.” 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014): 152-159.

[39] Cheng, De, Y. Gong, Sanping Zhou, J. Wang and Nanning Zheng. “Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 1335-1344.

[40] Zheng, Zhedong, L. Zheng and Y. Yang. “Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro.” 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 3774-3782.

[41] Lin, Chen-Hsuan and S. Lucey. “Inverse Compositional Spatial Transformer Networks.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 2252-2260.

[42] Szegedy, Christian, W. Liu, Y. Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, D. Erhan, V. Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 1-9.

[43] Huang, Gao, Zhuang Liu, and Kilian Q. Weinberger. “Densely Connected Convolutional Networks.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 2261-2269.

[44] Sun, Y., L. Zheng, Y. Yang, Qi Tian, and S. Wang. “Beyond Part Models: Person Retrieval with Refined Part Pooling.” ArXiv abs/1711.09349 (2018): n. pag.

[45] Quan, Ruijie, Xuanyi Dong, Yuehua Wu, Linchao Zhu and Y. Yang. “Auto-ReID: Searching for a Part-Aware ConvNet for Person Re-Identification.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 3749-3758.

[46] Choi, Seokeon, Sumin Lee, Y. Kim, Tae-Kyung Kim, and Changick Kim. “Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 10254-10263.

[47] Zheng, Zhedong, L. Zheng and Y. Yang. “A Discriminatively Learned CNN Embedding for Person Reidentification.” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 (2018): 1 - 20.

[48] Liao, Shengcai, Yang Hu, Xiangyu Zhu and S. Li. “Person re-identification by Local Maximal Occurrence representation and metric learning.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 2197-2206.

[49] Köstinger, Martin, Martin Hirzer, Paul Wohlhart, P. Roth, and H. Bischof. “Large scale metric learning from equivalence constraints.” 2012 IEEE Conference on Computer Vision and Pattern Recognition (2012): 2288-2295.

ABBREVIATIONS

CNN: Convolution Neural Network

CMC: Cumulated Matching Characteristics

DNNs: Deep Neural Networks

Def Conv: Deformable Convolution

FC: Fully Connected

HRNet: High-Resolution Network

IDE: ID Discriminating Embedding

IoT: Internet of Things

MAP: Mean Average Precision

MLP: Multi-Layer Perceptron

Re-id: Re-identification

ResNet: Residual Network

ReLU: Rectified Linear Unit

TDNN: Time Delay Neural Networks