Two Stream Deep CNN-RNN Attentive Pooling Architecture for Video Based Person Re-Identification

By
Wajeeha Ansar
2016-NUST-MS-CS-06
00000171500

Supervisor
Dr. Muhammad Moazam Fraz
Department of Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST), Islamabad, Pakistan.

(June 2018)

Approval

It is certified that the contents and form of the thesis entitled "**Two Stream Deep CNN-RNN Attentive pooling architecture for Video Based Person Re-Identification**" submitted by **Wajeeha Ansar** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Muhammad Moazam Fraz**

Signature: _____

Date: _____

Committee Member 1: Dr. Shahzad Saleem

Signature: _____

Date: _____

Committee Member 2: **Dr. Asad Waqar Malik**

Signature: _____

Date: _____

Committee Member 3: **Dr. Muhammad Shahzad**

Signature: _____

Date: _____

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Wajeeha Ansar**

Signature: _____

Acknowledgment

Thanks to **ALLAH Almighty**, for giving me an opportunity to increase my experience and skill levels through different phases of thesis. It's because of His blessings that I am able to complete my dissertation successfully. I would also like to give my sincere appreciation to my supervisor **Dr. Muhammad Moazam Fraz** for his excellent guidance and support. Without his guidance and motivation the completion of this dissertation was not possible. I would also like to thank my committee members Dr. Muhammad Shahzad, Dr. Shahzad Saleem and Dr. Asad Waqar Malik for their guidance and support.

I am also thankful to National University of Sciences and Technology (NUST) for providing me a great atmosphere. I am honored to be a student of School of Electrical Engineering and Computer Sciences (SEECS) where I was able to learn new things and meet highly skilled professors that added to my knowledge.

Last but not least, I'd like to thank **my parents** who have provided me every comfort of life and their utmost affection and support made me reach where I am today.

Certificate for Submission of Thesis – PG Students

It is certified that I have received 01x bound copy along with 01x CD containing soft copy of thesis work submitted by following student: -

Name of Student:	Wajeeha Ansar
Class:	MS (CS) – 6
Registration Number:	00000171500
Name of Supervisor:	Dr. Muhammad Moazam Fraz

_____ _____

Supervisor Assistant Librarian

_____

Mgr ACB (PG)

Proposed Certificate for Plagiarism

It is certified that MS Thesis Titled **Person Re-Identification based Visual Surveillance System** by **Wajeeha Ansar** has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.
b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.
c. There is no fabrication of data or results which have been compiled/analyzed.
d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

Dr. Muhammad Moazam Fraz

Signature: _____

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by **Wajeeha Ansar**, (Registration No **00000171500**), of School of Electrical Engineering and Computer Science (SEECS School) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

.

Signature: _____

Name of Supervisor: Dr. Muhammad Moazam Fraz

Date: _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Dedication

I dedicate this thesis to my parents and family members who were my greatest source of support.

Abstract

Person re-identification task is building up the correspondence between person images taken at different places and time from different cameras. It is an essential task for visual surveillance system. In this thesis, we propose a novel two stream convolutional - recurrent model with attentive pooling. Each stream of the model is a Siamese network and acquire different characteristics of feature maps. To fully utilize learned feature maps and to have some common features, we fuse the output of two streams. The use of attentive pooling in our model can choose only informative frames over the whole input video sequence. Then learn spatial and temporal information for these selective frames.

For proposed model we used [1] as our base model; but we perform a lot of changes in this base code as follow: (i) we make it two stream model (ii) we add attentive pooling layer in it, which is till now only used for action recognition tasks (iii) we add extra dropout layer in CNN base model (iv) RGB and optical flows are separately treat as input to learn spatial and temporal information separately (v) two stream fusion is done in proposed model to make one siamese cost feature for person re-ID. Fusion is done using weighted function. Which gives more weights to spatial features because these features are more discriminative as compare to temporal features.

Experiments are performed on three publically available person re-ID datasets: MARS, PRID-2011 and iLIDS-VID. Experimental results shows that our proposed model is considerably best for feature extraction, and it outperforms existing state-of-the-art supervised models. Results are more efficiently increased by using both RGB and Optical flows as input rather than using either of them independently. Proposed model gives 14.6%, 14.0% and 16% better accuracy for iLIDS-VID, PRID-2011 and MARS respectively at rank1 than base model.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Person re-identification

Individual re-ID is a process to recognize a person when he/she moves over disjoint non overlapping field of view [2-5]. It is important due to its application in action recognition, people tracking, and surveillance videos at public places (like airport, museums, train stations, shopping mall, roads, universities etc.). Individual re-id is a challenging computer vision task and can provide a powerful security tool in video surveillance applications [5]. Re-ID can for instance be utilized for following how individuals travel through an air terminal or museums. The data obtained from non-overlapping field of view, would then be able to be utilized for making insights portraying how individuals move around. Another utilization could be looking through numerous cameras for a man spotted in one camera field of view. This could for instance enable police to find a speculate utilizing less labor and time.

For making more pro-active surveillance, video analysis is very important task through which we can timely alert to related security person and predict undesirable event or activity of suspicious person [3]. In multiple camera network with non-overlapping field of view, matching is required when a person disappears from one camera view and same person appear in another camera field of view. System must assign same id to same person in when person move in different field of view shown in Fig 1 taken from [6], different colored dots represent different individuals and number beside the dots are IDs assign to individuals. Dotted line show the direction in which certain person move in the multiple camera network. A person may move in a specific location many times or may pick something unpaid, so person re-identification with single camera is also relevant in surveillance applications. Apart from surveillance it has many application in photo browsing, automatically photo tagging, robotics and multimedia.

Individual re-ID is the undertaking of finding the genuine match of a question subject over a scope of hopeful targets, which may show up fundamentally extraordinary in caught pictures [7]. A positioned list is produced of display pictures from well on the way to most drastically averse to be the test individual. Typically the best passage in positioned list is viewed as right that characterizes the exactness of calculations Fig 2 introduces the nuts and bolts of individual re-id where test picture is coordinated with display pictures to perceive the right parson.

We need automatic surveillance system, as multiple cameras are used and produce huge data so it is very difficult for humans to monitor all data manually. Person re-ID classified into

image based and video based. Till now, a lot of work have been done for image based re-ID. Variety of algorithms developed for image based re-ID including features representation learning [7-10], distance metric learning [11-16] and CNN based schemes [17-21]. More natural way for person re-ID is video based and currently researchers are more focused on this. Data volume for video based re-ID is larger than image based, because there are number of frames in each tracklet. For multiple frames, single match or multi match strategy is used for hand crafted features. These strategies induces higher computational cost. For this problem many authors proposed deep CNN based and pooling-based methods, which blends features of frame level into a global vector and gives better scalability [22-24].

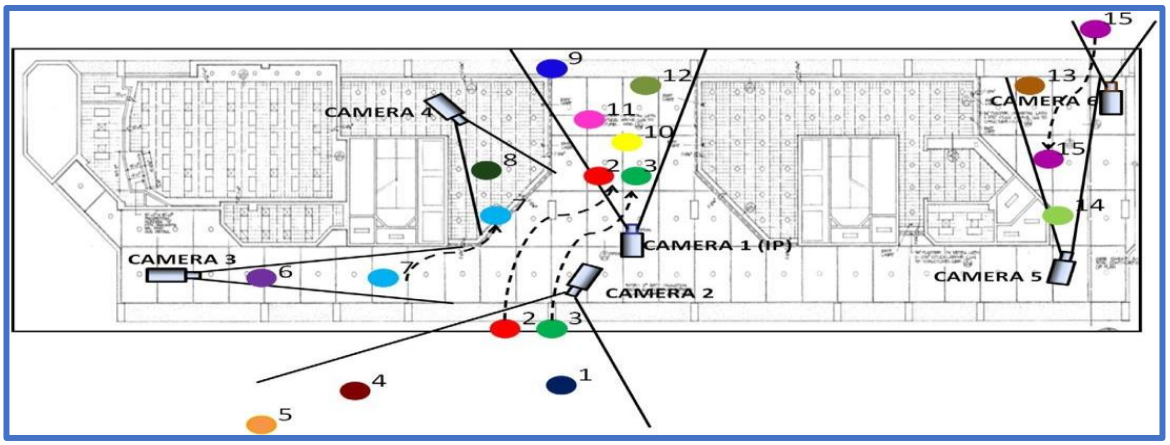

Figure 1: Multi camera surveillance and re- ID of a person [6].

Video based methods boost performance of person identification task due to multiples frames provide rich information about person and it involve temporal information associated to person motion can be taken using optical flows. Single stream or two stream architecture have been proposed for extracted optical flows. In single stream data fusion is done by concatenation of RGB images and optical are for the input of CNN-RNN model for training [1, 25, 26]. While in two stream architecture, model fusion is done. Two separate CNN are builds for optical flow and RGB images and fuse their output and feed it to RNN model [27-29].

Motivated from recent success of model fusion and data fusion, in this paper we proposed two stream CNN-RNN attentive pooling model architecture which is based on both data and model fusion to utilize learned features maps. The first component of video based methods is CNN model to extract automatically useful spatial features representations from input. Other important contributions of this work are: the use of temporal pooling operation at the output of CNN model for down sampling and to summarise long term appearance data with variable lengths and frame-rates [1]. To capture temporal information RNN model is used. Finally Siamese model is deployed on attention vectors, judge the extent of matching.

1.2 Re-identification Pipeline

Human agents entrusted by the measurable investigation of video from CCTV multi cameras systems confront several difficulties, comprising information over-burden from extensive numbers of cameras, restricted ability to focus prompting vital occasions and targets being missed, an absence of logical information indicating what to search for, and restricted capacity or, then again powerlessness to use correlative non-visual wellsprings of information to help the look prepare. Thusly, there is a particular requirement for an innovation to lighten the weight set on constrained HR and increase human abilities.

An automatic re-ID apparatus takes tracks or bounding boxes as input, holding sectioned pictures of single people, as produced by a confined following or discovery procedure of a visual observation framework. To consequently coordinate individuals at various areas after some time caught by various camera sees, a re-identification handle normally makes the accompanying strides:

- Extricating symbolism highlights that are more solid, hearty and compact than crude pixel information.
- Building a descriptor or portrayal, e.g. a histogram of elements, competent of both portraying and separating people.
- Coordinating determined test pictures or tracks against a display of people in another camera see by measuring the comparability among the pictures, or utilizing a few demonstrate based coordinating method.

Figure 2: Probe person is matched with candidate target in gallery set [30].

A preparation stage to enhance the coordinating parameters could conceivably be required relying upon the coordinating procedure. Such handling strides raise certain requests on calculation and framework outline. This has prompted both the improvement of new and the misuse of present vision methods used for tending to the issues of highlight portrayal, display coordinating what's more, deduction in setting. To accomplish of image/video based re-ID, some basic steps have to be done on raw input. The first step is detecting a person in image or video.

Second step is, extracting discriminating features from the frames, and last step is building a unique descriptor and matching for re-id as shown in Fig 3. Each step add a significant challenge and increase the complexity of the task discussed in up-coming section. Hence, the mission of re-ID of individual over a network of cameras is not easy because of the many challenges faced from describing features to designing system. These challenges are discussed in the upcoming section.

1.2.1 Detection

Contemporary ways to deal with re-distinguishing proof ordinarily misuse low level elements, for example, shading, surface, spatial structure or blends thereof. This is on the grounds that these components can be moderately effortlessly and dependably measured, what's more, give a sensible level of between individual segregation organized with between cameras invariance. These components are additionally encoded into settled size individual descriptors, e.g. as histograms, co-variances or fisher vectors.

1.2.2 Representation

Once a reasonable portrayal has been acquired, closest neighbor or, then again show based coordinating calculations, for example, bolster vector positioning might be utilized for re-ID. For each situation, a separation metric (like Euclidean or Bhattacharyya) must be measured the likeness among two examples. Model-based coordinating methods and closest neighbor remove measurements can both be discriminatively advanced to augment re-distinguishing proof execution given commented on preparing information of individual pictures. Crossing over these two phases, a few reviews have additionally endeavored to learn discriminative low-level components specifically from information.

1.2.3 Matching

Other integral parts of the re-recognizable proof issue have too been sought after to enhance execution, for example, enhancing vigor by consolidating numerous frames worth of components along a direction, set-based examination, considering outside setting, for example, gatherings of people, and learning the topology of camera systems with a specific end goal to lessen the coordinating inquiry space and consequently lessen false-positives.

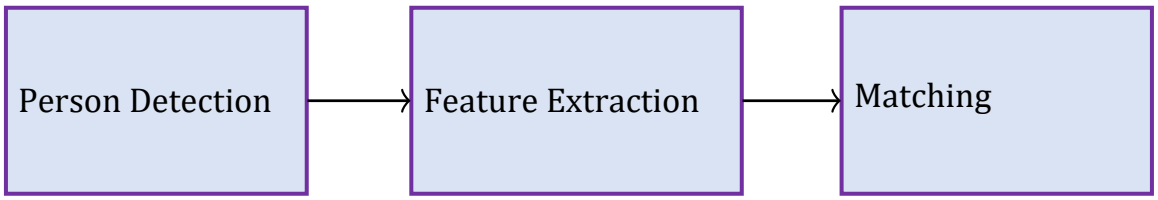

Figure 3: Re-identification basic steps

1.2.4 Classification of Methods

Diverse methodologies (as represented in various sections of this thesis) utilize marginally unique scientific classifications in arranging existing individual re-ID strategies. By and large, when just a picture match is matched, this model is deliberated as a single shot acknowledgment technique. In the event that coordinating is led among two arrangements of pictures, e.g. outlines acquired via two isolated directions, the strategy is identified as a multi-shot acknowledgment approach. An approach is ordered as a managed strategy if earlier to application, and it misuses named tests for tuning model parameters such as separation measurements, highlight weight or choice limits. Generally a strategy is viewed as an unsupervised approach in the event that it concerns the extraction of strong visual highlights and does not depend on preparing information. Obscuring these limits fairly are techniques which do gain from preparing information preceding arrangement, yet don't depend on explanation for these information.

1.3 Challenges in Person Re-id

In our daily life we met many people and for us, as human beings person re-identification is very simple and quite easy task without any struggle. Our brain and eyes are expert to detect, identify and re-identify different people, animals and objects in this world. For re-identification we use unique descriptor of person like face, eye color, hair color, clothing, height, walking style, voice; however, for automated system for human re-ID is very difficult task without human intervention. Human re-ID methods can be categories into biometric-based (iris, finger print, face, gait) and appearance-based (height, shape, clothes, walking style etc.). In biometric based methods we need special type of sensors and human cooperation to acquire biometric information. Therefore biometric based method are not suitable for person re-identification over wide area surveillance system. Appearance based methods are more suitable in surveillance system because they are less constrained as compared with biometric based methods. For appearance based methods there are two types of challenges which can be categories as intra-camera issue and inter camera issue. Fig 4 show some challenging like background clutter, variations in pose, variations in illumination and change in view angle.

1.3.1 Intra-camera issues

Multiple cameras which are used in surveillance task are low-resolution cameras. In addition change in light in a day (illumination) will cause change in background. Due to light effect and low resolution a person may disappear from scene or may be color or texture of clothes and background is same. We can take color and texture as feature when image is taken by camera a few minutes or hour apart (called as short period of time). This mean clothes (color and texture) are reasonable one but not used for long period of time in which images are taken month apart. Occlusion will cause as some parts of body may disappear or may be hand or leg of another person is added to different person as in images there are multiple persons are present. All of these challenges may cause variation in human appearance as some features are observed in one camera field of view and may disappear when person move to another camera field of view.

1.3.2 Inter-camera issue

As in multiple camera system with non-overlapping field of view, person images are taken with different cameras so it is very difficult for system to re-identify same person because of different angle of view, change in human pose and change in object position carry by human. Angle of view is changed because camera is located at different places. Human enter in camera range in different pose because human changes its pose while walking or even in a standing position. A person may look different due to change in object position carry with person like bag, jacket (zipped, unzipped), glasses, cap etc. Re-identification is a pipe line process consist of different processing steps like image segmentation, feature extraction and feature classification. Each of these task present a vast area of research in image processing and computer vision, but in this thesis we only consider concatenation of these steps.

Apart from this tracking is another big issue in multiple camera surveillance system with non-overlapping field of view. Tracking provides consistent labeling and establish corresponding between each detected person appears across multiple frames. Till now sufficient amount of work has been done for person detection and multiple person tracking within single camera field of view [15]; but multiple person tracking with in multiple camera with non-overlapping field of view remain an open problem. In person re-identification there are numerous open issues, some general issues are discussed in this section below:

1.3.3 System Design Challenges

In multi shot circumstance, highlights are mined from a video succession and from a picture, same as in the event of single shot re-ID. In both the situations, it is required to accurately identify the individual and after that restrict for proficient descriptor age. A solid correspondence should be worked among various picture edges of a similar individual for multi shot re-id. The correspondence recognizes that the numerous examples have a place with a similar individual and the procedure is known as following. Lately, broad research has been directed on issue of individual location and furthermore on individual following. Both the procedures are muddled issues and present their very own few impediments. Indeed, even following quite a while of research, there is opportunity to get better to devise proficient identification and following calculations.

1.3.4 Descriptor Challenges

After individual location, the following stage is to separate visual highlights from pictures. It is the most difficult and a pivotal piece of individual re-id errand. The need of re-id errand is to have hearty highlights for exact individual coordinating. In any case, it isn't generally conceivable because of uncontrolled situations as reasons like low determination, camera edges and settings differ crosswise over system and in a large portion of cases can't be controlled. Accordingly, the nature of descriptors exceptionally relies upon the earth in which information is gathered. What's more, wasteful individual identification and following calculations include blunders in descriptor age step. Hence, the broken component extraction process drops the productivity of re-id.

Numerous techniques are fused in extricating highlights and can be isolated into two sorts, single shot and multi shot. The previous techniques more often than not indicate low exactness rate as single picture is inadequate to accurately speak to a character. Then again, multi shot show high precision since numerous pictures are utilized for every individual to manufacture a viable descriptor. However, there is a confinement on number of IDs in a dataset as calculation cost increments. In both the situations, the extricated highlights ought to be discriminative in nature and vigorous to changes in enlightenment, posture, foundation mess, misalignment and impediment and so on. A few strategies have just been recommended that can accomplish sensibly high correctness's for individual re-ID. Most of feature extraction methods target a specific re-id challenge such as illumination or pose. The need to deal with multiple challenges simultaneously in a single descriptor therefore still exists and is open for improvement.

1.3.5 Matching Challenges

The second step of individual re-id is contriving strategies that will discover remedy highlight coordinate through metric learning or separation techniques. The component coordinating is unpredictable assignment as it is hard to separate over complex system of cameras with expansive datasets. Coordinating individuals over substantial system of cameras likewise progresses toward becoming non-minor since pictures are gathered on shifting areas, distinctive examples of day and time. Because of this, connecting individuals to amend ID winds up troublesome. The multifaceted nature additionally increments if there should arise an occurrence of multi shot strategy because of cost calculations as talked about before. Furthermore, when number of IDs increments in a dataset, uniqueness of descriptors decreases and calculations turns out to be expensive regarding memory and calculations. Different difficulties incorporate absence of explained information, little example estimate, appearance changes or comparative appearance, bury and intra camera issues and so on. Current techniques proposed somewhat take care of the issues of re-id however can't completely handle these difficulties. The accompanying are couple of issues that add to the difficulties of re-id.

1.3.6 Occlusion

The objective of intrigue get halfway or totally blocked because of close-by objects e.g. other individuals may impede them or when the objective is conveying pack or different s. In genuine situation, impediment by encompassing items is a typical event and it adds to difficulties of re-id. Because of impediment, it turns out to be relatively difficult to catch unmistakable highlights which ordinarily add to the nature of the descriptor.

1.3.7 Variation in illumination

Enlightenment condition is another testing factor that expansion the multifaceted nature of individual re-id. A similar individual saw by camera at day time will have distinctive enlightenments and shine if saw at some other time. Because of the distinction in enlightenment conditions extraordinary changes happen in target appearance. Numerous ongoing descriptors [31, 32] take light changes in thought and handle this test by utilizing some brilliance modifying

calculation like retinex calculation utilized as a part of before extricating highlights. This is a decent practice particularly in situations where appearance based descriptor is utilized as shading observation is principle data is such descriptors.

1.4 Person Re-identification Scenarios

This segment talk about the two situations of re-id a) close set and b) open set and each have issues of their own. A re-id errand comprises of an exhibition that incorporates marked pictures and a test that contain unlabeled pictures that are should have been perceived. The process of individual re-id framework is shown in Fig 5.

Figure 4: Challenges in person Re-ID [33].

Assume a gallery $(g_1,g_2,...,g_m) \in G$ and a probe $(p_1,p_2,...,p_n) \in P$ where IDs display in G are named and unlabeled in P. Typically in re-id system, given test is coordinated over every exhibition picture and a similitude score is processed. Based on comparability score a rank rundown is produced of display pictures that decides the genuine test coordinate. Display set might possibly be settled and in situation where exhibition does not develop after some time is known as close set re-id. This utilizes test is a piece of exhibition that is in close set re-id an ID is to be perceived which is as of now show in display, that could possibly have comparative appearance. Though in open set situation, display isn't settled and it is obscure if the test picture

to be contrasted has a place with exhibition. Accordingly, it is important to discover genuine match simply in the wake of checking the reality if the test is available in display. In setup where test isn't in exhibition, the test picture should be enlisted in display.

1.4.1 Close Set Re-id

By and large, re-id issue is encircled as close set re-id where test individual is the very same individual present in the display set with varieties of stances, hindrances and enlightenment settings and so on. The exhibition contains individual IDs catch over the system of cameras at various time and area. The numerous events caught of test subject over the system of cameras utilize that test appearance changes from camera to camera. Be that as it may, when the exhibition set stay settled after some time, the individual re-id issue winds up one to numerous re-id coordinating issue. Here best match is to be found to take care of the issue. The nearby set re-id is required where different cases of target individual is to be distinguished over a system of cameras.

1.4.2 Open Set Re-id

Open set re-id is required when following a specific focus over a system of disjoint cameras. Here exhibition set develops with time and the test set have vast number of non-target individuals alongside target individuals significance there is no certification that given test has a place with display. Besides, in following model there is a probability that in excess of one subject exists in scene and must be recognized in the meantime. In open re-id situation, discovering best match does not take care of the issue on the grounds that there is a plausibility that the test individual is absent in the display set. The best match is discovered just if closeness coordinating score fulfill following edge, $pr(g_m|p_n) > T$. The open set re-id issue is numerous to numerous coordinating issue and not very many endeavors have been made to explain it [34-36].

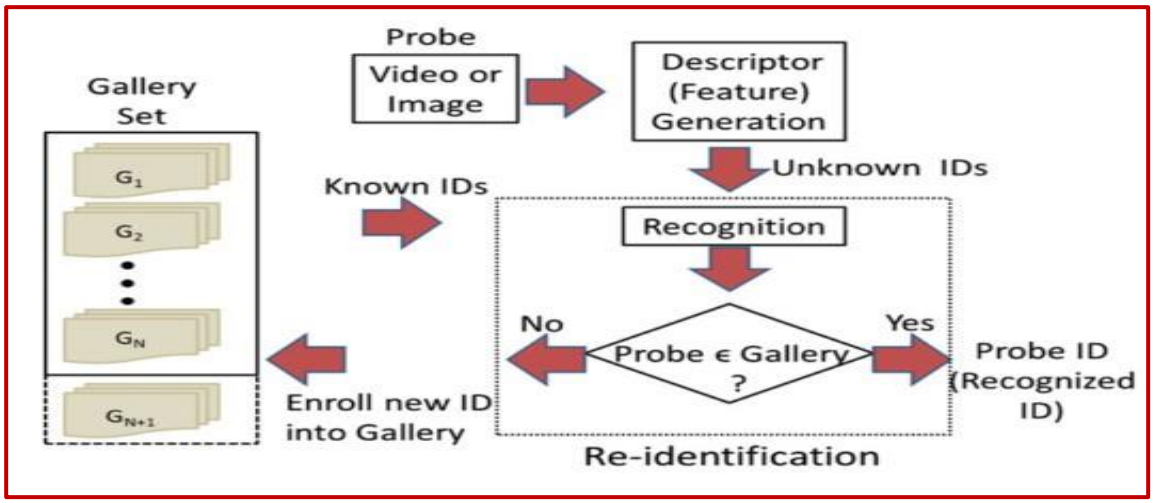

Figure 5: Person re-identification system framework [6].

1.5 Economic Benefits of Person Re-ID in Pakistan

Research is important and considered to be good if it contributing in the field as well as giving some economic benefit to the country. Same is the case here we are discussing the economic importance of person Re-Id with respect to Pakistan. In this section our focus is to answer all those question which are related to the importance of the topic and its role in the society especially in Pakistani society. Here are some of the examples through we try to show or to highlight the importance that if we can achieve this so we can easily get the benefits. The use of automated person re-identification can has significant benefits in Pakistan. Irregular activates like disturbance, unauthorized persons detection and terrorism can be monitored. From security to productivity it has following benefits.

Automated person re-identification offers wider viewing angels with fast data processing and provides quality of surveillance. So it reduces cost and time as compared to manual system. Remote monitoring through cameras saves a lot of human effort. It is saleable system we can grow our coverage area. Camera can be added to the network easily. With automated system keeps a sharp eye on the activities in public places it is impossible to manually watch the input data (images/videos) over long period of time. So automated software system can monitor activities and behaviors and alerts can be generated and send to security teams. The summary of above points is: with automated person re-identification monitoring activities in the workplace on daily basis the employees in the workplace feel safer, can take breaks efficiently knowing that the system is working in advance.

1.5.1 Camera Surveillance 24/7

The general public of today is inconceivably security cognizant because of the rising rate of violations over the world. The risky idea of the present world request a fundamental insurance to guarantee the wellbeing of family and business. Having a camera reconnaissance is incredible approach to dishearten the wrong goals of bystander or crooks. The account in cameras go about as a strong verification against the burglary or any criminal movement. Camera observation go about as obstruction for illicit exercises in organizations, workplaces, colleges and even homes. Video observation helps locate the missing individual, ensures the edge of your property, screen suspicious exercises and are likewise equipped for creating alerts. Cameras don't need rest and monitor our place 24/7 over multiple areas. It work for us in working hours as well as in off hours. While related security guards need break and time for sleep and don't monitor multiple places at a time. It will reduce our cost as it replace many security guards and easy to manage.

1.5.2 Security Application

Presently a-days surveillance cameras are sent all over, in parks and other open spots for security reasons. The CCTV cameras are sent to turn away security dangers and now and then to spy purposes. These CCTV cameras are ordinarily sent at shopping centers, government workplaces, banks and other such places to enhance security level. Movement police make utilization of surveillance cameras to peruse auto number plate and figure out who is mindful of overstepping

activity laws. Surveillance cameras are additionally helpful in deciding the present area of criminal free to move around at will.

1.5.3 Tracking People

Re-id framework ended up being exceptionally helpful in following the suspicious exercises of conceivable suspect in any criminal occasion. Instead of ordinary strategy where covert officer takes after could possibly be a suspect. Along these lines, squandering part of exertion and valuable time. Following battles the wrongdoing in a proactive way that is to foresee the conceivable criminal occasion before it happens. Numerous such techniques have been advanced where body developments may demonstrate conceivable risk like individual raising hands to shoot, his standing stance.

1.5.4 Robotic Applications

Numerous administration robots are composed as of late and are sent in healing centers, eateries and shopping centers, with the motivation behind conveyance, observation and cleaning. To satisfy the assignment close by the robots should know how to interface with encompassing people, for that they utilize systems like face acknowledgment, laser trackers and biometric highlights and so on. Robots in a large portion of nations are taking up human employments and are exceptionally proficient at it as well. One such application is medicinal services robots, where robots deal with patients by watch their condition and deciding when the patient need healing center visit.

1.5.5 Person Re-Id in Shopping Malls

Shopping Malls are the crowded places where people go for shopping, the importance of person Re-Id in the shopping Malls is that if we observe the visiting of individual customers so we can infer different meanings from that, we can detect good and advantageous information from it as well as some criminal meanings also we can take if we want. For example, after identifying the customers try to know their interests in the products and goods which they like and import more products and goods to better operate your business, and also it can avoid the loss by identifying the criminals and robbers in the premises of the Malls.

1.5.6 Deter crime

This is the main and biggest ease for installing video cameras in our homes, workplaces and at public places. Video cameras will help us to prevent crimes, illegal activities and provide us solution to problems like punctually of worker at offices. Apart from this cameras will provide us feeing of safety, inner satisfaction which is exquisite. As cameras are available in different shapes and size so it is very easy to place them anywhere (like in plants, on wall corners, photo frames etc.).

1.5.7 Gather Evidence

Cameras will help us to gather all information of any illegal activity, which will later use in legal places (like court, police station etc.). Wherein spectator may forgotten or may be not give correct detail due to pressure of someone else; so camera footage will help us to collect all evidence and making right decision in both domestic and professional scenarios. Apart from this it will also help us in a situation where we face controversy within your family, between your working staff and customer and ultimately increase productivity.

1.6 Motivation

The field of camera observation is quickly growing and is currently demonstrating a consistently expanding enthusiasm for the assignment of individual re-id. Vast system of cameras are conveyed out in the open spots like colleges, workplaces, air terminals and shopping centers and so forth. With the reason for checking suspicious exercises, following specific target and legal examination. In late decades, psychological oppressor brutality is on rise and there has been developing worry about open wellbeing. It is indispensably vital to put CCTV (close circuit TV) cameras on open places as video reconnaissance can assume a key part in battling fear mongering. Such an observation framework can give experts an extraordinary instrument to keep up abnormal state security. The systems of cameras deliver enormous measure of information and physically experiencing the recordings/pictures is wrong and tedious assignment. It is basic to have self-sufficient framework as human checking decrease viability of reconnaissance. The human autonomous reconnaissance framework helps recognizes target individual quicker in a proactive way and give considerably more effective security.

Depending on human administrator manual re-recognizable proof in vast camera systems is restrictively exorbitant and erroneous. Administrators are frequently relegated a greater number of cameras than they can practically screen all the while, and even inside a solitary camera, manual coordinating is defenseless against inescapable attentional holes. Besides, benchmark human execution is controlled by the individual administrator's understanding among other variables. It is hard to exchange this skill straightforwardly amongst administrators, and it is hard to acquire steady execution because of administrator inclination. As open space camera systems have grown quickly lately, it is ending up plainly progressively clear that manual re-ID is not adaptable. There is consequently a developing interest inside the PC vision group in creating computerized re-ID arrangements.

1.7 Thesis Contribution

An end-to-end two stream convolutional-recurrent model with attentive pooling is proposed in this paper. In which both two streams learn different feature maps, and finally merge output of two stream to obtain union of characteristics. This model can act as features extractor for video based re-ID, as well as it produce hidden unit representations for measuring similarity score for time series input. Another contribution of our work is use of attention pooling for two stream CNN-RNN. Which solve problem occur in other DNN based models due to max pooling. Till

now attention pooling work best for action recognition task [37], image caption generation [38] and questioning-answering model [39].

Another main contribution of two stream network is, not applying fully connected layers at the end of CNN and RNN. We replace fully connected layer of data fusion model with simple convolutional layer and achieved state-of-the-art performance. By replacing fully connected layer, we get image level representation which is more meaningful and have more information of image due to high dimension of features space. Comprehensive tests on PRID-2011, iLIDS-VID and MARS datasets, our result show that proposed novel model achieve greater performance to the existing state-of-the-art re-ID models. Our architecture (TSCRA) is easy and simple to implement. It require little more computational time as compared to single stream networks, but provide superior performance on real, versatile and large datasets. Which make it most desirable model for re-ID in future.

For proposed model we used [1] as our base model; but we perform a lot of changes in this base code as follow: (i) we make it two stream model (ii) we add attentive pooling layer in it, which is till now only used for action recognition tasks (iii) we add extra dropout layer in CNN base model (iv) RGB and optical flows are separately treat as input to learn spatial and temporal information separately (v) two stream fusion is done in proposed model to make one siamese cost feature for person re-ID. Fusion is done using weighted function. Which gives more weights to spatial features because these features are more discriminative as compare to temporal features.

Experiments are performed on three publically available person re-ID datasets: MARS, PRID-2011 and iLIDS-VID. Experimental results shows that our proposed model is considerably best for feature extraction, and it outperforms existing state-of-the-art supervised models. Results are more efficiently increased by using both RGB and Optical flows as input rather than using either of them independently. Proposed model gives 14.6%, 14.0% and 16% better accuracy for iLIDS-VID, PRID-2011 and MARS respectively at rank1 than base model.

1.8 Thesis Organization

In next *Chapter* 2, literature is discussed. The *Chapter* 2 briefly discusses the related work in domain of person re-id and their comparison. In *Chapter* 3, proposed novel model with experimental setting is discussed. *Chapter* 4 give details on the dataset considered for evaluation and *Chapter* 5 demonstrate the performance of proposed method and comparison with state-of-the-art approaches. Conclusion is presented in *Chapter* 6.

Chapter 2

Background and Related Work

2.1 Typical architecture for re-id

In this chapter the re-ID issue is depicted in a general point of view, taking into account every one of the modules and usefulness required for a high level of autonomy in video reconnaissance frameworks. One conceivable general architecture of a robotized re-identification is displayed in Fig 5. Each re-ID framework depends on gallery set and a test. An exhibition set is made by pictures or successions of pictures from a man to be perceived over the cameras of the system. The test is a picture of a person to be re-distinguished against the exhibition pictures.

A gallery set is either obtained offline or on the web. In the offline rendition individuals are enrolled to be permitted to enter the space. In the online variant the display is refreshed as individuals enter and leave the framework. In the online variant we can likewise recognize shut spaces, where the exhibition cases are obtained at uncommon passages in the camera system, and open-spaces, where the exhibition illustrations are obtained anytime. At runtime, people are identified from the camera system's pictures. Recognitions are typically spoken to as jumping boxes around the people's pictures and can be gotten either by the framework's administrator manual intercession or consequently, by person on foot location calculations or foundation subtraction techniques. The procedure of re-ID at that point comprises in partner runtime individual location to the exhibition illustrations. Investigation should be possible at singular edges (single-shot) or with numerous outlines from tracks inside a similar camera (multi-shot). Investigation is ordinarily performed taking a gander at a few highlights extricated from the people bouncing boxes, e.g., shading, shape, surface, or movement.

These highlights are then related to cases in the display through suitable classifiers. Classifiers extend from as straightforward as NN to more mind boggling regulated or semi-administered strategies. At long last the ordered people on foot are followed all through the camera arrange abusing however much as could reasonably be expected the limitations in the system topology what's more, human movements, e.g., by means of Multiple Hypothesis Tracking (MHT). In the accompanying area are portrayed in more detail the most vital segments essential for certifiable applications, which envelop some of the difficulties handled in this postulation.

2.2 Components for re-id

Literature review of components for person re-ID are discussed below:

2.2.1 Person Detection Techniques

Some individual re-id techniques require earlier information of human area in the picture that is portrayed by bouncing boxes. This pre-hand data causes them to extricate includes just from human body parts barring foundation clamor. The most generally utilized individual recognition strategy histograms of arranged inclinations is proposed by [40]. Numerous individual re-id strategies [41, 42] have utilized their model for human identification and some [43, 44] have additionally utilized it for distinguishing singular body parts. This indicator can be utilized as a part of situations where single casing is available, however require expansive preparing set. Best outcomes are accomplished with this finder in situations where there are changes in light settings or multi scale procedure is utilized. However, with posture varieties vast mistake is presented in identification. Disentangled LBP (SLBP) identifier is presented in [45] that utilized Ada Boost classifier for preparing purposes. This finder have an edge over histogram of arranged inclination indicator as it handles the multi scale challenge. An ongoing indicator approach is presented in [46] which is utilized for individual re-id purposes by [47] giving them energy to do continuous following. The locator separate signs data by looking at neighborhood pixels, encode it and figure histograms. Another constant philosophy is proposed by Aziz et al. [48] that utilize skeleton chart for performing effective identification in swarm. Impediment is a typical event in a swarmed put and when such situation is delivered their strategy catches the individual straightforwardly before the camera as it were.

Shading and spatial information caught in pictures are basic data in following a moving individual. The technique proposed by Javed et al. [49] join the spatial data keeping in mind the end goal to identify a moving individual and furthermore utilized shading histograms to catch shading data. The mark is relegated to the pixel if the result of shading and spatial qualities is greatest of that pixel. A frontal area district is then dispensed to moving item if adequate number of pixels have most extreme qualities.

2.2.2 Feature Extraction and Representation Techniques

It centers on speaking to importance full fixes in a picture into numerical vectors. These vectors go about as a component descriptor for that picture. Numerous endeavors have been made to extricate highlights that are vigorous to varieties in brightening, perspectives, impediment and intra-individual appearances [50]. The exploration did from this point of view of re-id framework are talked about in following section.

2.2.2.1 2D Appearance Features

Another method for diminishing the impact of foundation mess is by disintegrating a full person on foot picture into verbalized body parts, e.g. head, middle, arms and legs. Along these lines, one wishes to concentrate specifically on similitudes between the appearances of body parts while sifting through as a significant part of the foundation pixels in vicinity to the forefront as could be expected under the circumstances. Actually, a section based re-recognizable proof

portrayal shows better heartiness to fractional (self) impediment and changes in nearby appearances.

2.2.2.2 Salient Feature Selection

Two inquiries emerge: (1) Are all elements break even with? (2) Does the value of an element (sort) all around hold? Lamentably, not all elements are similarly vital or valuable for re-distinguishing proof. A couple of segments are more discriminative for personality, while others more tolerant or invariant to camera see changes. It is essential to choose both the conditions and the level of the estimation of every part. This is considered as the issue of highlight weighting or highlight determination. Current re-ID procedures for the most part expect certainly a component weighting or determination instrument that is worldwide, i.e. an arrangement of bland weights on highlight sort's invariant to a populace. That is, to expect a solitary weight vector or separation metric (e.g. Mahalanobis separate metric) that is comprehensively ideal for all individuals. For example, one regularly accept shading is the most essential (instinctively so) and all around a great component for coordinating all people. Other than heuristic or exact tuning, such weightings can be learned through boosting, positioning, or separation metric learning.

2.2.3 Feature Matching Techniques

The second step of individual re-id manages finding right component coordinates through separation equations and metric learning. Liao et al. [51] proposed a powerful metric learning strategy which ventures separated highlights into another subspace that is dimensionally effective and discriminative in nature. The greater part of metric learning approaches plan to diminish intra class contrasts yet the thought proposed in [52] by Zheng et al. centers around relative separation correlation streamlining. Nonetheless, the strategy is that it ends up unmanageable on vast datasets. Since camera settings shift from camera to camera in a multi camera arrange situations, utilizing a summed up metric learning is a trade off on precision. This issue of varieties in multi camera arrange is handled in [53]. Mama et al. [53] proposed a various metric learning methodology where Mahalanobis separations are ascertained from pictures of a solitary ID taken from an indistinguishable camera from well as from various cameras in the system. The strategy proposed in [54] is straightforward yet proficient and has been used widely in other individual re-id technique. Notwithstanding, substantial calculation is one of the confinements of this technique when utilized with positive semi-clear requirement.

2.3 Literature Review for video based person re-ID

In literature, researchers have explored various algorithms for single shot/image based individual re-ID. These techniques can be characterized into two classes: (i) appearance based features extractor like color histogram, texture histogram and local binary pattern [10, 13, 15, 55-58] (ii) metric learning based methods like large margin nearest neighbor (LMNN), Rank SVM and

Mahalanobis distance [10, 59-61]. However all of these images based methods are not using temporal information.

Video based re-ID use temporal information as multiple images are used for probe vs gallery images. Multi images provide us rich information about person; but dataset is large and image based methods can't perform well [55]. To replace image based methods, deep neural network (DNN) model become popular for video based re-ID [18, 62-64]. Typical DNN architectures can be decomposed into two parts: in first part Convolution neural networks (CNN) or recurrent neural networks (RNN) are used for features extraction. In second phase for making final prediction that query person is same or not, multiple matric learning layers are used.

In writing, individual re-ID is for the most part investigated with single pictures (single shot). As of late, video-based re-identification has turned out to be mainstream because of the expanded information abundance which actuates more research potential outcomes. It shares a comparative definition to picture based re-identification. As essential as the jumping box highlights seem to be, video-based strategies give careful consideration to multi-shot coordinating plans and the incorporation of fleeting data.

Lin et.al [64] Proposed first Siamese based CNN architecture, which leveraged SCNN model to the covered parts of the individual picture. Y. Yan et.al [19] Present a new recurrent feature aggregation architecture for person re-ID, which can learned sequence level representation from multiple frames. Similarly Jesus Martinez et.al [1] proposed another model combination of CNN and RNN, which also capable of learning features representation from multiple frames. In their model CNN extract spatial information and fed features to RNN to effectively learn temporal information between different time frames. However, these models are not fully utilize all information available in tracklets; due to use of pooling either max pooling (which utilized just dynamic component outline one time step) or average pooling (averaged over all time steps and give us a single feature representation vector).

To solve the problem because of max and average pooling, [64] proposed a model which utilizes pyramid and attention pooling layers. Similarly [65] present temporal attention model (TAM) and spatial recurrent model (SRM). To accumulate motion information [29] proposed two stream CNN and [26] proposed Bidirectional RNN for action recognition and event recognition. Many work have been proposed to understand features for both spatial and temporal dimensions through RNN [66-68]. However, their techniques depend on ImageNet pre-trained CNN maps.

An end-to-end two stream convolutional-recurrent model with attentive pooling is proposed in this thesis. In which both two streams learn different feature maps, and finally merge output of two stream to obtain union of characteristics. This model can act as features extractor for video based re-ID, as well as it produce hidden unit representations for measuring similarity score for time series input. Another contribution of our work is use of attention pooling for two stream CNN-RNN. Which solve problem occur in other DNN based models due to max and average pooling. Till now attention pooling work best for action recognition task [37], image caption generation [38] and questioning-answering model [39]. Table 1 shows some recent challenger models.

The vast majority of the past research defined the undertaking of re-id as a genuine match recovering issue or accurately perceiving errand. Accept an exhibition with vast number of named targets and a solitary or numerous occasions of test subject, the objective of re-id framework is to recover genuine match. In view of similitude score a rank rundown is processed of the exhibition set as for the comparability to the given test subject. By and large, it is normal that the main picture in the rank rundown is genuine ID. Such is the situation when test is a piece of display that characterizes close set situation. As specified above the greater part of the methodologies consider close set re-id. For the most part, the re-id undertaking can be part into three noteworthy stages a) portion individual from undesirable information in picture, b) remove individual highlights and c) process similitude between target subject and display. The forthcoming area talks about the work did in re-id regarding chain of command of individual re-id as appeared in Fig 3.

2.3.1 Hand-crafted Systems

In 2010 initial two trials [69, 70] were hand-created frameworks. These papers essentially utilize shading based descriptors and alternatively utilize frontal area division to identify the person on foot. They utilize comparable picture elements to picture based re-identification strategies, where the significant distinction is the coordinating work. As specified in Section 1.2, both techniques usually compute the base Euclidean separation among two arrangements of bouncing box includes as the set comparability.

In substance, such techniques ought to be grouped into "multi-shot" individual re-identification, where the likeness between two arrangements of casings assumes a basic part. This multi-shot coordinating procedure is embraced by later work [71]. In [72], numerous shots are used to prepare a discriminative boosting model in light of a set of covariance elements.

In [73], the SURF neighborhood highlight is used to recognize and portray intrigue focuses inside short video arrangements that are thus filed in the KD-tree to speed up coordinating. In [74], a spatial-fleeting diagram is created to recognize spatial-transient stable areas for forefront division. The nearby depictions are then figured utilizing a grouping strategy after some time to enhance coordinating execution. [75], utilize the complex geometric structures from video groupings to build more minimal spatial descriptors with shading based elements. [76] propose utilizing the restrictive irregular field (CRF) to join limitations in the spatial and transient areas. In [43], hues and chose confront pictures are utilized to fabricate demonstrate over casings that catch the trademark appearance and additionally its varieties after some time. [77], make utilization of multi-shots for a man and recommend that the test highlight be exhibited as a straight mix of the same individual in the exhibition. Various shots of a personality can likewise be utilized to upgrade body part arrangement.

2.3.2 Deep Learning based Systems

In video-based re-identification, information size is ordinarily bigger than picture based datasets, on the grounds that each track let contains a number of edges. A fundamental distinction among

video based and picture based re-identification is that with different pictures for each coordinating unit (video succession), either a multi-match system or a single match methodology after video pooling ought to be utilized. The multi-coordinate methodology is utilized as a part of more seasoned mechanism [69], which prompts higher computational cost and might be risky on extensive datasets. Then again, pooling-based strategies totals outline level elements into a worldwide vector, which has better adaptability. As an outcome, current video-based re-identification strategies commonly include the pooling step.

This progression can be max pooling as [1], or learned by a completely associated layer [19]. In framework of [78], worldly data is not expressly caught; rather, outlines of a personality are seen as its preparation tests to prepare an arrangement CNN show with soft max misfortune. Outline highlights are collected by max pooling which yield aggressive exactness on three datasets. These strategies are turned out to be powerful, but then there is a lot of space for development. As for this point, the re-identification people group can obtain thoughts from the group of activity/occasion acknowledgment. For illustration, [79] propose conglomerating the section includes in the fifth convolutional layer of CaffeNet into Fisher vectors [80] or VLAD [81], in direct CNN include exchange. [82] propose a figuring out how to-rank model to catch how outline highlights develop after some time in a video, which produces video descriptors of all-inclusive worldly elements. In [83], implant a multi-level encoding layer into the CNN model and create video descriptors of shifting succession lengths. Another great practice comprises of infusing worldly data in the last portrayal.

For hand crafted frameworks, [84, 85] utilize immaculate spatial-transient includes on the iLIDS-VID and PRID-2011 datasets and report focused precision. In [23], in any case, it is demonstrated that the spatial-temporal components are not adequately discriminative on the MARS dataset, in light of the fact that numerous people on foot share comparative waling movement under a similar camera, and on the grounds that movement highlight of a similar individual can be unmistakable in various cameras. The point shown up elements are basic in a huge scale video re-identification framework. All things considered, this overview points out for the current works of [1, 19, 55], in which appearance highlights (e.g., CNN, shading. what's more, LBP) are utilized as the beginning stage to be sustained into RNN systems to catch the time stream between casings. In [1], elements are separated from back to back video outlines via CNN, and after that sustained via an intermittent last layer, with the goal that data stream between time-steps is permitted. The components are then joined utilizing max or normal pooling to yield an appearance include for the video. Every one of these structures are joined into a Siamese system. A comparable design is utilized as a part of [55]. In [19], hand-made low-level elements, for example, shading and LBP are sustained into a few LSTMs and the LSTM yields are associated with a soft max layer. In real life acknowledgment, [86] propose separating both appearance and spatial temporal highlights from a film and assemble a mixture organize to meld the two sorts of elements.

Table 1: Competitor models for individual re-ID.

N0.	Technique Name	Year	Main Contribution	References
1	**DVR**	2014	This model prosed for person re-id, it uses discriminative space and time features to inevitably exploit most consistent video fragments.	[28]
2	**STA**	2015	Spatio-temporal body action architecture, which yield input as walking person video sequences and forms a spatio temporal appearance depiction for pedestrian re-ID.	[84]
3	**AFDA**	2015	This model hierarchically group's image sequences and uses the characteristic data samples to acquire a feature subspace exploiting the Fisher benchmark.	[87]
4	**SRID**	2015	Sparse re-Id, which block sparsity for person re-Identification. In this method feature vector of query person exist in the linear extent of the consistent gallery feature vectors in a learned embedding space.	[77]
5	**DVDL**	2015	Person re-Ide with discriminatively trained viewpoint invariant dictionaries, which acquire skilled dictionary of discriminatively and sparsely encoding features demonstrating different people.	[88]
6	**RFA**	2016	Recurrent feature aggregation model, depends on LSTM. This approach sums the frame-wise person region description at every time stamp and yields a sequence-level demonstration.	[19]
7	**CNN-RNN**	2016	Proposed unique convolutional recurrent model with temporal pooling.	[1]
8	**ASTPN**	2017	Jointly spatial and temporal pooling model with attentive for video-based person re-ID.	[17]
9	**RNNPR**	2018	Recurrent neural networks for person re-ID revisited model in which RNN are revisited, leading to a simpler feed-forward architecture and exposes the small effect of recurrent connections.	[2]

Chapter 3

The Proposed Model

For video based re-ID, this paper present a new end-to-end two stream convolutional-recurrent model with attentive pooling (TSCRA). To obtaining image level feature maps for one time step, RGB images and optical flow are separately passed to two stream CNN's as shown in Fig 6. Then this images level representation fed into recurrent neural network, to extract temporal information of video sequence. Last step is combining all time steps of resulting RNN to get sequence level representation, through using attentive pooling. Following subsections will discussed two stream model and attentive pooling in more detail.

For proposed model we used [1] as our base model; but we perform a lot of changes in this base code as follow: (i) we make it two stream model (ii) we add attentive pooling layer in it, which is till now only used for action recognition tasks (iii) we add extra dropout layer in CNN base model (iv) RGB and optical flows are separately treat as input to learn spatial and temporal information separately (v) two stream fusion is done in proposed model to make one siamese cost feature for person re-ID. Fusion is done using weighted function. Which gives more weights to spatial features because these features are more discriminative as compare to temporal features. Experiments are performed on three publically available person re-ID datasets: MARS, PRID-2011 and iLIDS-VID. Experimental results shows that our proposed model is considerably best for feature extraction, and it outperforms existing state-of-the-art supervised models. Results are more efficiently increased by using both RGB and Optical flows as input rather than using either of them independently. Proposed model gives 14.6%, 14.0% and 16% better accuracy for iLIDS-VID, PRID-2011 and MARS respectively at rank1 than base model.

3.1 Two-Stream CNN-RNN

All previous DNN approaches for person re-ID are based on single stream model. In which pooling is generally used for down sampling. However these approaches can't utilize full feature maps due to use of max/average pooling. As already discussed max pooling ignored many learned features and only focus on maximum values features. Similarly average pooling treat all features equally, although some features are not properly learned and decrease overall performance.

To solve this problem and to fill this gap, we use two stream CNN-RNN model. For our model we use color and optical flow, which allow to register short-term spatio-temporal information [1]. We used multiple convolutional layer for directly down sampling through using of stride of 2, same as used in [89]. Unlike max and averaged pooling, it focus on learned feature maps with fixed position and give us better performance as shown in 4.2.1 section. By using

stride =>2 in two stream CNN-RNN model we can utilize all feature maps. Each feature map resized and fed to RNN, recurrent layer is formulized by equation (1).

$$o^t = Ur^t + Ws^{t-1} \quad , \quad s^t = \tanh(o^t) \tag{1}$$

Where $r^t$is input of recurrent layer for time t, s^{t-1} is the hidden state which contain information for prior time step, and o^t is the output. Through matrix U recurrent layer implants high dimensional feature vector into low dimensional feature vector. For first time step hidden state (s^0) is set to zero; t*anh* activation function is used to pass hidden state between different time steps. In our proposed model each stream help other stream to learn multiple different aspect of feature maps. Which is not possible in single stream models. Another main contribution of two stream network is, not applying fully connected layers at the end of CNN and RNN. We replace fully connected layer of data fusion model with simple convolutional layer and achieved state-of-the-art performance. By replacing fully connected layer, we get image level representation which is more meaningful and have more information of image due to high dimension of features space.

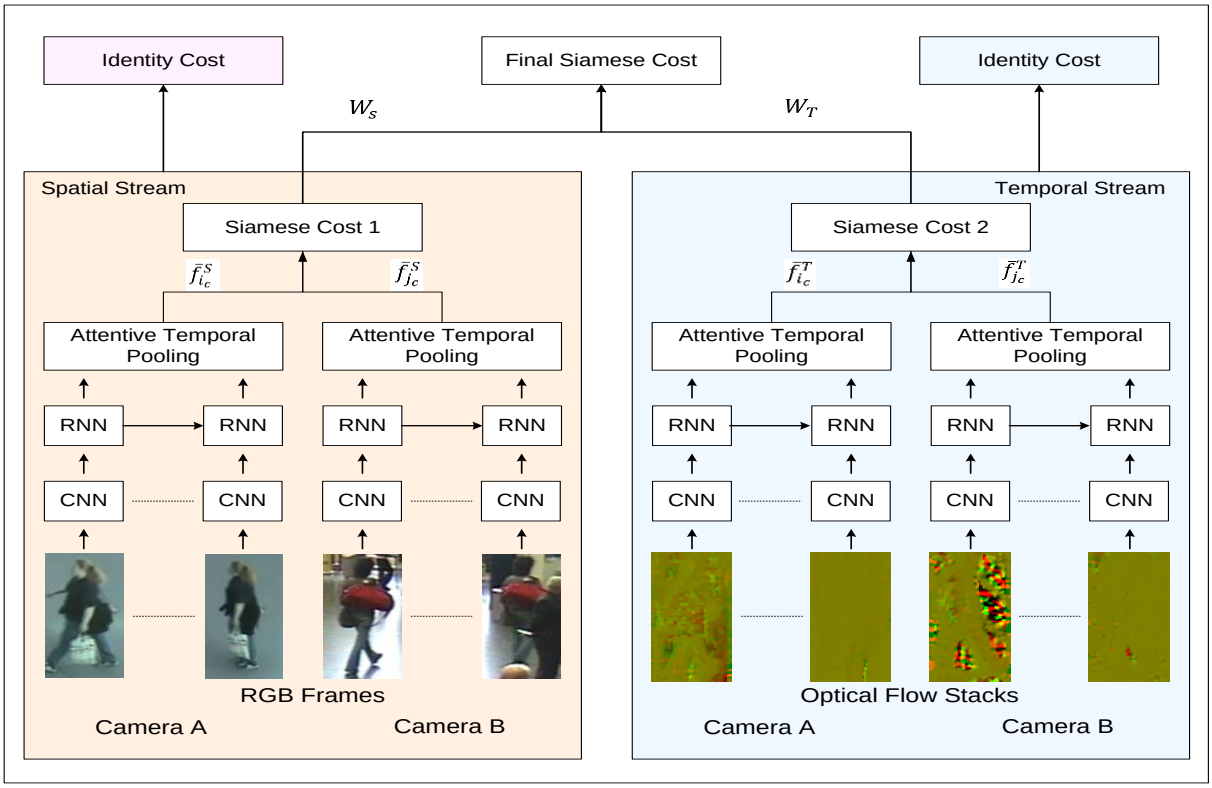

Figure 6: General model of the proposed two stream CNN-RNN.

3.1.1 Base CNN model

Fig 7 demonstrates the CNN structure and related hyper-parameters for it. For our model we modify base CNN [1], our CNN model takes one person video sequences as input (S(t) or T(t)) and produces the vector fS or fT for RNN. Our base CNN is made out of three convolution layers where each layer has convolution, non-linear activation and padding steps. For activation we use

non liner activation function, hyperbolic-tangent (tanh). Toward the finish of the three convolution layers, a completely associated layer is put to have mapping to every one of the enactments from the last convolution layer. Dropout [90] is additionally used to diminish the model over-fitting.

3.1.2 Base RNN model

The standard recurrent unit with a direct actuation work, $\sigma_{linear}(a) = \beta a + e = a$, can be communicated by equation (2):

$$h_t = \sigma_{linear}(Wh_{t-1} + Ay_t + b) = Wh_{t-1} + Uy_t + e \quad (2)$$

Where (W, A) are weight frameworks related with the state h and the outer info x, b is an inclination term. The standard intermittent unit is pained by two undesirable issues, detonating inclinations furthermore, vanishing inclinations. These issues can without much of a stretch be comprehended by assist examination of equation (3). In the event that the outer info (y, e) are disregarded from the condition, the intermittent unit can be re-communicated as $h_t = Wh_{t-1}$. By extending the repetitive relationship, the state at a given time step t is rearranged to [91],

$$h_t = W^t h_0 \quad (3)$$

On the off chance that the lattice W is square (N × N), with N straightly autonomous eigenvectors an Eigen decomposition can be performed same as in [91], with the end goal that W is communicated as equation (4).

$$W = M\Lambda M^{-1} \quad (4)$$

Where the segments in M speak to eigenvectors and the inclining grid Λ speaks to the eigenvalues. Equation 5 would thus be able to be re-communicated as below,

$$ht = M\Lambda^t M^{t-1} h_0 \quad (5)$$

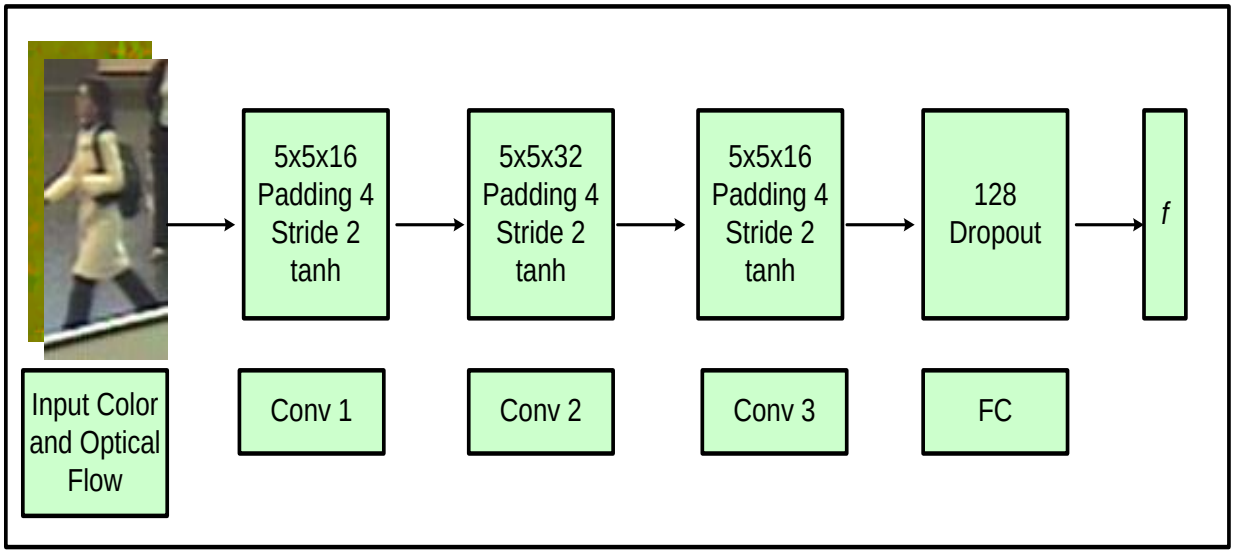

Figure 7: Proposed CNN structure with hyper parameters.

Every eigenvalue λ_i is therefore duplicated with itself t times. The angle $\frac{\partial h_t}{\partial h_0} = M\Lambda^t M^{-1}$ will consequently rot towards zero or detonate as t turns out to be vast, contingent upon the greatness of the eigenvalues. The normally utilized actuation capacity of the standard intermittent unit in equation (3) is tanh. The vanishing and detonating slope issue is likewise present for tanh and other non-direct initiation capacities with outside information sources and predispositions not quite the same as zero [92]. The detonating inclination issue can be reclaimed by thresholding the standard of the angle. It has been appeared that such an imperative does not seriously influence the system's capacity to learn worldly connections inside the outside information [92, 93]. The vanishing slope issue is a more difficult issue to be explained. The capacity of a system to engender data starting with one time step then onto the next is hampered by brief time vacillations which have a tendency to rule the commitments to the shrouded state [91]. The vanishing inclination issue forces accordingly an extreme issue when long haul conditions ought to be educated.

3.1.3 Attentive Temporal Pooling Layer

To capture temporal information, recurrent layer is used with hidden states, but there are a lot of redundant information as clothes and cluttered background. There are very few changes in continuous frames. To tackle with this problem we present attentive temporal pooling in our model; which only focus on effective information. Attentive pooling perceive data pairs for both gallery and query person, and permit the query input sequence to straight affect the calculation of gallery sequence illustration vg. Attentive pooling layer is placed between RNN and distance measuring layer same as proposed in [17, 39, 94] shown in fig 8. By convolution and recurrent layer we obtain P and G matrices for probe and gallery respectively; whose i^{th} row denotes output in the i^{th} time step of the recurrent. Then attention matrix is computed by equation (6):

$$A = \tanh(PUG^t) \tag{6}$$

Where both P and Q ε $R^{T \times N}$, U ε $R^{N \times N}$ and A ε $R^{T \times T}$. Matrix U is learned by network and intent for information sharing. Attention matrix (A) calculates weight scores in temporal dimension and it is capable to have vision on both query and gallery sequence features.

Next, temporal weight vector for probe (t_p) and gallery (t_g) are computed by column and row wise max pooling respectively on matrix A. Finally softmax function is applied on temporal weight vectors. Which transform the i^{th}weight $[t_p]_i$ and $[t_g]_i$to the attention ratio $[a_p]_i$and $[a_g]_i$using the following equation (7):

$$[a_p]_i = e^{[t_p]_i} \Big/ \sum_{j=1}^{T} e^{[t_p]_j} \tag{7}$$

Where equation 8 applies to both two stream CNN-RNN similarly with changed sort of input. To acquire the sequence-level representation v_p and v_g, apply dot product among the feature matrices P, G and attention vectors$a_p$as shown in equation (8):

$$v_p = P^T a_p \quad , \quad v_g = G^T a_g \tag{8}$$

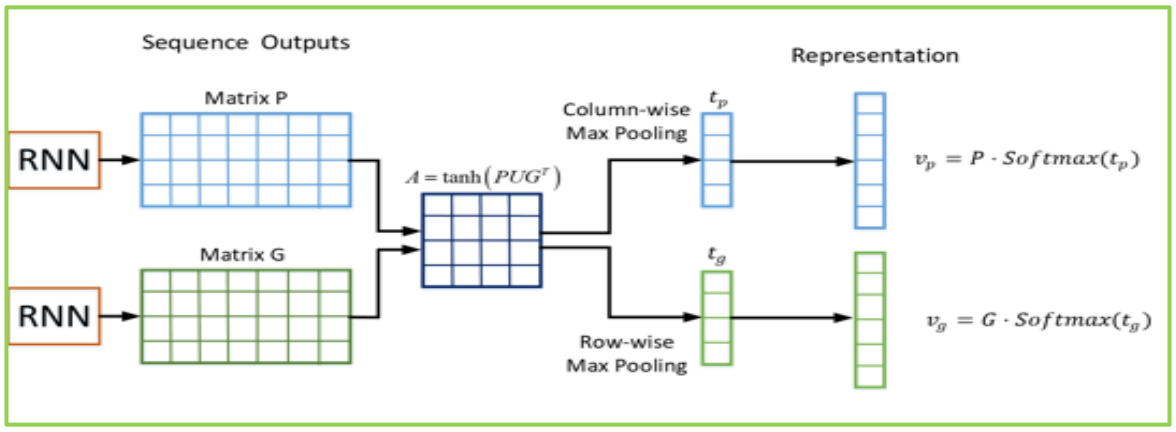

Figure 8: Attentive Temporal Pooling [17]

3.1.4 The Siamese Network

Siamese systems are made with two sub-systems with shared weights [95]. Euclidean distance is used by each of sub-networks to learn features from pairs. For training, model tries to decrease the separation between those element sets which are from a similar class and endeavor the separation between include sets when they have a place with various classes. Because of this property, Siamese systems have been widely utilized for individual re-ID. As stated earlier, we use a Siamese network for both two streams shown in Fig 6. Moreover, the general Siamese cost of our proposed architecture can be described as follows in equation (9). Where $f^-_{i_c}$ and $f^-_{i_c}$ are the attentive pooled feature vectors for person i and j separately. Similarly m denote Siamese margin.

$$D(f^-_{i_c}, f^-_{j_c}) = \begin{cases} \frac{1}{2}\,||f^-_{i_c} - f^-_{j_c}\,||^2 \quad , & \text{if } i = j \\ \frac{1}{2}\{\max(m - \|f^-_{i_c} - f^-_{j_c}\|, 0)\}^2 \quad , & \text{if } i \neq j \end{cases} \tag{9}$$

This system is utilized to lead tries different things with sets of pictures. Like the majority of the profound learning models face to face re-recognizable proof writing, we utilize the siamese design for our system since it can acknowledge test competitor sets of pictures as info. To separate the highlights from the pictures we utilize a progression of convolutional layers, trailed by max pooling, at that point went through an actuation work and standardized utilizing cluster standardization. After that the extricated combine of highlight vectors gets converged to one element vector and prepared by the comparability learning segment. The individual pictures are very little, so we don't require numerous convolutional layers to extricate the highlights.

3.1.5 Softmax Regression

On final features are V_p and V_g. we use softmax regression on to guess the identity of individuals. Via cross-entropy, identity loss I (V_p) and I (V_g) is computed on these final features. Meanwhile mutual learning through identity loss and Siamese loss is a great raise. Our ultimate training goal is grouping of Siamese and identity loss L (V_p,V_g) = E (V_p, v_g) + I (V_p) + I (V_g). Where E denote Euclidean distance [17].

3.1.6 Fusion

Fusion can be done any time in the two systems, with the main limitation that the two feature maps I_T^a and I_T^b belong to R^{HXWXD}, at time t, have the same spatial measurements. This can be accomplished by utilizing an "up-convolutional" layer [96], or if the measurements are comparative, up-sampling can be accomplished by padding with zeros. We do padding in CNN layers. From the RNN re-ID techniques it was observed that to enhance the re-ID accuracy, joining of Siamese cost with model is required. From each stream we have two Siamese cost functions, though RNN models have just a single Siamese cost. Thusly, we characterize the joined cost function J_f as equation (10):

$$J_f = \omega_s D\left(f_{i_{S_c}}^-, f_{j_{S_c}}^-\right) + \omega_T D\left(f_{i_{T_c}}^-, f_{j_{T_c}}^-\right) \quad (10)$$

ω_s, ω_T are the weights for Spatial-Net and Temporal-Net. We propose utilizing unique weights for each stream to have the capacity to underline the spatial features when contrasted with the optical features For re-ID, despite the fact that motions features includes discriminative capacity to the accuracy but spatial features for example, appearance, shading or surface are moderately more critical as far as re-recognizing individuals. Hence, we set the weights observationally with the condition $\omega_s \geq \omega_T$.

3.2 Model Details

This segment presents usage insights with respect to the preparation of the created systems.

3.2.1 Setup

The systems will be prepared and tried on two personal computers, they are both outfitted with NVIDIA GPUs for speedier preparing of the systems. The particulars of the PCs can be seen beneath in Table 2. The structure utilized, for planning the systems, is Keras [97] with Tensor Flow [98] as backend. Keeping in mind the end goal to change the pixel force esteems into a predefined extend each picture I_i was changed by equation (11):

$$I^*{}_i = \frac{I_i}{\max(I_i)} \quad (11)$$

The rescaled picture $I^*{}_i$ winds up changed into the regularly utilized interim [0, 1]. The ADAM enhancer was utilized to upgrade the weights of the systems. A learning rate of 10^{-3} was at first utilized. The learning rate was diminished directly as indicated by equation (12):

$$a_t = (1-\epsilon)a_0 + \epsilon a_\tau \quad (12)$$

Where $\epsilon = \frac{t}{\tau}$, a_τ the static learning rate and α0 is the underlying learning rate. The taking in rate will directly diminish from α0 to ατ as t → τ. usually, the learning rate is kept consistent when t > τ. The diverse systems were prepared for a settled measure of emphases t_{MAX}, contingent upon the extent of the system. The quantity of emphases for the diverse systems are presented additionally down. So as to moderate the danger of over-fitting, the execution on the approval dataset was checked, the system was spared persistently as the execution expanded on this dataset. The system was approved in ventures of 200 cycles amid preparing. The systems were likewise spared naturally in augmentations of $t_{MAX}/5$. Toward the finish of the preparation, the re-distinguishing proof.

Execution was estimated all in all approval dataset, the best spared arrange was picked. The best systems for every decision of misfortune work were likewise prepared on expanded pictures, bringing about broadened preparing datasets and expanded number of preparing cycles. Information growth modifications were connected to all pictures in the preparation datasets. Two diverse picture increase tasks were utilized, focus trimming and flat flipping of the picture. The trimming was led through lessening of the pictures, an inside picture with estimate traversing from (144×54) to (128 X 64) was decided for every case in the group. The lessened picture was from that point rescaled to the standard size (128 X 64).

Table 2: System Specifications.

Computer 1:	**Computer 2:**
➢ CPU: Intel I7-3770k	➢ CPU: Intel I5-6500
➢ RAM: 24 GB	➢ RAM: 32 GB
➢ GPU 1: NVIDIA GTX 1070 8 GB	➢ GPU 1: NVIDIA GTX 1060 6 GB
➢ GPU 2: NVIDIA GTX 1060 6 GB	➢ GPU 2: NVIDIA GTX 1060 3 GB

3.2.2 Input

We express the input video sequence as V_s, s belongs to **a, b** used for camera A and camera B respectively. Input for first stream of our model are RGB frames: $V_s = (R^{(1)}. . . R^{(L)})$ and for second stream optical flow images are used, $V_s = (T^{(1)}. . . T^{(L)})$. Here L is length of input sequence. Lucas-Kanade technique [58] is used to compute optical flows. Then by utilizing the convolutional network depict in Fig 7; we obtain feature maps set $C_s = (C^{(1)}. . . C^{(L)})$ and $C_t = (O^{(1)}. . . O^{(L)})$ for RGB frames and for Optical flow respectively for input. The use of Optical flow will increase the efficiency to learn temporal features as present in [5, 17].

The crude info pictures are first resized to 128 X 64 pixels. On the off chance that the picture is littler than 128 X 64 pixels at that point zero cushioning is added to fill the holes. On the off chance that the picture is bigger than 128 X 64 pixels then the picture is re-scaled. Since the width-stature proportion in generally individual re-recognizable proof is comparable, re-scaling the bigger pictures won't prompt abnormal body extents.

The pictures are gone through a cluster standardization layer to ensure that it is standardized to have 0 mean and a standard deviation of 1. We utilize the tanh as our nonlinearity since we found that observationally it beats the all the more usually utilized Exponential Linear Unit (ELU) [99]. A convolutional unit comprises of a convolutional layer with a walk of 1. Since the pictures are little, a little 3 X 3 bit is utilized to catch more detail. This is trailed by a maximum pooling activity with a 2 X 2 piece, trailed by the tanh lastly group standardization is connected to decrease inside covariate move as the highlights are proliferated through the branch. Notice that in the primary convolutional unit a maximum pooling activity of 4 X 2 is utilized. This was done to have a one-dimensional component vector as the yield of convolutional unit 6. The two branches of the siamese system share the same weights. Each branch yields a one-dimensional 512 component highlight vector. These are consolidated utilizing a component astute subtraction and after that by taking the outright esteem per component, bringing about the supreme contrast include vector. This vector is at that point passed on to the similitude learning piece. The likeness learning square comprises of an arrangement of FC layers joined with a dropout of 0:5 between the layers to give regularization. Toward the end we have a yield layer comprising of two one-hot-encoded yields: [0; 1] for coordinate and [1; 0] for a bungle. The forecasts are gone through the softmax work to standardize the yield: Prediction P = [p1; p2] and if p2 > 0:5 then the system predicts that the combine is a match.

3.2.3 Training

Our training set comprises of named picture sets, demonstrating whether a couple of pictures have a place to a similar character. A mark of [0; 1] showing that the given combine is a match, or a [1; 0] demonstrating that the given match is a crisscross. By doing this we cast the re-ID issue as a grouping issue comprising of two classes: match or confound. All systems are prepared for 1000 epochs. A characteristic class irregularity emerges in the information since we work with sets of pictures to prepare the model. This outcomes in a considerably higher number of negative (befuddling) picture sets contrasted with the quantity of positive (coordinating) picture sets. When we influence sets of pictures we to get the accompanying:

$$p = \frac{(IC)^2 - IC}{2} \tag{13}$$

$$p_{pos} = I\frac{C^2 - C}{2} \tag{14}$$

$$p_{neg} = p - p_{pos} \tag{15}$$

$$r_{p,n} = \frac{C-1}{C(I-1)} \tag{16}$$

Where p is the aggregate number of sets conceivable, C is the quantity of cameras in a system, I is the quantity of IDs in the dataset obvious on all cameras, p_{pos} is the quantity of positive sets

conceivable and $p_{neg}$is the quantity of negative sets conceivable. When we think about the proportion amongst positive and negative sets we can see plainly that there is a colossal lopsidedness, see Table 3 which was figured from equations 13- 16.

The regular method to handle this issue is to utilize information expansion procedures to enlarge the quantity of positive sets, for example, reflecting, pivoting and zooming. Anyway we needed to perceive how well the datasets would perform alone, without utilizing increase so we didn't utilize information enlargement. To beat this lopsidedness we perform under sampling in the negative class: For each preparing age, the arrangement of negative sets is rearranged and an arbitrary subset is examined from the arrangement of negative picture sets. This examined subset has an indistinguishable number of sets from the number of sets in the arrangement of positive picture sets, influencing the last preparing to set similarly adjusted for the two classes.

Note that adjusting the information implies that we disregard the earlier conviction that the event of a match is an uncommon occasion. Rather, the subsequent model accept that a match happens with meet recurrence as a crisscross. On the off chance that we were utilizing the neural system as an independent identifier, at that point it would bode well to not adjust the information 50-50, since in the utmost, we would get a sensor demonstrate $P(Z_t|x_t)$ forced with the earlier conviction that an event of a match is uncommon, ideal for this present reality utilize case. Be that as it may, the neural system is utilized as part of a bigger estimation process, where the priors of a particular individual being available at the sensor is figured. $P(x_t|z1:t-1)$ Can be seen as the earlier conviction that the neural system has identified a match at the area of the sensor at time t. Since the earlier is unequivocally figured, the probability given by the sensor display $P(Z_t|x_t)$ ought to accept that a match happens with parallel recurrence as befuddle.

Table 3: Increase in number of IDs cause increase in class imbalance.

Camera =1				
IDs	**Total pairs**	**Positive pairs**	**Negative pairs**	*rp,n* **%**
10	190	10	180	5.56
50	4950	50	4900	1.02
100	19900	100	19800	0.51
500	499500	500	499000	0.10
1000	1999000	1000	1998000	0.05

3.2.4 Testing

We used evaluation protocol same as in [1]. For all three data sets (iLIDS-VID, PRID2011 and MARS) we randomly split each data set part into two subsets with a similar size. One is utilized for training the model and other is used for testing. For the testing, the arrangements from the principal camera are utilized as the tests while the successions from the second camera are utilized as the gallery. We approve the execution of our proposed technique furthermore, look at the execution against different techniques utilizing the Cumulative Matching Characteristic

(CMC) bend which demonstrates the likelihood of finding the right match in the best N ranks. The test is rehashed ten times by randomly split the dataset into preparing and testing and the normal outcome is accounted for. In our proposed strategy, we have two additional hyper-parameters (ω_T, ω_S). To see the adequacy of proposed strategy, we perform tries different things with different hyper-parameters settings. We perform tries different things with ω_T= 1 at the point when ω_S is set to 0 or 1 with a specific end goal to confirm the individual commitment of Temporal-Net. We likewise perform tests with ω_S = 2, 3 when ω_T = 1 to see the relative commitment of the spatial highlights when contrasted with the worldly highlights.

We also test our model by different experiments. In first we give only RGB images to our model and perform training and testing. In second experiment we only give optical flows as input to our model and in last third experiment we give both RGB to one stream and optical flows to second stream. Our results show that best result can be produce by experiment 3, which we discussed in detail in our next chapter.

Chapter 4

Experiment Results

In this portion, we assess and compare execution of our proposed novel procedure for video-based individual re-ID with cutting edge models for three datasets: PRID-2011[100], iLIDS-VID [101] and MARS [102]. We also examine how the two stream network with attentive temporal pooling and other collective spatial pooling bring advantage to the proposed model. Table 4 show characteristics of datasets used for training and testing in this thesis.

Table 4: MARS, iLIDS-VID and PRID-2011 dataset statistics.

Dataset	MARS	iLIDS	PRID
#ID	1,261	300	200
#tracklets	20,478	600	400
#bboxes	1,191,003	43,800	40k
#distractors	3,248	0	0
#cam./ID	6	2	2
Produced by	DPM+GMMCP	hand	hand
Evaluation	mAP+CMC	CMC	CMC
Links	http://www.liangzheng.com.cn/Project/project_mars.html	http://www.eecs.qmul.ac.uk/~xiatian/downloads_qmul_iLIDS-VID_ReID_dataset.html	https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/PRID11/

4.1 Datasets

Numerous customary benchmark datasets mirror a 'shut world' situation, like precisely two cameras sees through precisely one example of every individual per camera and one to one correct character correspondence among the cameras. It is as opposed to a more reasonable 'open-world' situation, where people in every camera might be just in part covering and the amount of cameras, spatial magnitude of the earth and number of individuals may be obscure and at an essentially bigger gauge. In this way the pursuit space is of obscure size and contains a possibly boundless number of contender matches for object. Re-distinguishing proof of focuses in such open conditions can conceivably scale to self-assertive levels, casing immense spatial

territories traversing not simply extraordinary structures but rather diverse urban communities, nations or even main lands, prompting a staggering amount of 'enormous information'.

4.1.1 iLIDS-VID

The iLIDS-VID dataset consist of 600 images of 300 pedestrians, where every single individual is represented by two sequences taken at arrival hall of airport via two separate camera. The dataset consists of static images and sequences images. Lighting variations and occlusions from different viewpoints was considered in this dataset. Each image is scaled to 128 x 64 pixels. Size of sequences differs from 23 to 192 frames, with a normal size of 73. It is very challenging dataset, because of dress resemblances for different persons, change in illumination, viewpoint deviations and arbitrary occlusions shown in Fig 9.

Figure 9: Sample images of a single person from the iLIDS-VID dataset [103].

4.1.2 PRID-2011

PRID-2011 dataset comprises of 749 folks, taken by two disjoint cameras. The dimension of frames differs from 5 to 675 for each person image sequence, with 100 sequences as average. It has simple backgrounds and less occlusions as compared with the iLIDS-VID dataset. As in [17], we utilized only first 200 individuals captured by both cameras. It contains many images of each pedestrian, from two camera views looking at a street crossing as shown in Fig 10.

Figure 10: Example images of PRID-2011 dataset for a single person [100].

4.1.3 MARS

MARS is considered as the largest dataset for video-based person re-ID. It contain 1261 different persons, captured by 2-6 cameras and on average each person has 13 sequences. Each person has at least two image sequences, which are automatically acquired by using DPM detector and GMMCP tracker. It is very challenging dataset, because of dress resemblances for different persons, change in illumination, viewpoint deviations and arbitrary occlusions shown in Fig 11. The first three rows each corresponds to an identity, and tracklets in each column belong to different cameras. The last row presents four examples of false detection and tracking results.

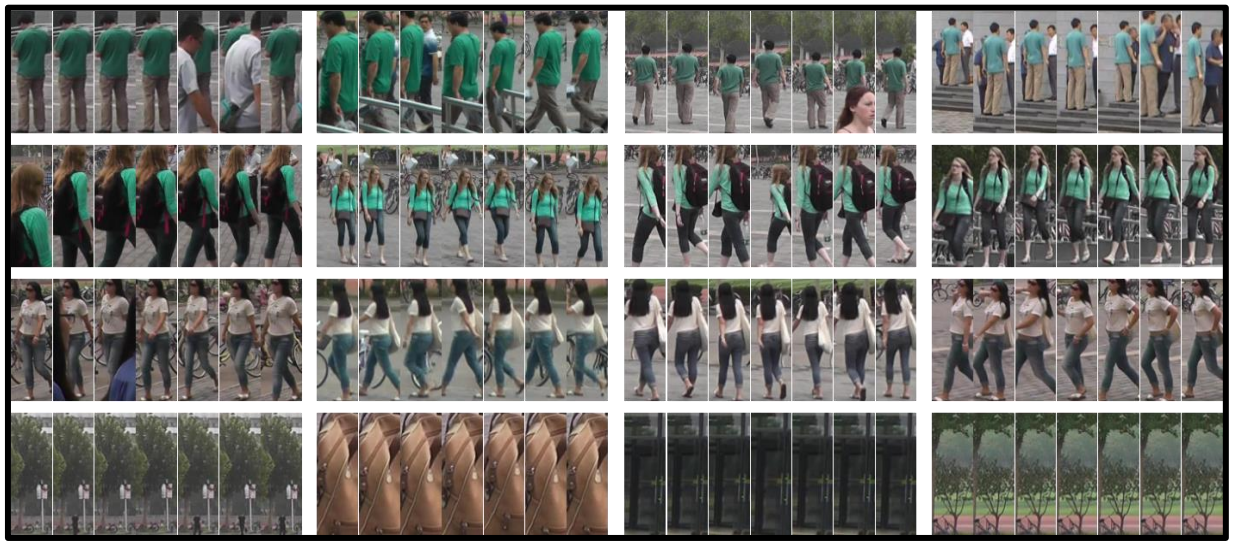

Figure 11: Example of tracklets in MARS dataset. [23].

4.2 Performance Measures

There are an assortment of measurements that are helpful for evaluating the adequacy of a re-ID framework. The two most regular measurements are 'Rank-1 precision', and the 'CMC bend'. Rank-1 exactness alludes to the regular thought of classification exactness: the rate of test pictures which are flawlessly coordinated to their relating exhibition picture. High Rank-1 exactness is famously difficult to get on testing re-id issues.

All the more reasonably a model is required to report a positioned rundown of matches which the administrator can investigate physically to affirm the genuine match. The question is the means by which high genuine matches ordinarily show up on the positioned list. The CMC (Combined Match Trademark) bend abridges this: the possibility of the genuine coordinate showing up in the main 1, 2. . . . N of the positioned list (the primary point on the CMC bend being Rank-1 precision). Different measurements which can be gotten from the CMC bend incorporate the scalar region under the bend, and expected rank (by and large how far down the rundown is the genuine match). Which of these two measurements is the most important apparently relies on upon the particular application situation: Regardless of whether a (likely low in supreme terms) possibility of impeccable match or a decent normal positioning is favored.

This polarity brings up the further fascinating issue of which assessment rule is the significant one to streamline when outlining discriminatively prepared re-recognizable proof models.

4.2.1 Cumulative Matching Characteristic Curve

Cumulative matching characteristic (CMC) curve is utilized in person re-ID methodologies to evaluate how well the 1: n identifying ability of the current system is. CMC curves measures ranking power of re-id techniques. Let's assume a gallery set $G = \{g_1, g_2, g_3 \ldots g_n\}$ with n discriminating feature descriptors and a probe set $P = \{p_1, p_2, p_3 \ldots pm\}$ with m descriptors. Given query $Q \in P$ and person descriptor $g_1 \in G$, the matching algorithm assign a score $S(Q, g_1)$ on similarity basis. In similar way, every probe entry is matched and scored with gallery entries to compute $n \times m$ CMC curves. All the scores are then ranked in an ascending order that is from least score to highest. Fig 12 illustrates CMC curves of two different models, where performance of model 1 (which is in blue color) is better than model 2 (in red color) depending on their ranking ability.

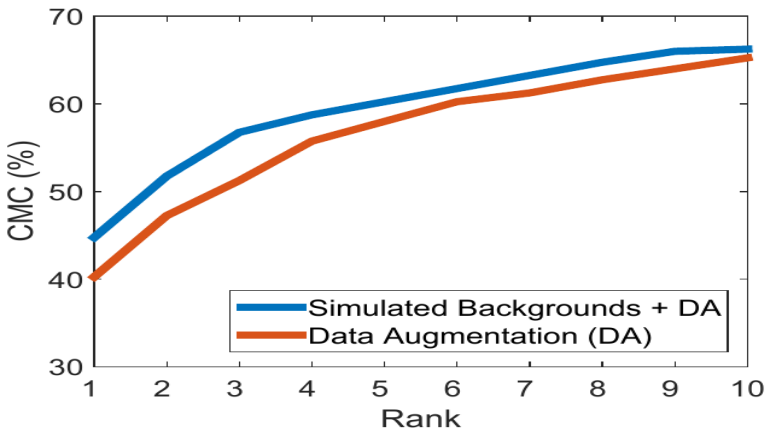

Figure 12: CMC curves of two systems [104] .

4.2.2 Area under the Curve

Area under the curve (AUC) is another execution metric used to assess execution of twofold classifiers. In a paired classifiers, the yield is in positive or negative and the yield is generally represented in a table called disarray framework. For the most part, the ROC (Receiver Operating trademark) bends are utilized to gauge exhibitions in machine learning condition. The zone under the ROC bends give us insights about how well a classifier is and furthermore give us the chance to think about the capacities of numerous classifiers. Then again, AUC is a likelihood that the chose positive passage will be positioned higher in contrast with negative section by the classifier. In procedure of highlight coordinating, standardized AUC is utilized where full region secured by CMC bend is considered. This creates a general score on how great the procedures perform.

4.2.3 Mean Average Precision

Mean average precision (mAP) is figured alongside CMC bends for better comprehension of calculation re-recognizing capacity. The mAP is utilized to assess re-id framework execution with expect to display all the positive matches to the administrator. Let's say that the two systems have equal capabilities of identifying the ground truth but retrieving orders are different. In such a case, mAP has more discriminating power which is lost in CMC curve as shown in Fig 13. Given a query image $q_n \in Q$ first R pictures are recovered from exhibition set. Accuracy is registered on these recovered pictures for each question in Q and after that normal is figured over n inquiries to get a mean normal single score called mAP. In all the ranked list CMC is 1 but mAP is 1, 1 and 0.41 depending on location of correct match.

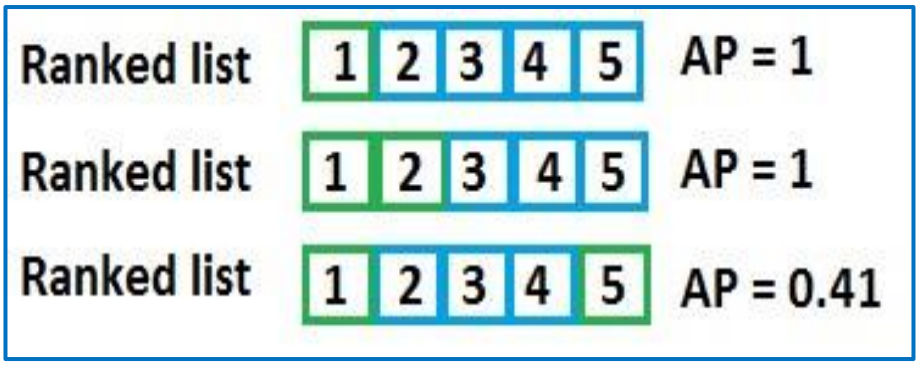

Figure 13: Ranked List [105]

4.3 Experiment Settings

Similar to the protocol used for evaluation in [17], we randomly select two cameras of the same person, where the case was condensed to experiences with PRID-2011 and iLIDS-VID. Then we follow [1], where we randomly divided the whole set of persons sequence pairs of MARS, PRID-2011 and iLIDS-VID into two subsets of the same size. One subset is utilized for training model, and the other is utilized for testing.

For all three datasets, experiments were repeat 10 times with altered test and train splits to ensure constant results. Section 4.4 show the average CMC (Cumulative Matching Characteristics) curves for our results. Preprocessing was done in several following forms [1]: i) All probe and gallery pictures were changed over to YUV shading space before being passed to the model. Each of the three shading channel were standardized to zero mean and unit variance. ii) Lucas-Kanade method [106] is used for computing both horizontal and vertical optical flow of each pair of attached images, and then we standardized all optical stream channels to the range - 1 to 1. iii) When the model was trained with stochastic gradient descent, learning rate was set at 0.001. Introduction of hyper-parameters of our model was performed in light of our base paper [1], which is adjusted on the challenging VIPeR individual re-ID dataset.

Data augmentation include the following various forms: Initially, since we have the gallery and probe sequences are of different size, so for training and testing we fixed the length of each person sequence to 16 and 128 respectively for the fairness of experimentations. Furthermore,

positive pair was made out of a sub-grouping from camera 1 and camera 2 containing a similar individual A, and antagonistic match was made out of a sub-succession from camera 1 of individual A and a sub-arrangement from camera 2 of individual B, who was named arbitrarily from the rest of the general population in preparing set. Then again, these positive and negative arrangement sets were sent to our framework, with the goal that model is sufficiently skilled to separate between right and wrong match, and a full epoch contained same number of positive and negative groupings. Lastly, we set margin to 4 and learning rate to 2e-3 for training Siamese network. Then after 500^{th} epoch we multiply it with 0.5, and finish the training process at 1000^{th} epoch. For 150 persons, training takes about half day using the Nvidia GPU. After training phase we have stored features vectors for all gallery sequences. For testing person re-ID can then be done more efficiently, because only the new sequence of probe person passed through the model to yield a feature vector. Less than one second is required to match probe feature vector with gallery features vectors using a single matrix vector product.

4.4 Result for iLIDS-VID

4.4.1 Sample Results

One query picture of one individual and the five most plausible matches among 100 people in the display. The rundown is arranged in plunging request, from the most likely to more improbable are shown in Fig 14-16. Correct result shown in green. For the situation found in Fig 14, the right individual is matched in rank1. It depict that **proposed model perform best for occluded persons**, also other rank results shows that our model is sensitive to color.

Figure 14: iLIDS-VID best result sample for proposed model.

One can unmistakably see in Fig 15 and 16 that the system has figured out how to discover comparable looking individuals. Unmistakably, dress is one basic characteristic the system makes utilization of when a man is re-distinguished. In the second case, showed by Fig 15, the correct individual has the second biggest likeness score and correct matched found at rank2. For the situation found in Fig 15 all people among the five probably have comparable dark dress. Separately, the two people are re-related to rank one and two. From Fig 15, we can say that

proposed model best for pose variation as well, because correct matched is also found at rank3 with different pose.

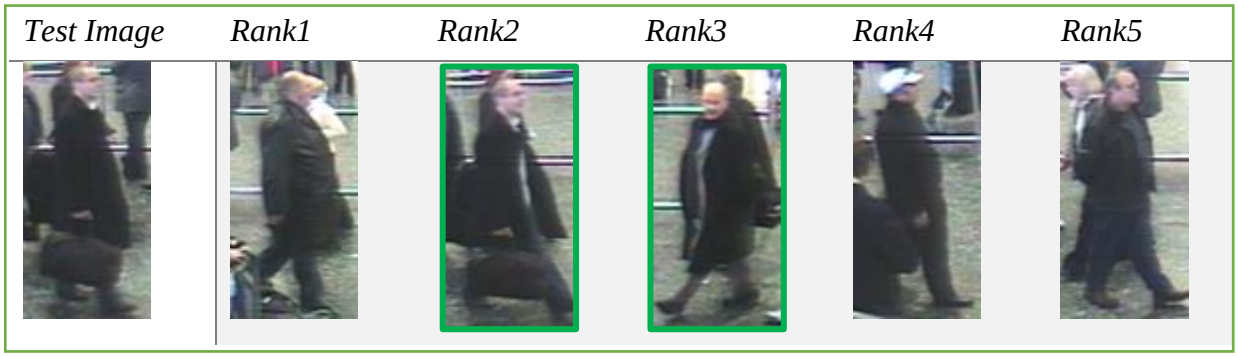

Figure 15: iLIDS-VID average result sample for proposed model.

Figure 16: iLIDS-VID worst result sample for proposed model.

Fig 16 shows one re-recognizable proof situation when the right individual isn't among the five most plausible. In any case, the most plausible people all have comparative qualities as the test individual. The test individual was re-related to rank 10. Fig 16 demonstrates certain situations where the system experienced difficulty re-recognizing the people, because all are men and wear same color jacket and carry a bag with them. The re-ID cases give a thought of the conditions the system find troublesome.

4.4.2 State-of-the-art comparison

In this section, we compare the results of our proposed architecture with state-of-the-art models for video-based re-ID. Our proposed model accomplish the best performance against state-of-the-art models for iLIDS-VID datasets as shown in Table 5. We train and test proposed model with different inputs for iLIDS-VID. Firstly, we give only RGB images to proposed model and it gives us better performance. Secondly, we test our model by giving only optical flow and it shows relatively better performance as compares to first experiment. Lastly, we test model by giving RGB and optical flow as input and it gives us superior performance as shown in Fig 17.

The CMC results of our model with other models are compared in Table 5. With old algorithm of optical flow, our novel proposed two-stream multi-scale model can achieve a matching rate of rank-1 of about 72.6% on iLIDS-VID dataset, which is greater than all other methods. The performance can be more improved when we use both optical and color images as input. The improvements are 2.7% and 2.2% for rank-1 and rank-5 respectively.

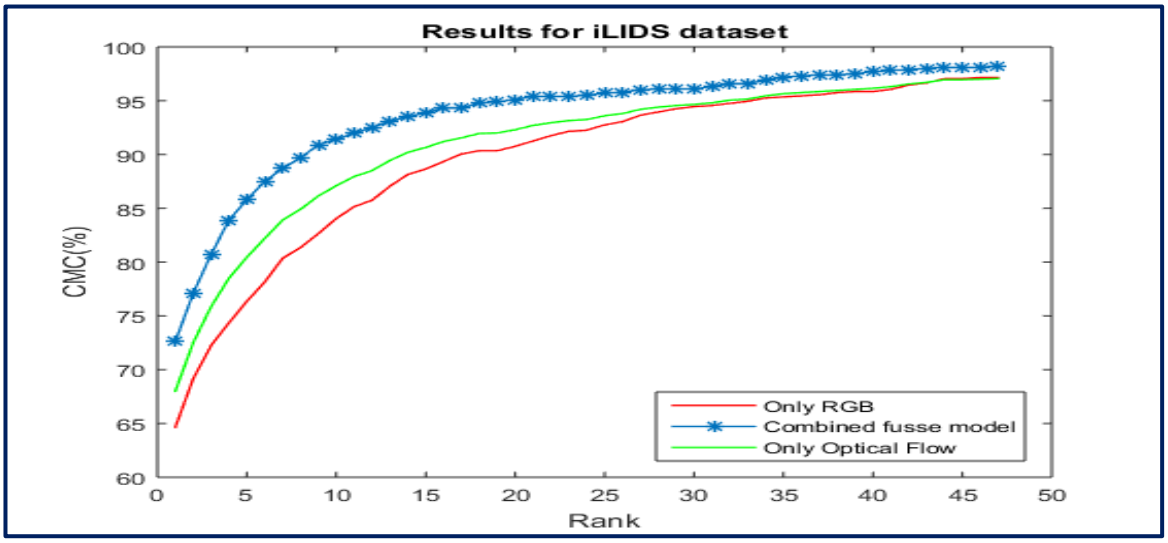

Figure 17: Proposed model CMC curve for iLIDS-VID dataset.

4.5 Result for PRID-2011

4.5.1 Sample Results

Distinctive re-recognizable test cases from PRID-2011 dataset are shown in Fig 18-20 below. Each succession is spoken to by their first picture. The grouping length utilized amid testing was 80. To one side is the test arrangements spoke to. To the privilege are the ten most plausible matches among 55 people in the exhibition. The rundown is arranged in slipping request, from the most likely to more improbable. Over each picture is a similitude score, scores set apart in green show amend matches.

Figure 18: PRID-2011 best result sample for proposed model.

In the primary case, outlined by Fig 18, the system has discovered people with comparative white dress as the test individual. The test individual was re-related to rank 1. In the second case, represented by Fig 19, the five in all probability people have comparative blue coats as the test individual. The test individual was re-related to rank 3. Also as in the single shot re-recognizable proof case, the multi shot systems make utilization of shade of apparel when people are re-recognized. The case represented by Fig 20 demonstrates a case when the test individual was not re-distinguished among the five in all probability people. All people among the top five probably have low likeness scores, because test image is man and due to occlusion its look like a woman with long hairs, so in results its shows all images for women, and correct matched was found at rank 14.

Figure 19: PRID-2011 average result sample for proposed model.

Figure 20: PRID-2011 worst result sample for proposed model.

4.5.2 State-of-the-art comparison

Our proposed model accomplish the best performance against state-of-the-art models for PRID-2011 datasets as shown in Table 5. We train and test proposed model with different inputs for PRID-2011. Firstly, we give only RGB images to proposed model and it gives us better performance. Secondly, we test our model by giving only optical flow and it shows relatively better performance as compares to first experiment. Lastly, we test model by giving both RGB and optical flow as input and it gives us superior performance as shown in Fig 21.

For PRID-2011 dataset, our mode beats other methods, with rank-1 accuracy of 84.0%. Our architecture still beats other models prominently as shown from Table 5, with rank-1 accuracy achieving 84.0% surpassing the RNN-CNN model by 14%. Moreover, our system is more competent and robust as its CMC rank rate reaches 97% at level of rank 5 and further goes up to the meeting 99.2% at rank 10. The tendency of accuracy exhibits that our model is an efficient space and time feature extractor, capable to acquire more discriminative sequence-level person representation over learning process.

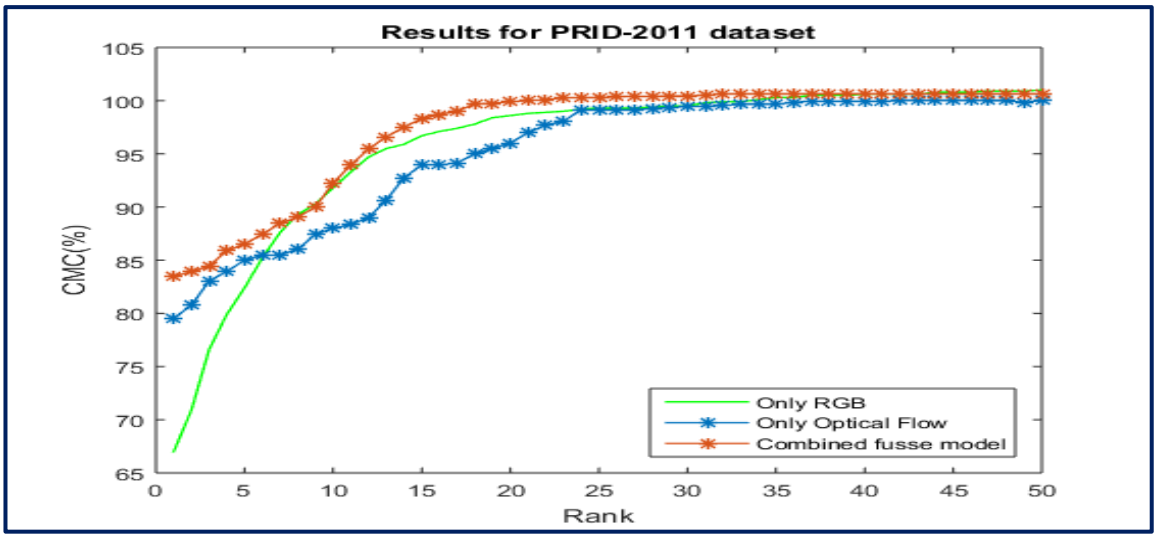

Figure 21: Proposed model CMC curve for PRID-2011 dataset.

4.6 Result for MARS

4.6.1 Sample Results

The bounding boxes for the MARS dataset are discovered utilizing the Deformable Parts Model, which implies the jumping boxes can differ fundamentally from casing to outline. Because of the programmed jumping encloses the MARS dataset it is utilized the PRID2011 dataset the bouncing boxes from the PRID2011 dataset are made by hand, and in this way does not have similar issues with bounces in jumping box sizes. It is thusly chosen to prepare on the PRID2011 dataset with an indistinguishable settings from above.

Fig 22 shows best result computed for MARS data set. Probe person is correctly recognized at rank one as well as on rank two. Further result images are quite similar to test image, because shirt color is changed for one person as well due to different lighting conditions. Similarly, in Fig 23, test image is matched at rank 2, but all images at other ranks are very similar shirt and half sleeves and shorts with same type of sleepers. Proposed model showed all result images are pretty much closed to each other. Which shows that proposed model perform best on such big data set.

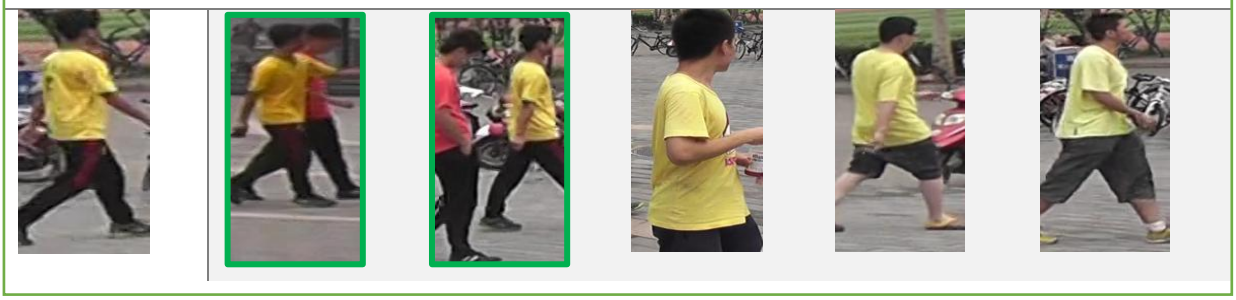

Figure 22: MARS best result sample for proposed model.

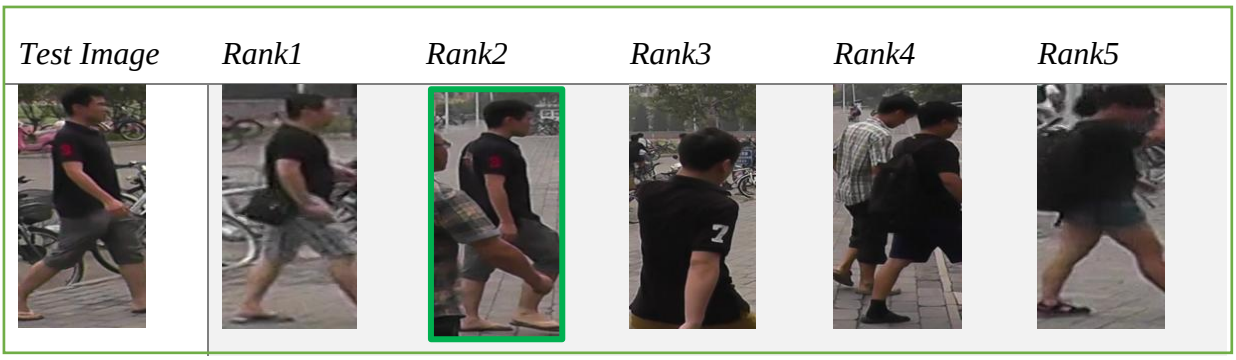

Figure 23: MARS average result sample for proposed model.

MARS is very challenging dataset, because of dress resemblances for different persons, change in illumination. Apart from that the bouncing boxes for the MARS dataset are discovered utilizing the Deformable Parts Model, which implies the jumping boxes can differ fundamentally from casing to outline, this can be found in our result shown in Fig 24. Correct matched was found at 11 rank, because all result images at top five ranks look very similar to test image. Specially result image at rank 1 due to different bounding boxes.

Figure 24: MARS worst result sample for proposed model.

4.6.2 State-of-the-art comparison

We also perform experiments on realistic and large MARS dataset, to further evaluate the proposed model as shown in Fig 25. Table 6 show the performances comparison of our model with ASTPN and RNN-CNN. As we can see from Table 6, our proposed model attains the best accuracy. It again demonstrates the effectiveness of our proposed model that uses different streams to yield full use of feature maps. As our two stream model has less parameters so it can be trained very well for less challenging dataset. The main cause for such good performance of our model is that different multi-streams emphasis on different characteristics of the feature maps and each stream help other to learn some union features.

Result for MARS dataset in Table 6 shows that proposed model perform 16 times greater than our base model at rank 1; similarly 3%, 7% and 13% greater than base model for rank 5, 10 and 20 respectively.

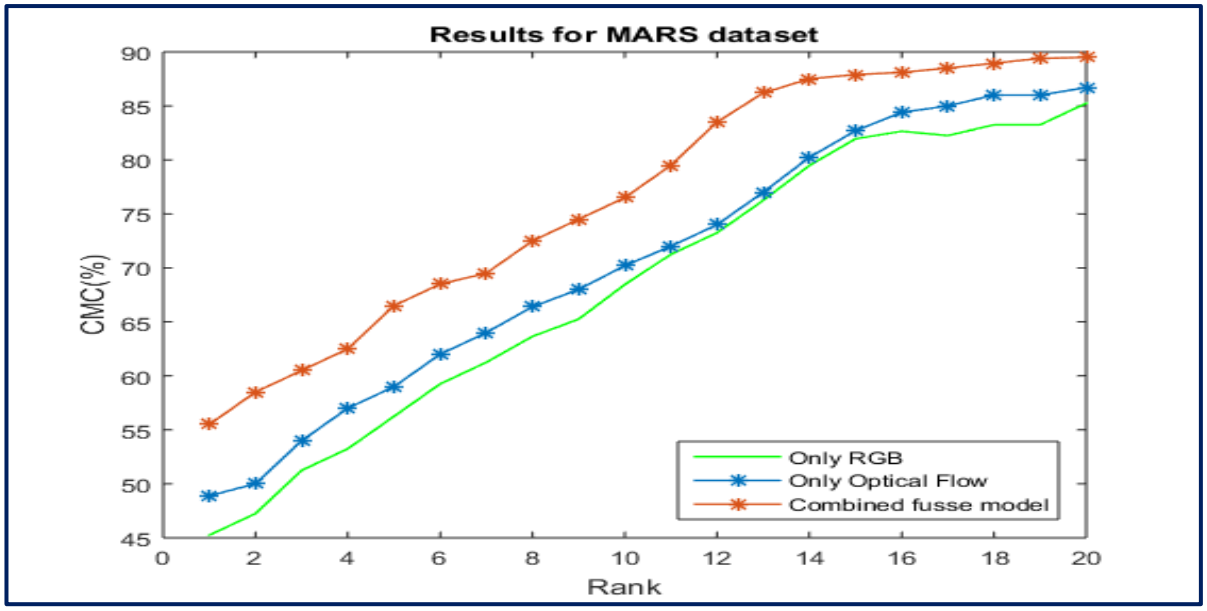

Figure 25: Proposed model CMC curve for MARS dataset.

Table 5: Evaluation of our approaches with other state-of-the-art methods on iLIDS-VID and PRID-2011.

Methods	**Year**	**Dataset**							
		iLIDS-VID				**PRID-2011**			
		Rank 1	**Rank 5**	**Rank 10**	**Rank 20**	**Rank 1**	**Rank 5**	**Rank 10**	**Rank 20**
STA [84]	2015	44.3	71.7	83.7	91.7	64.1	87.3	89.9	92.0
DVR [28]	2014	39.5	61.1	71.7	81.8	40.0	71.7	84.5	92.2
SRID [77]	2015	24.9	44.5	55.6	66.2	35.1	59.4	69.8	79.7
AFDA [87]	2015	37.5	62.7	73.0	81.8	43.0	72.7	84.6	91.9
DVDL [88]	2015	25.9	48.2	57.3	68.9	40.6	69.7	77.8	85.6
CNN-RNN [1]	2016	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0
CNN-BRNN [26]	2017	55.3	85.0	91.7	95.1	72.8	92.0	95.1	97.6
ASTPN [17]	2017	62.0	86.0	94.0	98.0	77.0	95.0	99.0	99.0
AMOC + Epic-Flow [29]	2017	68.7	94.3	98.3	99.3	83.7	98.3	99.4	100
AMOC + LK-Flow [29]	2017	65.3	87.3	96.1	98.4	78.0	97.2	99.1	99.7
Multi-TSCN [5]	2017	67.5	90.4	97.2	98.6	78.8	96.7	99.1	99.6
RNNPR [2]	2018	58.0	87.5	93.7	97.5	76.4	95.3	98.0	99.1
Our Multi stream with RGB	-	**64.6**	**76.5**	**84.6**	**90.7**	**65.7**	**81.5**	**90.5**	**97.3**
Our Multi stream with Optical Flow	-	**68**	**80.4**	**87.2**	**92.6**	**80.8**	**90.7**	**92.8**	**95.6**
Our Combined Multi stream	-	**72.6**	**85.8**	**95.2**	**97.5**	**84.0**	**97.5**	**99.2**	**100**

Table 6: MARS dataset results.

Methods	**Year**	**MARS Dataset**			
		Rank = 1	**Rank = 5**	**Rank =10**	**Rank = 20**
CNN-RNN [1]	2016	40.0	64.0	70.0	77.0
MARS: A Video Benchmark for Large-Scale Person Re-identification [23]	2016	30.6	46.2	59.2	15.5
ASTPN [17]	2017	44.0	70.0	74.0	81.0
Multi-TSCN [5]	2017	45.6	72.4	75.4	82.6
LCAR [107]	2017	55.5	70.2	-	80.2
Our Multi stream with RGB	-	45.2	56.2	73.2	85
Our Multi stream with Optical Flow	-	45.9	57	67	83.7
Our Combined Multi stream	**-**	**56**	**67**	**77**	**90**

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we present a novel deep end-to-end two stream CNN-RNN architecture with attentive pooling (TSCRA). Which use two separate streams for RGB frames and optical flows, and learn feature maps with different aspects from CNN. Then these features fed to RNN for extracting temporal information. Finally to obtain some union features from two stream, we merge output of both stream and pass it to attentive pooling. Only informative frames over full sequence are selected through attentive pooling. We also demonstrate that max/average pooling and fully connected layer is not necessary for two stream model.

The comprehensive tests on PRID-2011, iLIDS-VID and MARS datasets, our result show that proposed novel model achieve greater performance to the existing state-of-the-art re-ID models. Re-id accuracy has been improved on PRID-2011, iLIDS-VID and MARS datasets to 84.0%, 72.6% and 56.0%respectively. Our architecture (TSCRA) is easy and simple to implement. It require little more computational time as compared to single stream networks, but provide superior performance on real, versatile and large datasets. Which make it most desirable model for re-ID in future.

We perform training and testing on iLIDS-VID, MARS and PRID-2011 dataset because, (i) these three datasets are based on multi-shot scenario (ii) person images sequences in these datasets contain a lot of occlusion, viewpoint and illumination changes (iii) indoor and outdoor both scenario are available in these datasets (iv) these datasets are large as compare to other person re-ID datasets, hence these model are best for deep CNN models (v) our base model and other CNN based models used only these three datasets, so we can compare our results with these models.

Individual re-recognizable proof, predicted in the most established stories, is increasing broad enthusiasm for the present day academic group. In this thesis, a study of individual re-distinguishing proof is introduced. Initial, a concise history of individual re-identification is presented and its likenesses and contrasts to picture characterization and case recovery are portrayed. At that point, existing picture and video-based techniques are looked into, which are arranged into hand-made and profoundly learned frameworks. Situated in between picture arrangement and occasion recovery, individual re-identification has far from turning into a precise and effective application. Accordingly, leaving from past overviews, this paper puts more accentuation on the immature yet basic future conceivable outcomes, for example, the end-to-end re-identification frameworks that incorporate walker location and following, and individual

re-identification in huge exhibitions, which we accept are fundamental strides toward reasonable frameworks.

We likewise highlight some critical open issues that may pull in further consideration from the community. They incorporate comprehending the information volume issue, re-identification re-positioning techniques, and open re-identification frameworks. With everything taken into account, the coordination of discriminative component learning, indicator/following streamlining, and productive information structures will prompt an effective individual re-recognizable proof framework.

5.2 Future Work

As a result of the sheer measure of various recurrent units that were assessed in this study, other system designs other than the Siamese classifier organize engineering were not tried for person re-ID. Additionally triplet network together with the triplet soft max misfortune demonstrated guarantee in the single shot case. Perhaps would such a system yield tantamount or far superior execution contrasted with the systems utilized as a part of this investigation. As of late a few endeavors have been made to tackle end-to-end re-id. The conclusion to-end characterizes a re-id framework that work with crude recordings rather than specifically working with pictures with distinguished bouncing box. The methodologies proposed give break even with significance to location step like each other advance of re-id assignment. The execution of such methodologies are additionally subject to nature of identifier, better the location higher will be the execution utilizing same list of capabilities.

While tests have just been directed on shut set re-id datasets, a similar descriptor ought to perform well in an open set setting too. For the future, multi-modular individual re-id will be performed, joining anthropometric measures and warm highlights. Planning to investigate their effect on exactness rates and assessing their combination with visual descriptors. The domain of person re-id is moving towards more deep learning approach with combination of multi streams used for upper, lower, fully body etc. and it should be since their performance is better than distance or metric learning approaches, provided large training set. So, the future research should be focused on improving the efficiency and effectiveness of human re-id methods on massive datasets unlike the current datasets.

As of recently the greater part of the examination is centered on enhancing two parts of re-id, descriptor age and metric learning. Accepting exact jumping boxes are accessible, that are either handmade or distinguished through machine learning techniques. Be that as it may, both the presumptions are false. With headway in look into exhibition sizes are winding up expansive and making bouncing boxes physically is tedious. If there should arise an occurrence of programmed recognition, the identifiers are not proficient and rely upon limit esteems. The low limit esteem identify undesirable expansive number of jumping boxes and the other route round.

References

[1] N. McLaughlin, J. M. del Rincon, and P. Miller. (2016). *Recurrent convolutional network for video-based person re-identification*. Available: https://github.com/niallmcl/Recurrent-Convolutional-Video-ReID

[2] J.-B. 1Boin, A. Araujo, and B. Girod, "Recurrent Neural Networks for Person Re-identification Revisited," *arXiv preprint arXiv:1804.03281,* 2018.

[3] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded Person Re-identification," *arXiv preprint arXiv:1804.02792,* 2018.

[4] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007, pp. 1-7.

[5] Z. Yu, T. Li, N. Yu, X. Gong, K. Chen, and Y. Pan, "Three-Stream Convolutional Networks for Video-based Person Re-Identification," *arXiv preprint arXiv:1712.01652,* 2017.

[6] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing,* vol. 32, pp. 270-286, 2014.

[7] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 1622-1634, 2013.

[8] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *European Conference on Computer Vision*, 2012, pp. 413-422.

[9] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?," in *European Conference on Computer Vision*, 2012, pp. 391-401.

[10] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144-151.

[11] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3685-3693.

[12] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE transactions on pattern analysis and machine intelligence,* vol. 35, pp. 653-668, 2013.

[13] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European conference on computer vision*, 2014, pp. 1-16.

[14] Z. Zhang, Y. Chen, and V. Saligrama, "Group membership prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3916-3924.

[15] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 2528-2535.

[16] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846-1855.

[17] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly Attentive Spatial-Temporal Pooling Networks for Video-based Person Re-Identification," *arXiv preprint arXiv:1708.02286,* 2017.

[18] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 2014, pp. 34-39.

[19] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *European Conference on Computer Vision*, 2016, pp. 701-716.

[20] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Advances in Neural Information Processing Systems*, 2016, pp. 2667-2675.

[21] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1345-1353.

[22] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984,* 2016.

[23] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang*, et al.*, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*, 2016, pp. 868-884.

[24] R.Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083,* 2015.

[25] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly Attentive Spatial-Temporal Pooling Networks for Video-based Person Re-Identification," *arXiv preprint arXiv:1708.02286,* 2017.

[26] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology,* 2017.

[27] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," 2016.

[28] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568-576.

[29] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan*, et al.*, "Video-based person re-identification with accumulative motion context," *IEEE Transactions on Circuits and Systems for Video Technology,* 2017.

[30] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE transactions on pattern analysis and machine intelligence,* vol. 37, pp. 1629-1642, 2015.

[31] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European conference on computer vision*, 2014, pp. 536-551.

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

[33] N. Wojke and A. Bewley, "Deep Cosine Metric Learning for Person Re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 748-756.

[34] N. Martinel, G. L. Foresti, and C. Micheloni, "Distributed and Unsupervised Cost-Driven Person Re-identification," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, 2016, pp. 1225-1230.

[35] S. Liao, Z. Mo, J. Zhu, Y. Hu, and S. Z. Li, "Open-set person re-identification," *arXiv preprint arXiv:1408.0872,* 2014.

[36] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2016, pp. 769-773.

[37] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119,* 2015.

[38] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov*, et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048-2057.

[39] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *arXiv preprint arXiv:1512.05193,* 2015.

[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893.

[41] M. I. Khedher, M. A. El-Yacoubi, and B. Dorizzi, "Probabilistic matching pair selection for surf-based person re-identification," in *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*, 2012, pp. 1-6.

[42] I. PARTICIPANTS, "R4-A. 3: Human Detection and Re-Identification for Mass Transit Environments."

[43] A. Bedagkar-Gala and S. K. Shah, "Part-based spatio-temporal model for multi-person re-identification," *Pattern Recognition Letters,* vol. 33, pp. 1908-1915, 2012.

[44] E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using haar-based and dcd-based signature," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, 2010, pp. 1-8.

[45] E. Corvee, S. Bak, and F. Bremond, "People detection and re-identification for multi surveillance cameras," in *VISAPP-International Conference on Computer Vision Theory and Applications-2012*, 2012.

[46] J. Wu, C. Geyer, and J. M. Rehg, "Real-time human detection using contour cues," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 860-867.

[47] M. Eisenbach, A. Kolarow, K. Schenk, K. Debes, and H.-M. Gross, "View invariant appearance-based person reidentification using fast online feature selection and score level fusion," in *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, 2012, pp. 184-190.

[48] K.-E. Aziz, D. Merad, and B. Fertil, "People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, 2011, pp. 303-308.

[49] O. Javed, Z. Rasheed, O. Alatas, and M. Shah, "KNIGHT/spl trade: a real time surveillance system for multiple and non-overlapping cameras," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, 2003, pp. I-649.

[50] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys (CSUR),* vol. 46, p. 29, 2013.

[51] 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.

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

[53] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Transactions on Image Processing,* vol. 23, pp. 3656-3670, 2014.

[54] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in neural information processing systems*, 2003, pp. 521-528.

[55] L. Wu, C. Shen, and A. v. d. Hengel, "Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach," *arXiv preprint arXiv:1606.01609,* 2016.

[56] 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), 2010 IEEE Conference on*, 2010, pp. 2360-2367.

[57] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*, 2008, pp. 262-275.

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

[59] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2288-2295.

[60] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3610-3617.

[61] 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.

[62] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908-3916.

[63] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152-159.

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

[65] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017, pp. 6776-6785.

[66] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," *arXiv preprint arXiv:1511.06432,* 2015.

[67] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110-1118.

[68] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.

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

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[71] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *British Machive Vision Conference*, 2012, p. 11 pages.

[72] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*, 2011, pp. 91-102.

[73] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, 2008, pp. 1-6.

[74] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 1528-1535.

[75] D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi, "Video sequences association for people re-identification across multiple non-overlapping cameras," in *International Conference on Image Analysis and Processing*, 2009, pp. 179-189.

[76] S. Karaman and A. D. Bagdanov, "Identity inference: generalizing person re-identification scenarios," in *European Conference on Computer Vision*, 2012, pp. 443-452.

[77] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 33-40.

[78] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang*, et al.*, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*, 2016, pp. 868-884.

[79] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1798-1807.

[80] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*, 2010, pp. 143-156.

[81] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 3304-3311.

[82] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, pp. 773-787, 2017.

[83] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal Pyramid Pooling-Based Convolutional Neural Network for Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 27, pp. 2613-2622, 2017.

[84] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015, pp. 3810-3818.

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

[86] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 461-470.

[87] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-Shot Human Re-Identification Using Adaptive Fisher Discriminant Analysis," in *BMVC*, 2015, p. 2.

[88] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4516-4524.

[89] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[90] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research,* vol. 15, pp. 1929-1958, 2014.

[91] M. B. Christopher, *PATTERN RECOGNITION AND MACHINE LEARNING*: Springer-Verlag New York, 2016.

[92] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310-1318.
[93] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *International Conference on Machine Learning*, 2015, pp. 2342-2350.
[94] C. N. 57dos Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *CoRR, abs/1602.03609,* vol. 2, p. 4, 2016.
[95] 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), 2010 IEEE Conference on*, 2010, pp. 2360-2367.
[96] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014, pp. 818-833.
[97] F. Chollet. (2015). *Keras*. Available: https://github.com/keras-team/keras
[98] *TensorFlow:* . Available: https://www.tensorflow.org/
[99] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289,* 2015.
[100] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. (2011). *Person re-identification by descriptive and discriminative classification*. Available: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/PRID11/
[101] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam*, et al.* (2017). *Person re-identification by unsupervised video matching*. Available: http://www.eecs.qmul.ac.uk/~xiatian/downloads_qmul_iLIDS-VID_ReID_dataset.html
[102] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. (2017). *Person re-identification in the wild*. Available: http://www.liangzheng.com.cn/Project/project_mars.html
[103] X. Zhu, X.-Y. Jing, F. Ma, L. Cheng, and Y. Ren, "Simultaneous visual-appearance-level and spatial-temporal-level dictionary learning for video-based person re-identification," *Neural Computing and Applications,* pp. 1-13, 2018.
[104] N. McLaughlin, J. M. Del Rincon, and P. Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, 2015, pp. 1-6.
[105] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian, "Person re-identification meets image search," *arXiv preprint arXiv:1502.02171,* 2015.
[106] B. D. 58Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," 1981.
[107] W. Zhang, S. Hu, and K. Liu, "Learning Compact Appearance Representation for Video-based Person Re-Identification," *arXiv preprint arXiv:1702.06294,* 2017.