PERSON RE-IDENTIFICATION USING DENSELY CONNECTED CONVOLUTIONAL NEURAL NETWORK

A Project Report Submitted as a Part of Major Project-2

Master of Technology

in

Information Systems

by

SHRADHA JAISWAL

2K17/ISY/14

Under the Supervision of

Dr. Dinesh K. Vishwakarma
(Associate Professor – Department of Information Technology)

Department of Information Technology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Bawana Road, Delhi – 110042
June-2019

DECLARATION

We hereby declare that the Major Project-2 work entitled "**PERSON RE-IDENTIFICATION USING DENSELY CONNECTED CONVOLUTIONAL NEURAL NETWORK**" which is being submitted to Delhi Technological University, in partial fulfillment of requirements for the award of degree of Master of Technology (Information System) is a bonafide report of Major Project-2 carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

SHRADHA JAISWAL

2K17/ISY/14

CERTIFICATE

This is to certify that Major Project report-2 entitled "PERSON RE-IDENTIFICATION USING DENSELY CONNECTED CONVOLUTIONAL NEURAL NETWORK" submitted by **SHRADHA JAISWAL (Roll No. 2K17/ISY/14)** for partial fulfillment of the requirement for the award of degree Master of Technology (Information Systems) is a record of the candidate work carried out by him under my supervision.

Dr. Dinesh K. Vishwakarma

Project Guide

Associate Professor, Department of

Information Technology

Delhi Technological University

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Dr. Dinesh K. Vishwakarma for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided without which the project could not have been a success.

Secondly, I am grateful to Dr. Kapil Sharma, HOD, Department of Information Technology, DTU for her immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

SHRADHA JAISWAL

2K17/ISY/14

ABSTRACT

A successful system for person re-identification inspired by [67], a Densely Connected Convolutional Neural Networks (DenseNet) have been developed. This architecture was proposed by Huang et el. [67] (2017) for object recognition. We are using this model for exploring the network configurations and settings for getting a better solution for person re-identification tasks. We will train and test different person re-identification datasets, search for optimal settings and other factors affecting the result. In this work, we are using two different network configurations of DenseNet model i.e., DenseNet-121 and DenseNet-161 with growth rate of 32 and 48 respectively. The model is trained and tested on different RE-ID datasets which are CUHK01 [31], MARS [44], VIPeR [27] and Market-1501 [29].

Keywords: *Person Re-Identification; Deep Learning; DenseNets; Convolutional Neural Networks; Re-Identification*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

LIST OF ABBREVIATIONS

S No	Abbreviated Name	Full Name
1	FOVs	Fields-of-views
2	RNN	Recurrent Neural Network
3	CNN	Convolutional Neural Network
4	DL	Deep Learning
6	RE-ID	Re-Identification
7	SVM	Support Vector Machine
8	NN	Neural Network
9	CMC	Cumulative Matching Characteristic
10	ReLU	Rectified Linear Unit
11	BN	Batch Normalization
12	SGD	Stochastic Gradient Descent
13	mAP	Mean Average Precision

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

1.1 Person Re-Identification

Person Re-Identification is gaining immense popularity because there is a need of public safety in areas such as malls, airports, office buildings, public places, etc. The first research work on person re-id was given by Huang and Russel in 1997 [1]. This work was based on Bayesian formula for appearance prediction of a person in one camera provided some information detected in other camera views. Multiple spatio-temporal features such as height, width, colour, length of vehicle, observation time and velocity were included in the appearance model. Many improved models and methods were proposed after that, which has increased its popularity in security areas.

The security and safety of public is one of the major concern in our society and it is increasing every day. Video surveillance systems are playing an important role in accomplishing this goal. It monitors and analyses the videos which are acquired by different cameras in a camera network. Tracking of people is very important task in analysing the activity detection across that area and hence person Re-ID plays a fundamental part in multiple camera tracking and multiple camera activity detection.

Humans can identify objects and people very easily as our brains and eyes are trained in this way, but for a machine it can be a quite challenging task to train the machine in such a way so that it can recognize objects easily with lesser errors. So, the task of any re-id system should be to minimize the error and increase the accuracy while tracking and recognizing the object.

"Person Re-Identification is the process of comparing a pair of input images taken across various cameras to predict whether they are same or not." A basic Re-Identification system comprises of 3 process i.e., person detection followed by person tracking and then person identification. There are various applications of person re-id such as video surveillance systems, multiple camera tracking, forensics, robotics, etc. [2]. It is a challenging task as the look of an individual changes due to change in viewpoint, pose, different camera angles, lightning variations, etc. So, to develop a highly robust person re-id model is a challenging task.

Various part-based body models have been proposed to handle the appearance changes as the person moves in the recorded video. There are several models which

uses local and global features, extraction of these features is often a difficult task. Recurrent patterns of motion are analysed when video data is available.

Feature extraction plays an important role in re-id systems. Conventional methods use hand-craft features but as the computational power has increased and datasets have become bigger and bigger, the demand of deep learning based systems has increased. Deep learning is demanding in these fields due to its high performance based methods to achieve higher accuracy and lower error rates [3].

Cameras are deployed in large numbers in various public urban places such as shopping malls, university campus, office buildings, airports etc. which records huge amount of videos, those video data are required in analysing the person's information. Task like human activity detection and unwanted event prediction requires high level video surveillance systems to predict and analyse the information accurately.

Fig. 1.1 shows the non-overlapping fields-of-views (FOVs) [4] which are monitored by multiple cameras in a surveillance area captured from the top view of the building. Persons are shown with coloured dots and numbers corresponding to those dots are IDs that are assigned to them. The movement of a person in the camera network is represented with dotted lines. Re-Id system determines the tracking across different areas as the movement of the individuals is tracked from one camera's field-of-view to another camera's field of view.

There are various datasets that are freely available to train and test re-id models such as MARS [44], Market 1501 [29], ETHZ [26], CAVIER [32], CUHK [31,20,21], iLIDS [28], ViPER [27], MSMT [33] etc. which consists of several images of different person captured with multiple cameras in the camera networks. They are broadly classified into two parts i.e., Image based person Re-Id and Video based person Re-Id [2]. There are many research work that has been proposed on these categories which are widely used in various practical applications nowadays.

Re-Id algorithms helps in tracking the movements of the individuals in a multi-camera network, due to this reason it is widely used in practical scenarios as well as in research areas such as modeling activities, human-robot identification and physical social network mining. A re-Identification system basically consists of gallery set i.e., collection of known person and the probe i.e., unknown person, on these two sets the identification is performed.

The following two steps of person re-identification system is very important for building any RE-ID system:

i) Extraction from descriptor on the basis of appearance

ii) Calculating and matching the distance between candidate images

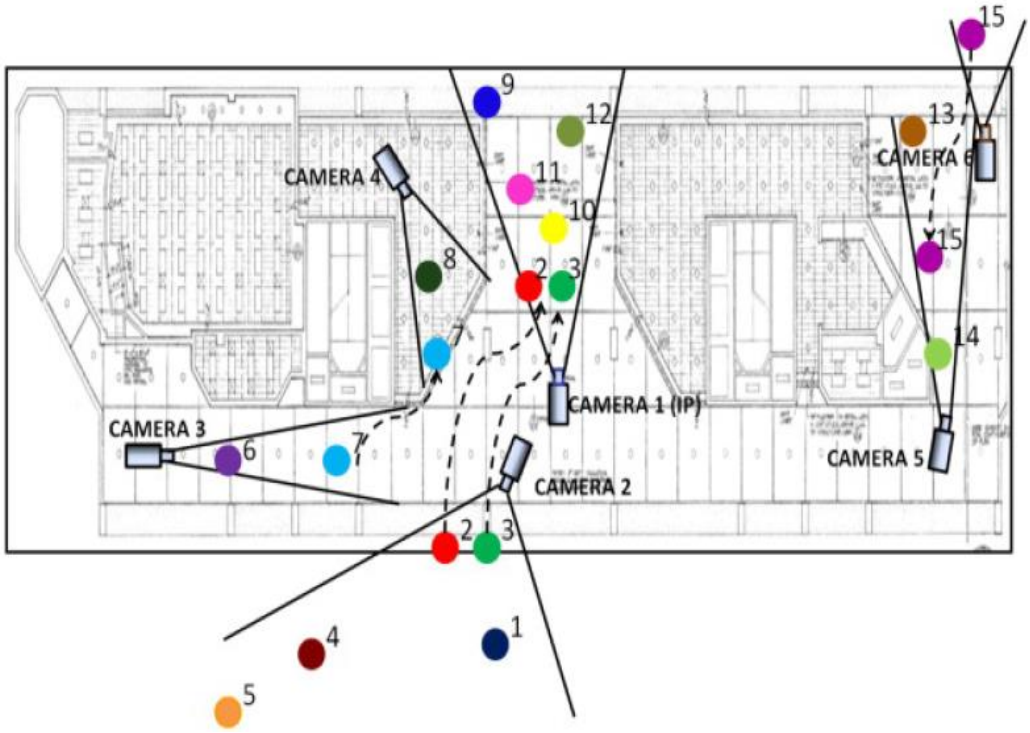

Fig.1.1 Person Re-Identification using Multiple-Camera surveillance system [4].

Two assumptions are applied on the first step i.e., information like fine cues (face or iris) cannot be retrieved properly because of the presence images having low resolution and clothes of the people remain same across different cameras. In the second step, initially a set of candidate images and a query image is given, the task is to find the minimum distance between probe image and the goal image (query image from different scene) from descriptors. Supervised or unsupervised learning helps in identifying pairwise distance metric.

1.1.1 Image based Person Re-Id

In 2006, a spatio-temporal segmentation model was given by Ghessari et al. [17] which uses video frames only for designing and matching. This work is based upon image matching on the basis of salient edge histograms and extracting colour features using Hessian-Affine interest point operator [10,11].

In video surveillance systems, which are deployed in public areas, we need to search images of the person according to its attribute description such as colour, age, gender, dress, etc. as provided by the user. If the person can be easily identified by the help of given attributes, then we can also find that person in the query attributes. But this is not an easy task as it seems to be, because there are several other factors which can affect the result such as occlusion, low resolution image, illumination changes,

lightning and pose variations.

Description of the person and distance metrics are the two important components for any RE-ID system. Firstly, a Proper description of person is very important task, colour is the most common feature that is used while extracting person's information from the captured image and foreground features are separated from background images for avoiding any miscalculation. Texture features are less frequently used. Secondly, a good distance metric learning is a challenging task because high-dimensional features cannot be fully extracted by the descriptors due to the low resolution of captured images [2]. So, there are various metric learning methods that have been proposed in various research work which basically includes supervised and unsupervised learning, local and global learning. Majorly, supervised and global distance metric learning is used.

There are various public available datasets which are used to train and test Re-Identification models such as Market-1501[29], ViPER [27], CUHK [31,20,21] and so on.

1.1.2 Video based Person Re-Id

Most of the person Re-Identification work was based only on image matching. A new work based on random selection of images for multi-shot re-id was developed by [19] 2010. It uses segmentation model and colour for the identification of the foreground features. A lot of problem occurs when the person moves between overlapping cameras, so it becomes difficult because of changing appearances of the person due to lots of environmental factors across different cameras. For still pictures, there are various models that have been developed but identification on sequence of images turn out to be more accurate and hence widely used in practical applications. Earlier, video re-identification seems to be difficult task due to lack of video re-id datasets but now there are many video datasets that are available which has helped in training complex machine learning algorithms. Using multiple frames per person has improve the performance of re-id systems as compared to single frame per person and a huge improvement in the accuracy when there is more number of selected frames.

The goal of video based Re-Id system is to determine the same person in different videos which are captured from multiple cameras. Image features are similar to image based re-id systems but the matching function will be different in case of video re-id. An important task in these systems is calculation of distance and metric learning while matching the videos. Various works of video re-id are based on

appearance models using multiple shots and also adding temporal cues. Because of temporal information in videos adds more details of the person which will help to clearly understand things so it is a challenging task compared to image-based re-id and it will also improve the accuracy but it also increases the complexity in designing the algorithm for such system [18]. So, RNN models are better in case of video Re-Id systems, as compared to Deep CNN models for Image based Re-Id Systems.

Recurrent Neural Networks (RNN) helps in processing the large time-series in a video using neural networks. Using recurrent connections helps in increasing the efficiency and accuracy of video based re-id systems by passing information between time-steps.

Computing the distance between query video and a candidate video is the main purpose of video re-id systems. These two videos id contain the same person, then the computed distance will turn out to be small, otherwise, not. Practical applications like video surveillance system, video Re-Id is gaining more popularity due to its better performance.

There are various publically available datasets to train and test video re-id models such as ETHZ [26], MARS [44], iLIDS-VID [28] and so on.

1.2 Basic Architecture of Person RE-ID System

Person Re-ID is the process of identifying a person in different views taken by several cameras in a camera network under changing lightning conditions and at different time intervals. The RE-ID process makes an attempt to find the person in various views and the system should be capable enough to differentiate between various humans in the same view.

Re-Identification systems generally depends on a probe image and set of images in gallery known as gallery set. A gallery set comprises of individual or sequence of images acquired from various cameras in a camera network. The image which is to be matched and identified from the gallery set is called as probe image. Gallery set can be available online or offline. In offline set, people first register themselves to enter any public space while in online set gallery is updated automatically as individual takes entry or exits the place [4]. During runtime, in the detection stage people are tracked and identified from the images captured in the camera network.

Bounding boxes of person's images are created automatically or manually by using pedestrian detection algorithms or various background elimination methods for person detection phase. Analysis is performed on single frames (single-shot) or

multiple frames (multi-shot) by extracting features like motion, shape, colour, texture, etc. from persons bounding boxes. These obtained features are then classified from the gallery set by the help of certain classifiers such as SVM [12], Random forest, Naïve Bayes, and so on. Classifiers can be simple or complex supervised or semi-supervised methods for classifying the persons features obtained from the gallery set.

There are basically two standard techniques that the researchers follow in this area that is, descriptor generation and similarity computation. There are various approaches for appearance based descriptors such as Color-histogram-based descriptors, Interest-point-based descriptors, Covariance-based descriptors, Textual-based descriptors and so on [16].

For similarity computation, computation of similarity scores is done between probe image from the gallery images. The various distance metric methods and nearest neighbour approaches are used to calculate the similarity score. A good metric learning algorithm helps in increasing overall performance of any re-identification system [4].

RE-ID system is categorised in three different stages, firstly detection a person, secondly tracking them in camera and thirdly identifying the person. All of these components will be discussed in brief in the section 2.2.

RE-ID system also requires a labelled dataset for evaluating and analysing the performance. The various image and video person RE-ID datasets will be discussed in the section 2.4.1. A good dataset must be very challenging so that proper testing and analysis of the system can be performed, highlighting the weakness and bottlenecks.

1.3 Person Re-Identification using Deep Learning

Deep Learning is a sub-category of machine learning which is based on how human brain works. Deep Learning algorithms are based on how human brain instantly learn new features and recognizes objects. Conventional algorithms were slower as compared to deep learning algorithms so they are less used in today's applications. Deep Learning algorithms uses larger labelled datasets and requires large computational power systems such as GPU based systems. With the help of GPU based systems, complex and time taking computations can be performed with less efforts which also saves a lot of time as compared to conventional methods. On the larger datasets, for training a neural network conventional methods on CPU will take years to train while deep learning methods will do all the computations in a week on

GPU based system.

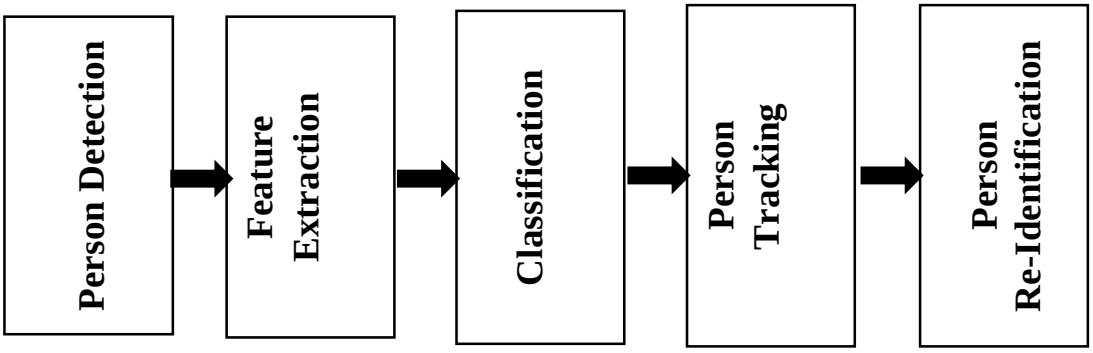

Fig: 1.2 Person re-id flow of process.

Deep learning methods work better than pre-Deep learning methods also called Traditional methods due to its adaptability and learning capabilities, which were not possible before. Deep Learning models have gained a lot of attention in various fields such as automated driving, aerospace, defence systems, medical researches, automation industries, video surveillance systems and so on.

Deep Learning models have achieved greater performances in state-of-the-arts methods, even better than human capabilities. Classification process is performed on images, videos and text in deeply learned systems. The deep learning models are trained on deeper and deeper architectures of neural networks, then tested on larger datasets. Neural network consists of input layer, initially the data is fed on this layer and finally the output is predicted, while in deep neural network, between input and output layers there is more number of hidden layers which helps to predict better results as compared to shallow neural networks [4]. With the increase in number of hidden layers, the architecture will be complex, resulting in overfitting of data and more computation time.

To avoid this drawback, there are various methods such as weight decay, sparsity, adding dropouts so that it removes some units in particular intervals and applying some data augmentation techniques such as mirroring/rotating, cropping due to which size of small training datasets can be increased. There are various parameters such as initial weight, rate of learning, optimizers, size of the neural network, batching, etc. will affect the computation time of the neural network. So deciding these parameters is a challenging task, there is no proper algorithm for that, it's a try and fail method. Fixing a value to parameters and checking its accuracy until we get the required results. After training the model on labelled data by setting the required parameters, testing is done on unlabelled data for checking the model's efficiency.

The purpose of any deep learning model is to learn accurate feature representation and evaluating those features at various measures [4]. Basically, there are two types of deep learning models that have been developed in this research field:

- Classification model for Person Re-ID
- A triplet/pair comparison based Siamese model

Yi et al. [20] and Li et al. [21] proposed a Siamese NN to test whether both the images belongs to the same person or not which basically helped in gaining huge success for image classification in this field. Deep Learning hence became a popular choice in person RE-ID and computer vision fields because of its better performance, efficiency and improved accuracy as compared to traditional re-id models.

For image-based re-id, CNN based deep learning models are used, basically classification model for classification of images and detection of objects, and Siamese model using pair of images. Currently, Siamese model is more popular and mostly used method for image based re-identification problems. In 2014, [21] proposed a "Deep Filter pairing neural network for person re-identification" which uses patching layers which multiplies the responses of convolution of an image pair. Later in 2015, [8] improved the earlier proposed Siamese model by calculating the cross-neighbourhood difference features from one input image to another image.

Wu et al. [23] proposed a "PersonNet" in 2016, which uses small sized convolution filters. Varior et al. [45] proposed an LSTM architecture based on Siamese networks for person e-id. There are various architectures that have been proposed after that with improved accuracy and lower error rates which are tested on various image datasets, and research is still going on.

For video-based re-id, as discussed earlier, RNN models seems to be the best choice. The datasets are larger in comparision to image re-id datasets due to each tracklet consists of so many frames and hence requires more efforts in analysing datasets in a proper way. In 2016, Zhang et al. [6] used frames in training samples for training the model with softmax loss. McLaughlin et al. [14] used features which are extracted by successive frames in video passed through a Convolutional Neural Network model and recurrent layer as final layer. Varior et al. [45] proposed a model which uses LSTMs connected to a softmax layer by extracting low-level hand-crafted features. A hybrid DL framework was proposed by Wu et al. [23] which extracts spatio-temporal features from a video. There are so many researches that are going on which uses spatio-temporal features in a video based person re-identification models.

There are lots of work based on deep learning methods that have been proposed

which are based on new approaches or some improvement in the existing approaches. Still it suffers from lack of training data, because of this reason Siamese model is preferred. Deep learning techniques for person re-id is briefly discussed in later sections.

1.4 Motivation and Problem Formulation

The work conducted in thesis is motivated by the fact of improving re-identification accuracy and obtain solutions for fully automated and computer based person RE-ID systems.

A successful system for person re-identification inspired by [67], a Densely Connected Convolutional Neural Networks (DenseNet) have been developed. This architecture was proposed by Huang et el. [67] (2017) for object recognition. We are using this model for exploring the network configurations and settings for getting a better solution for person re-identification tasks. We will train and test different person re-identification datasets, search for optimal settings and other factors affecting the result. In this work, we are using two different network configurations of DenseNet model i.e., DenseNet-121 and DenseNet-161 with growth rate of 32 and 48 respectively. The model is trained and tested on different RE-ID datasets which are CUHK01 [31], MARS [44], VIPeR [27] and Market-1501 [29].

CHAPTER 2
LITERATURE REVIEW

This chapter includes the detailed description of techniques of deep learning (DL), person re-id components, challenges and its evaluation algorithms are discussed.

2.1 Person Re-Id Challenges

Person re-id system has faced numerous challenges such as individual's changing movement in the camera, occlusions, lightning variations, changing poses, and different changing clothes could also challenge the re-id systems. Low resolution cameras can also affect the quality of re-id systems as it will not be able to extract high-dimensional features like eyes, face recognition, biometric features, etc. Robust and detailed description of an image is required which is basically a challenging task as there are various factors that are uncontrollable such as resolution, imaging conditions, imaging angles, frame rate and so on. False and incorrect detections may lead to bad quality re-id system. Therefore, identifying a unique description automatically is an extremely difficult task [13].

Another problem could be creating a labelled dataset for evaluating and analysing the performance of re-id systems [3]. Capturing the videos and manually labelling the appearances of the individuals is an expensive process.

2.2 Components of Re-Identification Systems

There can be found many components of person re-id systems as shown below in Fig. 1.2 such as person tracking, person detection, feature extraction etc.

2.2.1 Pedestrian Detection

For any RE-ID system, first and foremost thing is to detect the person in a camera network by using Part-based detectors that models connection of a person's parts of body, monolithic detectors which maps single descriptor to single detector window. The first detection algorithm was proposed by Boult et al. [47] which detects every person in an image manually but the problem was to get perfect and clean detections. So later on an automatic person detection algorithm was designed Taiana

et al. [51] in 2013 which analyses RE-ID algorithms with the non-perfect detections as input to them. Many algorithms have been proposed after that.

Nowadays an in-camera tracker is used to detect the person in the camera, even in the multi-shot situation, the algorithm is able to detect persons, also allowing more feature extraction, eliminating background noise and automatically selecting the cleanest image.

2.2.2 Features and Feature Extraction

Selective and persistent features are needed to be extracted for any re-identification systems which will be able to differentiate every person from other persons in an efficient way. Features can be manually chosen/designed or can be combined with multiple feature channels in some pattern or selecting the best feature from each test sample and then use those features for the classification. The most used feature for RE-ID system includes Histogram of Oriented Gradients, colors so on are some of the mostly used color feature extraction algorithms. Maximum Response Filter Bank (MR8) [55,56], Local Binary Patterns (LBP) [52], Gabor filters [57] and Schmidt [58] filters are used to extract texture features from an image because in some cases texture features are more convenient than color features.

For edge detection, Gabor filters [57] are mostly used. A dynamic feature selection method was given by Liu et al. [59] using the feature type for every type of clothes of a person and it works best for these types of features. For encoding local features, a Fisher vector turn out to a better model. DL models are based on extracting local features, global features or combination of both the features.

Earlier, hand craft features like color, shape, edge etc. (low level features) were used but nowadays high level features are extracted by CNN models resulting in higher performance compared to traditional methods. One of the broad surplus of exploiting DL models on person re-id systems is that the DL models extracts features automatically in convolution layers.

Various works have been proposed for feature extraction such as Pictorial Structures (PS) body-part detectors given by Andriluka et al. (2009) [60], Girshick et al. [61] proposed body-part detectors with grammar models in 2011.

2.2.3 Classification

The task of classification depends upon labelled datasets so that the neural network will be able to correlate between labels and data. Classification in person RE-ID algorithms is based on identifying the persons in images from the gallery set after feature extraction and person detection in an image. There are various algorithms proposed using neural networks for classification tasks such as SVM [7], K Nearest Neighbour, Learning Vector Quantization, Random Forests, and so on. Classification can also be performed using direct distance minimization or simply learning [62,64,65].

Zheng et al. [63] illustrated a different metric called the distance metric method that minimizes a distance called inter-class distance(ICD) and tend to maximizes the ICD by calculating the metric in such a way that for a given triplet image which contains a pair of image of one person and one image of distinct human.

2.2.4 Tracking and Re-Identification

After the person is detected and classified, then the next step is tracking and re identifying that person in the camera networks. A pedestrian cannot be present in different cameras at the same instance of time, so tracking step will help in ignoring or even correcting the mistakes during classification step. The overall efficiency and performance of the system increases if we combine the classification and tracking stages but the errors should be taken care of [66]. Person Re-Identification is the last step which is to be performed for matching the pair of image from the gallery set. At this step the main problem is that the detection and classification stages requires a good object descriptor and the models of learning step should be recognized by different cameras of the same person.

2.3 Techniques of Deep learning for person RE-ID

DL techniques are very popular among researchers for person re-id tasks for its high-performance with high accuracy. Many works of DL are fundamentally works on improving previous models or designing a new model. There are various types of deep learning models that have developed but in this section we will focus on the two popular models i.e. a classification model and a Siamese model [4]. DL models also face a lot of challenges because of lack of training data and few of the datasets like

VIPeR [27] contains only two images per person that is very challenging with respect of person re-id.

2.3.1 Classification Models

Classification based models require to compute the probability of the class to whom the image of a person belongs to. Fig. 2.1 shows the classification model based on convolutional neural networks where an input image of a person is fed to the network and probability is calculated for the person's corresponding class [4]. Classification models fails for re-identification problems where datasets contain less images per person because of overfitting.

A feature fusion deep neural network was proposed by Wu et al. [23] in 2016 which regularizes CNN features from hand crafted features. An image of 224 x 224 x 3 is passed as input to the network, standard descriptor was used to extract the hand craft features. Those features are passed to next steps that is to the fully connected layer then a buffer layer which are generally described as fusion layer. Then what passes to the fully connected layers is Softmax layer for minimizing the cost and outputs the required deep feature representation.

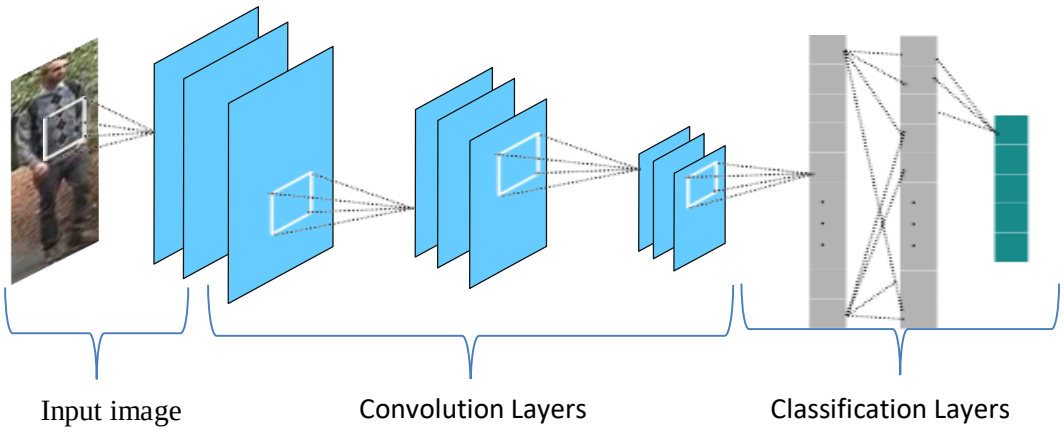

Fig. 2.1 A standard Classification model example [4].

By using CNNs and multiple datasets, a deep feature representation model was given by Xiao et al. [24] for effective training of neurons in the layers. For managing the problem caused due to misalignment and pose variations, Su et al. [49] worked Convolutional Neural Network model for re-id. This model has the capability to learn adaptive measurements of similarity and effectual representation of features from pedestrian's body-parts.

2.3.2 Siamese Models

The broadly used network by researchers is Siamese Network (SN) for person re-id tasks, this network consists of a set of two or more similar sub-networks: like pairwise when there are two sub-networks or triplet when three sub-networks included.

Pairwise models: Zhang et al. [6] illustrated that a model in which a pair of images is fed in the form of input to the SN which then connects to the first convolution layer and for measuring the similarity between input images, where linear SVM is opted.

A deep filter pairing neural network was proposed by Li et al. [21], which encodes photo-metric transformation in different camera views. An example of pairwise Siamese model is shown in Fig. 2.2. A multiple scale model which can learn discriminant feature representation in various resolutions across multiple scales was proposed by Qian et al. [9].

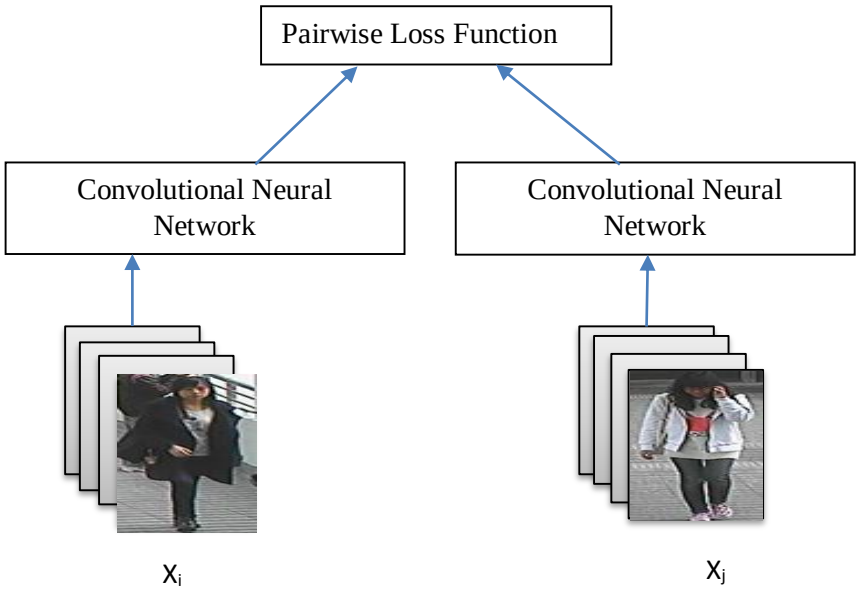

Fig. 2.2 A pairwise Siamese model.

Triplet models: A multi-scale triplet model composed by adding one deep CNN and two shallow NN was given by Lieu et al. [68] which shares the parameter between them. Deep CNN contains five convolutional layers, five max-pooling layers, two normalization fields upfronted by three fully-connected layers while both shallow NN have two convolutional and two pooling layers. Then output of both networks are combined with an embedding layer for the final feature representation. Lin et al. [36] worked on DL techniques that includes the entire camera network, also uses forward and backward propagation for computing and updating the features of the used CNN

model. Fig. 2.3 shows an example of triplet Siamese neural network.

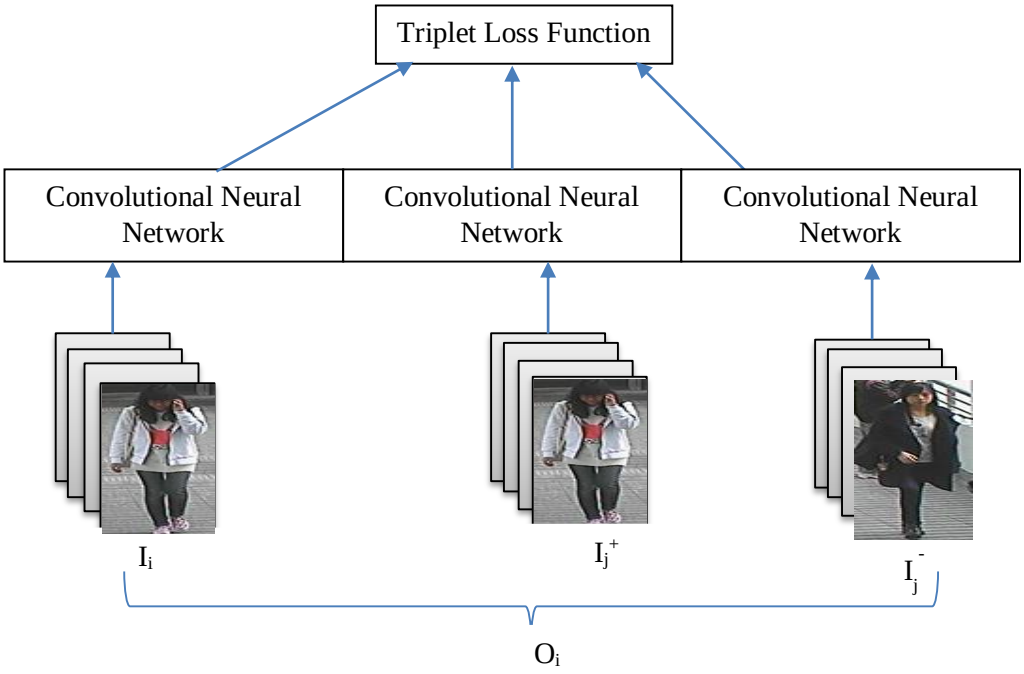

Fig. 2.3 A triplet Siamese model.

2.3.3 Loss Functions

A loss function associates cost with some event by mapping some values to a single real number. Neural networks use loss functions to minimize the cost associated with those values [4].

Below are some of the commonly used loss functions for pairwise and triplet models:

Pairwise loss function:

- Cosine similarity loss: For positive pairs, it maximizes the cosine values and decreases the angle between them. For negative pairs, it minimizes the cosine value.

$$I(\mathrm{x}_1, \mathrm{x}_2, \mathrm{y}) = \begin{cases} max(0, \cos(\mathrm{x}_1 - \mathrm{x}_2) - \mathrm{m}) & if\ y = 1 \\ 1 - cos(\mathrm{x}_1, \mathrm{x}_2) & if\ y = -1 \end{cases} \quad (2.1)$$

- Hinge loss: It is also called max-margin classification. The loss function's output comes out to be zero once the distance of the positive pairs have the more impact than negative pair.

$$I(\mathrm{x}_1, \mathrm{x}_2, \mathrm{y}) = \begin{cases} \|\mathrm{x}_1 - \mathrm{x}_2\| & if\ y = 1 \\ max(0, \mathrm{m} - \|\mathrm{x}_1 - \mathrm{x}_2\|) & if\ y = -1 \end{cases} \quad (2.2)$$

Triplet loss function:

- Euclidean distance: It is used as the distance metric of the triplet loss function. Difference of matched and mismatched pair of a single triplet is calculated by:

$$d(W, I_i) = \|F_W(I_i) - F_W(I_i^+)\|^2 - \|F_W(I_i) - F_W(I_i^-)\|^2 \quad (2.3)$$

2.4 Person RE-ID algorithm evaluation

2.4.1 Datasets

Datasets have a significant role to play in the evaluation where the improvement of the RE-ID algorithms carried out. Annotation of datasets should be properly done for the evaluation process. Most of the datasets of person re-id contains information about identity to the person, location on the camera network. The dataset should comprise of cropped images of persons with their specifications. The detailed description of the most commonly datasets such as ETHZ [26], VIPeR [27], CUHK [20,21,31], Market1501 [29] and so on for re-id is provided in section that is the up next.

- **PRID2011 [69]:** This dataset was created by Austrian Institute of Technology; the images of the pedestrians are captured in two camera views from single shot. Both cameras captured the images of 200 pedestrians. Camera A or camera B captured images captured images of 700 peoples separately. There are at least 5 or more cropped images per person.

- **VIPeR [27]:** In the year of 2008, the most challenging dataset was known as VIPer. Viewpoints, lighting conditions, pose and illumination changes etc. are some of the challenges proposed by this dataset. There are images of 632 people snaped by a pair of camera views, and each individual has set of pictures which are cropped to 128 x 48 pixels.

- **i-LIDS [28]:** The images in this dataset is captured from NOC (Non overlapping-cameras) on the airport. It contains 476 images having 119 individuals with strong

occlusions, lightning and pose changes. There are 2 images for every person and 4 images for some person.

- **CAVIER [70]:** The images in this dataset is captured by pair of camera angles on shopping malls with the crossover field with 90 degree angle. There are total images of 72 persons in which 50 people shows up in both the scenes and 22 peole shows up in only one scene. There are 5 pictures of a person per-view with lightning, pose and illumination changes.

- **ETHZ [26]:** There have 3 video sequence dataset in this category wiht two movable cameras on a crowded view. There are three sequences of every person in which every image is of different sizes. There are 83, 35 and 28 persons in sequence 1, 2 and 3 respectively. Images shows changes in illumination, occlusion and lightning variations.

- **Market-1501 [29]:** It is the huge dataset for person re-id. Market-1501 contains 32,643 fully annotated class of 1501 pedestrians. Deformable Part Model (DPM) is something that crop the images of the pedestrians which is clicked by six different cameras.

- **CUHK:** Chinese University of Hong Kong (CUHK) provided three different partitions with definitive setup for person re-identification task. There are three partitions of this dataset i.e. CUHK01 [31], CUHK02 [20] and CUHK03 [21] with specific structure. **CUHK01[31]** dataset consists of 3,884 pictures of 971 persons clicked by 2 different cameras and angles. Camera I has pose and viewpoints variations, and Camera B covers front and back views. There are two images per person. **CUHK02 [20]** dataset comprises of images of 1,816 persons by five pairs of cameras. Every pair has 971,306,107,193 and 239 pedestrians accordingly. There are pair of images per person in every angle view camera. When camera views are different in training and testing then this dataset will be used. **CUHK03 [21]** dataset consists of 13,164 images of 1,360 persons. Images are captured by six different surveillance cameras and has 4.8 pictures on an average for every person in every view by the disjoint cameras which is two in numbers. Every image is cropped manually which shows changes in pose, misalignment, missing body parts and illumination.

- **MARS [44]:** This is also the huge sequence-based datasets on video re-id which consists of images of 1,261 persons. Each image of a person is captured by minimum two cameras. The dataset includes 1,191,003 bounding boxes and 20,478 tractlets.

Table 2.1: Illustration of some currently available datasets for person RE-ID [54]

Datasets	Year	#Camera	#Persons	Label	Total Images	Crop Image Size
ViPER	2007	2	632	Hand	1264	128 x 48
ETHZ	2007	1	148	Hand	8580	Vary
GRID	2009	8	250	Hand	1275	Vary
iLIDS	2009	2	119	Hand	476	Vary
PRID 2011	2011	2	200	Hand	24541	128 x 64
3DPES	2011	8	200	Hand	1011	Vary
CAVIER	2011	2	72	Hand	1220	Vary
WARD	2012	3	70	Hand	4786	128 x 48
CUHK01	2012	2	971	Hand	3884	160 x 60
CUHK02	2013	10	1,816	Hand	7264	160 x 60
CUHK03	2014	2	1,467	Hand/DPM	13164	Vary
iLIDS-VID	2014	2	300	Hand	42495	Vary
RAiD	2014	4	43	Hand	6920	128 x 64
Market 1501	2015	6	32,668	Hand/DPM	32217	128 x 64
MARS	2016	6	1,261	DPM & GMMCP	1191003	256 x 128
DukeMTMC -reID	2017	8	1812	Hand	36441	Vary
DukeMTMC 4REID	2017	8	1852	Doppia	46261	Vary
MSMT17	2018	15	4101	Faster RCNN	126441	Vary
RPIfield	2018	12	112	ACF	601581	Vary

- **WARD [71]:** This dataset includes 4,786 pictures on 70 individuals that are on surveillance area with 3 non overlapping shot cameras with changes in pose, resolution and lightning variations.

- **RAiD [72]:** The extended version of RAID is Re-Identification Across Indoor-outdoor Dataset that consists 6,920 and 43 bounding boxes of persons captured by 4 different angles. Images show large variations in illumination and poses due to indoor and outdoor conditions.

- **MSMT [33]:** This dataset is presented in 2018 for person re-identification having the data where there were 126411 pictures of 4101 pedestrian taken by three outdoor and 12 indoor cameras, with changes in illumination, scale, pose and lightning variations.

CHAPTER 3
THE PROPOSED WORK

3.1 Objective

The proposed work is inspired from the work of [67], a Densely Connected Convolutional Neural Network (DenseNet) which was developed for object recognition. To use this model for person re-identification task, we are interested primarily on exploring the network configurations and settings of this architecture that would help to get a good solution for the RE-ID problem. DenseNet architecture was developed for object recognition problems, we will use this architecture to train and test the person re-identification problem and, search for optimal settings and other factors that will affect the solution. The model is trained and tested on 4 Person RE-ID datasets i.e. MARS [44], CUHK01 [31], VIPeR [27] and Market1501 [29]. Satisfactory results have been achieved in comparison to various state-of-the-arts methods. Evaluation is done on mAP score and rank-k (where k=1,5,10,20) recognition rates.

3.2 Densely Connected Convolutional Neural Networks (DenseNet)

Convolutional Neural Networks have been very popular DL approach for person RE-ID tasks. CNN architecture was introduced over 20 years ago, several improvements were made year after year for increasing the performance and accuracy of the model, also lowering the error rates and minimizing the computation time of the model. Initially LeNet5 [15] was proposed which consisted of 5 layers, followed by VGG [5] networks consist of 19 layers, Highway networks [30] and ResNets [25] (Residual Networks) have exceeded the 100 layers. The DenseNet architecture was proposed by Huang et al. [67] in 2017. In this architecture, in a feed-forward fashion every layer is connected to each other and passes its feature maps to all successive layers. In ResNets [25], features are never combined by summation before passing into further layers, instead features are combined by concatenating them. In DenseNets [67] for a L-layer network it has L(L+1)/2 connection, while there are only L connections in conventional architectures. Because of this, DenseNets needs less number of parameters than conventional architectures. The layers required in DenseNets are very narrow, so small set of feature map is needed for the “collective

knowledge" rest of the feature maps are not changed. So the final classifier takes all the feature-maps into consideration and makes a final decision.

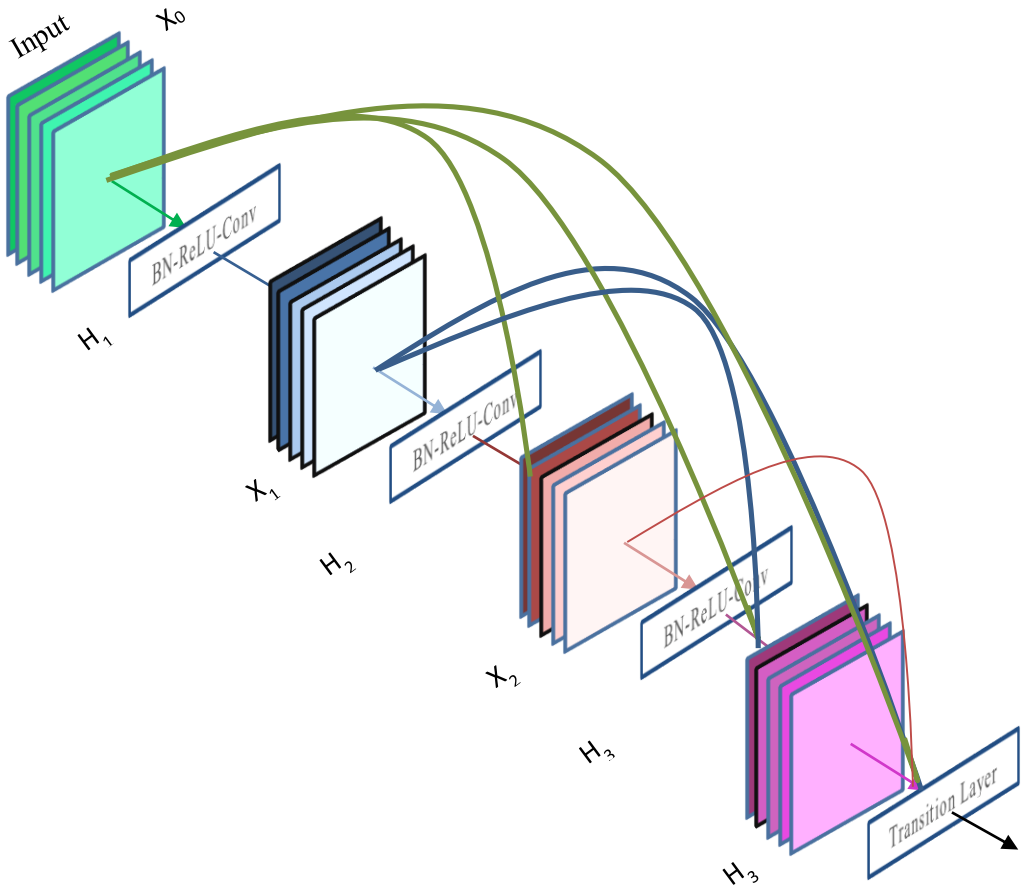

Fig. 3.1 A 4-layer dense block where each layer takes input as previous features.

There are several advantages of DenseNets such as improved parameter efficiency, for easier training the flow of gradients and information flow in whole network is better in this architecture. Signal from original input to the gradients and loss functions are directly accessible from every layer which helps in training the network deeper. It is also seen that overfitting has reduced while training smaller sets due to the regularizing effects of the dense connections. Fig.3.1 shows a 4-layer dense block where previous layers feed input to the next successive and connected layers for k=4 i.e. growth rate.

Consider an image x_0 which is fed to CNN. In an L layer convolutional network $H_l(.)$ implements a non-linear transformation where l denotes indexes for each layer. Composite function of operations such as rectified linear units (ReLU), Pooling, Convolution (Conv) or Batch Normalization (BN), is denoted as $H_l(.)$. Output of the l^{th} layer is denoted by x_l.

$x_0, x_l, \ldots, x_{l-1}$ are the inputs fed to the network and from the previous layers the l^{th} layer gets the feature maps.

$$x_1 = H_1([x_0, x_l, \ldots, x_{l-1}]) \quad (3.1)$$

where $[x_0, x_l, \ldots, x_{l-1}]$ specifies the concatenation between feature-maps of 0,…..,l-1 layers. This architecture is named as Densely Connected Convolutional Neural Network because of its dense connectivity between the layers in the network.

Composite function is denoted by H_l (.) which includes the following three process in order: BN, proceeded by ReLU and a 3 x 3 Conv layer. The network is divided into dense blocks, after every dense block a transition layer (performs Conv + Pooling) is added, which comprise of BN layer, followed by 1 x 1 Conv layer and 2 x 2 avg. pooling layer.

Growth rate is defined as follows: Growth rate of the network is defined by the hyper-parameter 'k'. Very narrow layers can be found in DenseNets e.g. for k=12, this differentiates DenseNets with other existing architectures [67].

3.3 Network Configurations

There two types of network configuration were used i.e., DenseNet-121 with growth rate, k=32 and Densenet-161 with the growth rate, k=48. Both the architecture uses 7 x 7 convolution layer and stride=2 as the first layer to the network with the output size of 112 x 112. Secondly, 3 x 3 max pooling and stride=2 is applied with the output size of 56 x 56. Both the layers are common in DenseNet-121 [67] and DenseNet-161 [67]. The difference in both the architectures is dense block configurations. DenseNet-121 uses block configuration of (6,12,24,16) whereas DenseNet-161 uses block configuration of (6,12,36,24). We have used 4 dense blocks for each architecture in our thesis work. Every dense block consists of convolution layers of sizes 1 x 1 and 3 x 3 followed by a transition layer with 1 x 1 convolution and 2 x 2 average pooling with stride=2. After 4^{th} dense block, there is a classification layer which consists of 7 x 7 global average pooling and 1000D FC layer followed by softmax function. Within a network, layer to layer replication is not required by the network unlike conventional networks because global state can be accessed everywhere in this architecture. The Network configurations of the Densely Connected Convolutional Neural Network (DenseNet-121 and DenseNet-161) that was used throughout this work are summarised in Table 2. The sequence of three continuous operations is represented by "conv" layer mentioned in the table i.e., Batch

Normalization followed by ReLU activation function and convolution operation.

Table 3.1 Configurations of DenseNet-121 and DenseNet-161 [67].

Layers	**Output Size**	**DenseNet-121 (k=32)**	**DenseNet-161 (k=48)**
Convolution	112×112	7×7conv, stride=2	
Pooling	56×56	3×3max pool, stride=2	
Dense Block (1)	56×56	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 6$	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 6$
Transition Layer (1)	56×56	1×1conv	
	28×28	2×2average pool, stride=2	
Dense Block (2)	28×28	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 12$	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 12$
Transition Layer (2)	28×28	1×1conv	
	14×14	2×2average pool, stride=2	
Dense Block (3)	14×14	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 24$	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 36$
Transition Layer (3)	14×14	1×1 conv	
	7×7	2×2average pool, stride=2	
Dense Block (4)	7×7	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 16$	$\begin{pmatrix} 1 \times 1\ conv \\ 3 \times 3\ conv \end{pmatrix} \times 24$
Classification Layer	1×1	7×7 global average pool	
		1000D fully-connected, softmax	

The distance between feature vectors is calculated using Euclidean distance and tested on several datasets for determining the best network configuration performance with several changing hyper parameters. Various other configurations were tested, in this thesis we presented the best performance configurations.

3.4 Evaluation Metrics

Cumulative Matching Characteristic curve (CMC) is the standard metric for evaluation of person RE-ID performance. ‘n’ denotes the total number of templates used for testing any RE-ID model. Probability of recognizing a correct individual in first ‘r’ ranks, where r=1,2, 3, n is called as CMC curve. As ‘r’ increases, CMC curve also increases and for r=n it will be 1. Basically, CMC curve shows how frequently the correct id is present in the best ‘r’ matches for each probe in the gallery set on average. In our thesis, we have shown Rank-1, Rank-5, Rank-10 and Rank-20 accuracy of the datasets for the evaluation of model’s performance.

3.5 Training Configurations

The following configurations were used for training on the image and video RE-ID datasets:

- NVIDIA TITAN RTX GPU’s were used for training.
- ImageNet pre-trained weights were used.
- Every step uses forward and backward propagation of every epoch depending upon the batch size and hardware configuration. Batch size ranges from 8 to 128.
- Cross dataset evaluation is also performed on some datasets which will be discussed later section.
- Training time for each epoch varies from 20 secs for smaller datasets such as VIPeR [27] to 20 min for larger datasets like MARS [1].

Other hyper-parameters/training configurations like learning rate, optimizers, batch size and so on were chosen in such a way that testing and validation loss should be minimised. Following are the configurations given for the above mentioned hyper-parameters:

- Several optimization algorithms such as Adam optimizer and stochastic gradient descent (SGD) algorithm were used but Adam optimizer resulted in better rank accuracies and minimised loss than SGD in this case.

- Different learning rates were randomly tested, ranging from 0.0003 to 0.03. Very large or very small learning rates turn out to be inefficient for training the used models.
- For testing one network configuration, number of epochs used in training was set in the range of 15 to 80.
- After every 20 epochs, learning rate is decreased by 10%. If there is no improvement after 20 epochs, then training is stopped early and different learning rate is set for training. No dropouts were used.
- Batch sizes used were 8,16,32 and 128.

In general, in most of the cases Adam optimizer turn out be fast converging than Stochastic Gradient Descent algorithm, usually learning stops after 20 to 30 epochs and optimal batch size for all input-settings is 32.

Random sampler is used for training i.e.; images are randomly selected for training. The advantage of using shuffled dataset is improvement of accuracy and model learns better.

3.6 Testing Procedure

Model is tested on the image dataset as well as video dataset. The process of testing is mentioned below:

- The testing was performed on query and gallery set of their respective datasets.
- Cross dataset evaluation [53] was also performed i.e., trained on different datasets and tested on different dataset.
- Euclidean distance is calculated between probe image to every image from gallery set, if the distance is minimum or close to zero then the both the images are identical otherwise not. Distance metric has higher impact on similarity calculation and accuracies.
- The percentage of correctly identified pairs in the given testing pairs is said to be the accuracy of the model.

- All the settings yielded a satisfactory result on testing on query and gallery set of same dataset, and some better results when tested with different datasets as well.

CHAPTER 4

EXPERIMENTAL WORK AND RESULTS

With Densely Connected Convolutional Neural Networks [67], we are primarily interested in exploring what configurations or settings of the architecture would lead to a satisfactory solution for problems related to re-id. We are trying to solve the problem of re-id by applying DenseNets, search for the optimal settings, and other factors that will affect the solution. Various experiments have been performed on challenging datasets to show the efficiency of our problem. Both image and video datasets are used for evaluation, image datasets that have been used in this experiment are CUHK01 [31], Market-1501[29] and VIPeR [27], video dataset that is used is MARS [44]. Rank recognition rates (1,5,10,20) are calculated on these datasets that is Cumulative Matching Characteristics curves. For measuring the quality of the image classification, Mean Average Precision scores is computed in all recall levels.

4.1 Setup and Datasets

The ***CUHK01*** [31] dataset consists of 3,884 pictures of 971 individuals with 2 images of every individual captured by two different cameras where each image is manually cropped to 160 x 60 pixels, and poses variations in pose and illumination. In our experiments, the probe samples are selected randomly. 485 identities are defined for training set where 1,940 images and for testing 486 identities are used. Batch size of 16 is used for both the architectures i.e., DenseNet-121 and DenseNet-161 [67]. Table 3 shows the mAP scores and, CMC with Rank-1, 5, 10 and 20 accuracies. Results in bold shows the highest accuracy achieved on that dataset between both the models.

Fig.4.1 Samples of image pairs of CUHK01.

VIPeR [27] dataset have total of 1,264 pictures of 632 persons having 2 pictures each person captured by two cameras in different views, A and B. The images are manually cropped to 128 x 48 pixels with variations in viewpoint and illumination. For training there were used 316 identities and 632 images which are randomly split and for testing other 316 identities were used. Batch size of 8 is used for both the architectures.

Fig.4.2 Sample images from VIPeR.

Dataset ***Market-1501 [29]*** includes a total of 32,668 pictures of 1,501 persons clicked by 6 distinct cameras which are configured manually. 751 identities and 12,936 pictures are intendent for training where 19,732 pictures of 750 identities exploited in terms of testing with a fixed split size. For both the networks we used Batch size of 64 (densenet-121 and densenet-161).

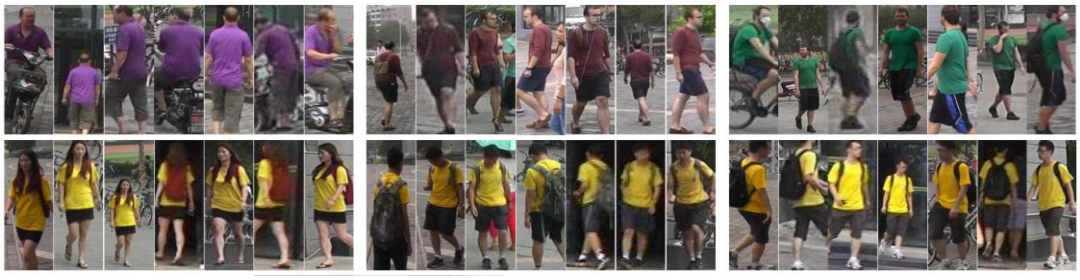

Fig.4.3 Sample images from Market-1501 dataset.

MARS [44] contains 1,191,003 bounding boxes with 20,478 tracklets captured by six different cameras which has video sequences of 1,261 pedestrians. In our experiment, 625 identities and 8298 tracklets are used for training, 626 identities are used for testing. Batch size of 8 i.e., number of tracklets and sequence length of 15 i.e., number of images in each tracklet is used. Average pooling method is used for

evaluation. (Batch size x Sequence length) gives the total images in each batch.

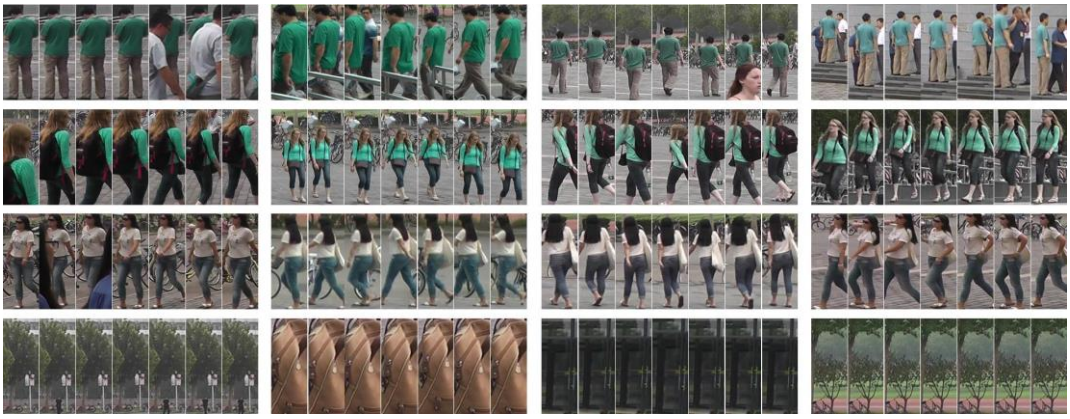

Fig. 4.4 Sample image sequences from MARS dataset.

4.2 Performance Evaluation:

Following setup is fixed for every dataset for evaluation of performance:

- The learning rate used for evaluation for all the datasets is initialised with 0.0003 and divided by 10 after every 20 epochs.
- Adam optimizer is used for all datasets.
- Single step scheduler with the step size of 20 and Random sampler was used for training.
- Each image in every dataset is resized to 256 x 128 pixels.
- Softmax loss function is used for calculation of the loss.
- Evaluation is performed after every 10 epochs.
- Growth rate of 32 and 48 is used for DenseNet-121 and DenseNet-161 model respectively [67].
- Euclidean distance metric was used for computing distance matrix between query and gallery set.

The Market-1501 dataset is trained for 30 epochs and achieved 80.9% Rank-1 accuracy on DenseNet-121 model, and 81.0 % Rank-1 accuracy on DenseNet-161 model.

CUHK01 dataset was trained for 60 epochs on DenseNet-121 model which achieved Rank-1 accuracy of 71.4 % and trained for 40 epochs on DenseNet-161 model which resulted in slightly higher Rank-1 accuracy than DenseNet-121 model i.e., 75.6%.

The performance of VIPeR dataset is evaluated on DenseNet-121 model by training for 60 epochs with Rank-1 accuracy of 25.9% and on Densenet-161 model, trained for 40 epochs and achieved Rank-1 accuracy of 36.1% shown in table 4.1.

Table 4.1: Recognition rates comparison with respect to our dataset and models.

Training Set + Model	**Testing**	**Rank1 (%)**	**Rank5 (%)**	**Rank10 (%)**	**Rank20 (%)**	**mAP score**
CUHK01 + densenet-121	CUHK01	71.4	88.2	91.5	95.6	71.6
CUHK01 + densenet-161	CUHK01	**75.6**	**89.7**	**93.2**	**96.2**	**75.5**
Market1501 + densenet-121	Market1501	80.9	**91.4**	**94.2**	96.1	61.4
Market1501 + densenet-161	Market1501	**81.0**	90.8	93.9	**96.3**	**61.7**
VIPeR + densenet-121	VIPeR	25.9	48.4	58.9	65.2	36.9
VIPeR + densenet-161	VIPeR	**36.1**	**56.6**	**67.7**	**75.6**	**46.4**
Market1501 + CUHK01 + densenet-121	VIPeR	27.2	40.8	49.1	56.0	34.8
Market1501 + CUHK01 + densenet-161	VIPeR	29.7	41.5	50.0	58.5	36.2
Mars + densenet-121	Mars	**76.8**	**89.9**	**92.9**	**95.0**	**67.3**
Mars + densenet-161	Mars	74.7	88.9	91.8	94.5	65.4

Cross dataset evaluation is also carried out on VIPeR as this dataset is considered to be challenging. Two datasets i.e., Market1501 and CUHK01 were trained together combined for 30 epochs and 60 epochs on DenseNet-121 and DenseNet-161 respectively. Training set consists of 14,876 images and 1236 training ids. The performance was tested on VIPeR and the accuracy of Rank-1 obtained onto the models was 27.2%.

Table 4.2: Single query comparison with market1501.

Methods	mAP(%)	Rank-1(%)
GoogLeNet [38]	48.24	70.27
VGG16Net [35]	38.27	65.02
ResNet50 [25]	51.48	73.69
DLCNN [40]	47.45	70.16
LSTM SCNN [45]	35.31	61.60
Gated SCNN [41]	39.55	65.88
MSCAN [42]	57.53	80.31
P2S [43]	44.27	70.72
CADL [36]	55.58	80.85
Ours(densenet-121)	61.40	80.90
Ours(densenet-161)	**61.70**	**81.0**

Table 4.3: Single query comparison with CUHK01.

Methodology	Rank-1(%)
FPNN [21]	20.65
Deep CNN [8]	65.0
CNN based on multi-model [39]	53.70
CNN with multiple domains [24]	66.60
Deep CNN based on Spatio-temporal features [34]	38.28
CNN based structured Laplacian embedding [50]	70.09
Ours(densenet-121)	71.40
Ours(densenet-161)	**75.60**

On MARS dataset, training was done for 20 epochs and 15 epochs on DenseNet-121 and DenseNet-161 respectively. The rank-1 accuracies achieved as 76.8% and 74.7% on DenseNet-121 and DenseNet-161 respectively. Comparison of CMC curves on rank-1 recognition rate on different datasets is shown in the Fig. 4.5. Table 4.2, 4.3 and 4.4 shows the comparison of rank-1 accuracy with various outstanding methods.

Table 4.4: Single query comparison with VIPeR.

Methods	Rank-1(%)
DeepReID [37]	19.90
Deep CNN with SVM [46]	12.50
CNN based 3D pyramidal NN [48]	18.04
Deep CNN [8]	34.81
Ours(densenet-121)	25.90
Market1501 + CUHK01 + densenet-121(Ours)	27.20
Market1501 + CUHK01 + densenet-161(Ours)	29.70
Ours(densenet-161)	**36.10**

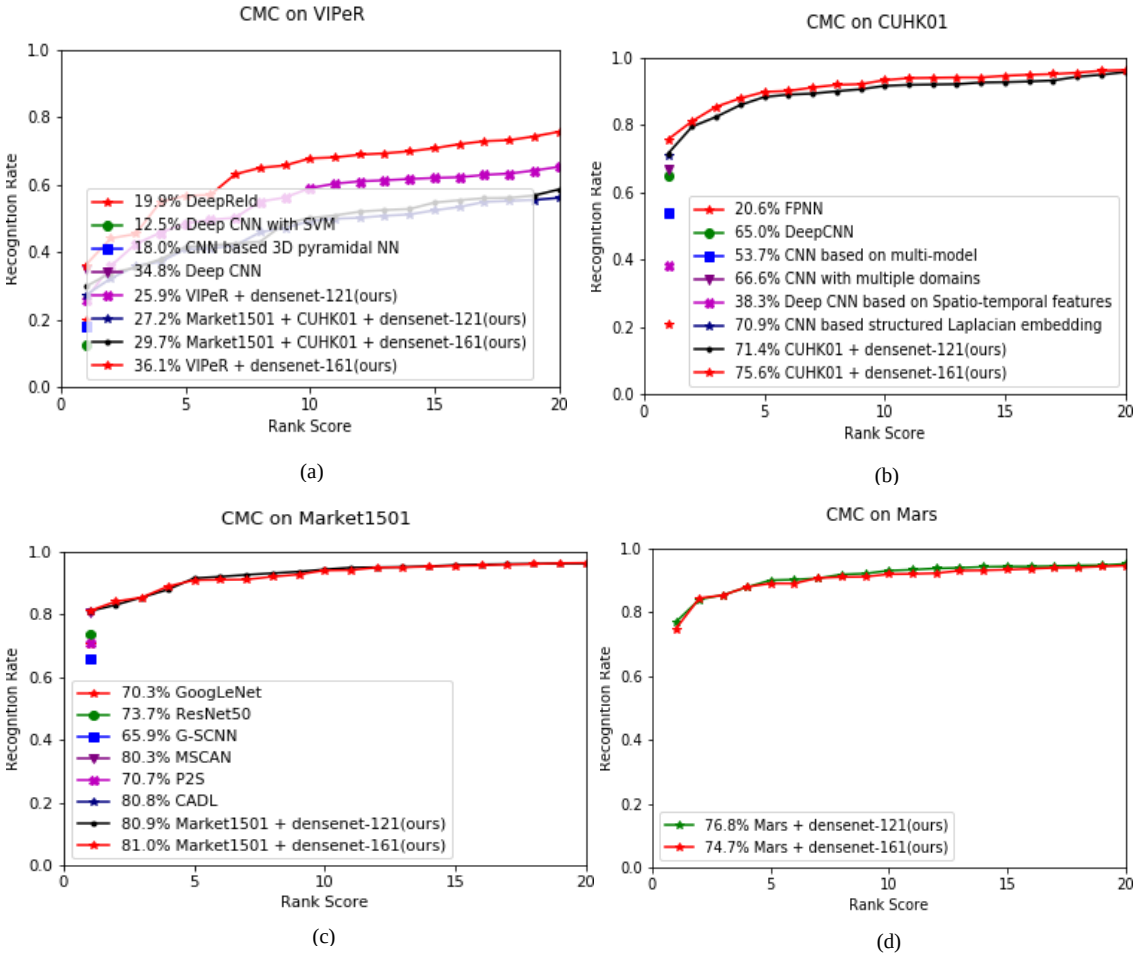

Fig. 4.5 CMC is compared with Rank-1 recognition on (a) VIPeR, (b) CUHK01, (c) Market1501 and (d) Mars datasets.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this thesis, we have proposed the use of Densely Connected Convolutional Neural Network (DenseNet) in order to solve the problem of person re-id which involves using the two different network configurations of DenseNet model i.e., DenseNet-121 and DenseNet-161 [67]. Even though the problem of person RE-ID is different to conventional person re-identification tasks in their problem domains, but because of insufficient information presents on the image it is considered as equally challenging.

Various person re-identification datasets were trained on both these models and yielded a positive and satisfactory results, which suggests that the DenseNets can be a potential solution to this challenging problem. For achieving state-of-the-arts performances, DenseNets require less computation and fewer parameters without performance degradation/overfitting.

5.2 Future Work

Person re-identifications are relatively new fields so there are many aspects that still needed to be explored and employed in real-world scenarios. We have experimented with various possible combinations of hyper parameters on both the DenseNet configurations, and computed their effect on the performance of the model, but there are various other possible configurations and combinations exists to be experimented which can help in achieving much better accuracies by detailed tuning of hyper parameters and learning rate schedules. For several tasks of computer vision based on deep learning, DenseNets may turn out to be a good feature extractor and in future it can be used for transferring features with DenseNets. There are several attempts that have been proposed in this regard, but still there exist many open issues that is required to be solved before any successful implementation of a real-time RE-ID systems.

References

[1] T. Huang and S. Russell, “Object identification in a Bayesian context,” in IJCAI, pp. 1276–1282, 1997.

[2] L. Zheng, Y. Yang and A. G. Hauptmann, “Person re identification: past, present and future,” arXiv preprint arXiv:1610.02984, 2016.

[3] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," Image and Vision Computing, pp. 270-286, 2014.

[4] B. Lavi, M. F. Serj and I. Ullah, "Survey on Deep Learning Techniques for Person Re-Identification Task," arXiv preprint arXiv:1807.05284, 2018.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., “Imagenet large scale visual recognition challenge”, International journal of computer vision, pp.211-252, 2015.

[6] Z. Zhang, T. Si and S. Liu, "Integration Convolutional Neural Network for Person Re-Identification in Camera Networks," in IEEE Access, vol. 6, pp. 36887-36896, 2018.

[7] M. Xu, Z. Tang, Y. Yao, L. Yao, H. Liu and J. Xu, "Deep Learning for Person Reidentification Using Support Vector Machines," Advances in Multimedia, pp. 1-12, 2017.

[8] E. Ahmed, M. Jones and T. K. Marks, "An improved deep learning architecture for person re-identification," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3908-3916, 2015.

[9] X. Qian, Y. Fu, Y. Jiang, T. Xiang and X. Xue, "Multi-scale Deep Learning Architectures for Person Re-identification," IEEE International Conference on Computer Vision (ICCV), pp. 5409-5418, 2017.

[10] A. Schumann and R. Stiefelhagen, "Person Re-identification by Deep Learning Attribute-Complementary Information," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1435-1443, 2017.

[11] Y. Zhou, W. Zheng, A. Wu, H. Yu, H. Wan and J. Lai, "Learning a Semantically Discriminative Joint Space for Attribute Based Person Re-identification," arXiv preprint arXiv:1712.01493, 2017.

[12] B. Hadjkacem, W. Ayedi and M. Abid, "A comparison between person re-identification approaches," Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), pp. 1-4, 2016.

[13] S. Gong, M. Cristani, C. C. Loy and T. M. Hospedales, "The re-identification challenge," in Person re-identification, pp. 1-20, 2014.

[14] N. McLaughlin, J. M. del Rincon and P. Miller, “Person re-identification using deep convnets with multi-task learning”, IEEE Transactions on Circuits and Systems for Video Technology, pp. 525-539, 2016.

[15] Y.LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86, pp. 2278-2324, 1998.

[16] X. Wang and R. Zhao, "Person re-identification: System design and evaluation overview," in Person Re-Identification, pp. 351-370, 2014.

[17] N. Gheissari, T. B. Sebastian and R. Hartley, “Person reidentification using spatio-temporal appearance,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), vol. 2, pp. 1528–1535, 2006.

[18] L. Bazzani, M. Cristani, A. Perina, M. Farenzena and V. Murino, “Multiple-shot person re-identification by hpe signature,” International Conference on Pattern Recognition (ICPR), pp. 1413–1416, 2010.

[19] M. Farenzena, L. Bazzani, A. Perina, V. Murino and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367, 2010.

[20] D. Yi, Z. Lei, S. Liao and S. Z. Li, "Deep Metric Learning for Person Re-identification," International Conference on Pattern Recognition, pp. 34-39, 2014.

[21] W. Li, R. Zhao, T. Xiao and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159, 2014.

[22] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Sackinger and R. Shah, “Signature verification using a Siamese time delay neural network,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 7, no. 04, pp. 669–688, 1993.

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

[24] T. Xiao, H. Li, W. Ouyang and X. Wang, “Learning deep feature representations with domain guided dropout for person re-identification,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 1249–1258, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition.” In CVPR, pp. 770-778, 2016.

[26] A. Ess, B. Leibe and L. Van Gool, “Depth and appearance for mobile scene analysis,” in IEEE International Conference on Computer Vision, pp. 1–8, 2007.

[27] D. Gray, S. Brennan and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, no. 5, pp. 41-49, 2007.

[28] T. Wang, S. Gong, X. Zhu and S. Wang, “Person re-identification by video ranking,” in European Conference on Computer Vision, pp. 688–703, 2014.

[29] L. Zheng, L. Shen, L. Tian, S.Wang, J.Wang and Q. Tian, “Scalable person re-identification: A benchmark,” IEEE International Conference on Computer Vision, pp. 1116–1124, 2105.

[30] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Training very deep networks.” In NIPS, pp. 2377-2385, 2015.

[31] W. Li, R. Zhao and X. Wang, “Human reidentification with transferred metric learning,” in Asian Conference on Computer Vision, pp. 31–44, 2012.

[32] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani and V. Murino, "Custom pictorial structures for re-identification," In Bmvc, vol. 1, no. 2, pp. 1-11. 2011.

[33] L. Wei, S. Zhang, W. Gao and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," IEEE Conference on Computer Vision and Pattern Recognition, pp. 79-88, 2018.

[34] S. Wang, C. Zhang, L. Duan, L. Wang, S. Wu and L. Chen, "Person re-identification based on deep spatio-temporal features and transfer learning," International Joint Conference on Neural Networks (IJCNN), pp. 1660-1665, 2016.

[35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2014, pp. 1–14.

[36] J. Lin, L. Ren, J. Lu, J. Feng and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5771-5780, 2017.

[37] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 152–159.

[38] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 1–9.

[39] D. Cheng, Y. Gong, S. Zhou, J. Wang and N. Zheng, “Person reidentification by multi-channel parts-based CNN with improved triplet loss function,” International Conference on Pattern Recognition, pp. 1335–1344, 2016.

[40] Z. Zheng, L. Zheng, and Y. Yang, “A discriminatively learned CNN embedding for person reidentification,” ACM Trans. Multimedia Comput., Commun. Appl., vol. 14, no. 1, p. 13, 2017.

[41] R. R. Varior, M. Haloi, and G. Wang, “Gated Siamese convolutional neural network architecture for human re-identification,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland, pp. 791–808, 2016.

[42] D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 384–393, 2017.

[43] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, “Point to set similarity based deep feature learning for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5028–5037, 2017.

[44] 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," European Conference on Computer Vision, pp. 868-884, 2016.

[45] R. R. Varior, B. Shuai, J. Lu, D. Xu and G. Wang, "A siamese long short-term memory architecture for human re-identification," in European Conference on Computer Vision, pp. 135-153, 2016.

[46] G. Zhang, J. Kato, Y. Wang and K. Mase, "People re-identification using deep convolutional neural network," International Conference on Computer Vision Theory and Applications (VISAPP), pp. 216-223, 2014.

[47] T.E. Boult, R.J. Micheals, Xiang Gao, and M. Eckmann. “Into the woods: visual surveillance of noncooperative and camouflaged targets in complex outdoor settings.” Proceedings of the IEEE, pp. 1382-1402, 2001.

[48] S. Iodice, A. Petrosino and I. Ullah, "Strict pyramidal deep architectures for person re-identification," in International Workshop on Neural Networks, pp. 179-186, 2015.

[49] C. Su, J. Li, S. Zhang, J. Xing, W. Gao and Q. Tian, "Pose-driven deep convolutional model for person re-identification," IEEE International Conference on Computer Vision (ICCV), pp. 3980-3989, 2017.

[50] D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann and N. Zheng, "Deep Feature Learning via Structured Graph Laplacian Embedding for Person Re-Identification," Pattern Recognition, pp. 94-104, 2018.

[51] M. Taiana, J. Nascimento and A. Bernardino. “An improved labelling for the INRIA person data set for pedestrian detection.” IbPRIA, pp. 286-295, 2013.

[52] T. Ahonen, A. Hadid, and M. Pietikainen. “Face description with local binary patterns: Application to face recognition”. Pattern Analysis and Machine Intelligence, IEEE Transactions on, pp. 2037 –2041, 2006.

[53] S. Rodionov, A. Potapov, H. Latapie, E. Fenoglio and M. Peterson. "Improving Deep Models of Person Re-identification for Cross-Dataset Usage." In IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 75-84, 2018.

[54] S. Jaiswal and D. Vishwakarma. "State-of-the-Arts Person Re-Identification Using Deep Learning." In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 238-243, 2019.

[55] T. Leung and J. Malik. “Representing and recognizing the visual appearance of materials using three-dimensional textons.” Int. J. Comput. Vision, pp. 29–44, 2001.

[56] C. Schmid. “Constructing models for content-based image retrieval”. In Computer Vision and Pattern Recognition. Proceedings of the 2001 IEEE Computer Society Conference on, p. 39-45, 2001.

[57] I. Fogel and D. Sagi. “Gabor filters as texture discriminator. Biological Cybernetics.”, pp. 103–113, 1989.

[58] C. Schmid. “Constructing models for content-based image retrieval”. In Computer Vision and Pattern Recognition, pp. 2-39, 2001.

[59] C. Liu, S. Gong, C. Change Loy and X. Lin. “Person Re-identification: What Features Are Important?”, In ECCV, pp.391-401, 2012.

[60] M. Andriluka, S. Roth and B. Schiele. “Pictorial Structures Revisited: People Detection and Articulated Pose Estimation”. Computer Vision and Pattern Recognition, pp. 1014-1021, 2009.

[61] R. Girshick, P. Felzenszwalb, and D. McAllester. “Object detection with grammar models”. PAMI, pp. 442-450, 2011.

[62] Z. Lin and L. Davis. “Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance”. In International symposium on visual computing, pp. 23-34, 2008.

[63] W. Zheng, S. Gong and T. Xiang. “Re-identication by Relative Distance Comparison”. PAMI, pp. 653-668, 2012.

[64] A. Yilmaz, O. Javed and M. Shah. “Object tracking: A survey”, pp. 13, 2006.

[65] K. Okuma, A. Taleghani, N. De Freitas, J.J Little, D.G Lowe. “A boosted particle filter: Multitarget detection and tracking”. In European conference on computer vision, pp. 28-39, 2004.

[66] A. Gilbert and R. Bowden. “Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity”. Computer Vision–ECCV, pp. 125–136, 2006.

[67] G. Huang, Z. Liu, L.V. Maaten and K. Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.

[68] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling and T. Mei. “Multi-scale triplet cnn for person reidentification”. pp. 192–196, 2016.

[69] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof. “Person re-identification by descriptive and discriminative classification”. In: Image Analysis, pp. 91-102, 2011.

[70] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino. “Custom pictorial structures for reidentification”. In: BMVC, p. 6,2011.

[71] N. Martinel, C. Micheloni. “Re-identify people in wide area camera network.” in: Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 31-36, 2012.

[72] A. Das, A. Chakraborty and A. K. Roy-Chowdhury. “Consistent re-identification in a camera network”. pp. 330–345, 2014.