

**SPOTTING EVENTS IN FOOTBALL VIDEOS WITH A  
COMBINATION OF TWO-STREAM CONVOLUTIONAL  
NEURAL NETWORK AND DILATED RECURRENT  
NEURAL NETWORK**

**BEHZAD MAHASANI**

**FACULTY OF COMOPUTER SCIENCE AND  
INFORMATION TECHNOLOGY  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2021**

**SPOTTING EVENTS IN FOOTBALL VIDEOS WITH A  
COMBINATION OF TWO-STREAM CONVOLUTIONAL  
NEURAL NETWORK AND DILATED RECURRENT  
NEURAL NETWORK**

**BEHZAD MAHASANI**

**DISSERTATION SUBMITTED IN FULFILMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF MASTER  
OF COMPUTER SCIENCE**

**FACULTY OF COMPUTER SCIENCE AND  
INFORMATION TECHNOLOGY  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2021**

**UNIVERSITY OF MALAYA**  
**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: BEHZAD MAHASANI

Matric No: WMA 170001 (17036019/1)

Name of Degree: MASTER OF COMPUTER SCIENCE

Title of dissertation in this work:

SPOTTING EVENTS IN FOOTBALL VIDEOS WITH A COMBINATION OF  
TWO-STREAM CONVOLUTIONAL NEURAL NETWORK AND DILATED  
RECURRENT NEURAL NETWORK

Field of Study: Artificial intelligence

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This Work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge, nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this work, I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date: 26/04/2021

Subscribed and solemnly declared before,

Witness's Signature

Date: 30/4/2021

Name:

Designation:

# **SPOTTING EVENTS IN FOOTBALL VIDEOS WITH A COMBINATION OF TWO-STREAM CONVOLUTIONAL NEURAL NETWORK AND DILATED RECURRENT NEURAL NETWORK**

## **ABSTRACT**

In this research, we address the problem of event detection and localization in football (soccer) videos. While the problem of event detection in videos is itself a research problem, event detection in sports, especially in football, has an important commercial impact as well. Football is played by more than 250 million players in 200+ nations. In addition, it has the highest television audience in sport. This makes football the most popular sport in the world. Considering the advancement in streaming technologies on mobile platforms, it is important to develop efficient and fast processing algorithms for thousands of videos captured and stored in the cloud. Unlike images, videos provide additional temporal information. While this additional information is helpful, it also makes the reasoning more challenging. On one hand, from the local correlation between adjacent frames, it is possible to identify the short-range correlation between player movements. On the other hand, one can identify the mid-range and long-range correlation between events that are seconds away from each other. One important challenge in analyzing long videos is how to consider all range of correlations (short - long) between video frames. Localizing (temporal segmentation) events in a football video is a challenging problem. While the general problem of temporal segmentation in videos have been extensively addressed in the literature, to the best of our knowledge this work is the among the first to address the event localization problem in “long” football videos using end-to-end deep learning techniques. Football videos are long and the correlation between frames in the video ranges from short to long. To model various range of correlations in football videos, we propose to use a combination of two-stream CNNs and dilated RNNs with LSTM cells, to capture short-range and long-range correlations. Our experimental

result shows 5.4% - 11.4% accuracy improvement compared to the state of the art and the baselines for the problem of spotting in long videos presented in the largest football dataset available for research community (i.e., SoccerNet).

Keywords: Deep Learning, Recurrent Neural Networks, Two-stream CNN, Sport Video Analysis, Activity Detection and Spotting.

Universiti Malaya

**MENGESAN PERISTIWA DALAM VIDEO BOLA SEPAK DENGAN  
KOMBINASI RANGKAIAN NEURAL KONVOLUSI DUA-SALURAN DAN  
RANGKAIAN NEURAL BERULANG**

**ABSTRAK**

Dalam penyelidikan ini kami menangani masalah pengesanan peristiwa dan penempatan dalam video bola sepak. Pengesanan peristiwa video merupakan satu cabaran bagi penyelidik, manakala pengesanan peristiwa dalam bidang sukan terutamanya dalam acara bola sepak. Ini membawa impak komersil yang agak penting. Bola sepak dimainkan oleh lebih 250 juta pemain di 200+ negara. Di samping itu, acara ini mempunyai penonton televisyen yang tertinggi dalam bidang sukan. Ini menjadikan bola sepak sebagai acara sukan yang paling popular di dunia. Memandangkan bahawa kemajuan teknologi streaming pada platform mudah alih, ini adalah sangat penting untuk mencipta proses algoritma yang mampan dan laju untuk beribu-ribu video yang ditangkap dan disimpan dalam cloud. Video berbeza dengan gambar, di mana ia membekalkan informasi temporal. Walaupun informasi ini bermanfaat, namun informasi ini juga menjadikan penaakulan lebih mencabar. Biasanya dari korelasi tempatan antara bingkai yang bersebelahan, ini adalah berkemungkinan untuk mengenalpastikan korelasi jarak pendek antara pergerakan-pergerakan pemain. Manakala, seseorang dapat mengenalpastikan korelasi jarak sederhana dan jarak jauh antara peristiwa yang berlaku dalam detik dari satu sama lain. Salah satu cabaran penting dalam menganalisis video panjang adalah bagaimana untuk mempertimbangkan semua korelasi (pendek - panjang) antara bingkai-bingkai video. Penempatan (segmentasi temporal) peristiwa video bola sepak ialah(adalah) satu masalah yang mencabar. Walaupun masalah umum segmentasi temporal dalam video telah ditangani secara meluas dalam kerja penyalidikan yang telah diterbitkan, untuk pengetahuan terbaik kami, penyelidikan ini ialah antara penyelidik yang pertama dalam menangani masalah peristiwa penempatan video bola sepak yang

“panjang” dengan menggunakan teknik pembelajaran mesin mendalam. Video bola sepak adalah panjang dan korelasi antara bingkai dalam video berbagai dari pendek hingga panjang Untuk memodel pelbagai korelasi dalam video bola sepak, kami mencadangkan untuk menggunakan gabungan two-stream CNNs dan dilated RNNs dengan sel LSTM, untuk menghasilkan korelasi jarak pendek dan jarak jauh. Keputusan ujian kami menunjukkan pembaikan ketepatan yang ketara 5.4% - 11.4% (berbanding dengan keputusan terbaik yang ada dan keputusan asas) terhadap masalah spotting dalam video panjang didapati dari dataset bola sepak terbesar yang sedia ada untuk komuniti penyelidikan (iaitu SoccerNet).

Kata Kunci: Pembelajaran Mesin Mendalam, Rangkaian Neural Berulang (RNNs), CNN dua aliran (Two-stream CNN), Analisis Video Sukan, Pengesanan Aktiviti dan Spotting.

## **ACKNOWLEDGEMENTS**

First and foremost, I would like to thank my advisors, Dr. Erma Rahayu Binti Mohd Faizal Abdullah and Dr. Ram Gopal Raj, for their support during all stages of my research. Their encouragement and guidance have been invaluable. Without their continuous guidance and supervision, I would not have been able to complete my research and made it through my master's degree. I would also like to acknowledge the faculty of computer science and information technology at the University of Malaya for providing great research and academic environment.

Last but not least, I am thankful to my parents for their unconditional love and continuous support. I dedicate this thesis to my mom for her endless love and sacrifices, and to my dad for his never-ending encouragement. I also want to thank my brother, who has always been a good friend and a reliable advisor.



## TABLE OF CONTENTS

Abstract .....	iii
Abstrak .....	v
Acknowledgements .....	vii
Table of Contents .....	viii
List of Figures .....	xii
List of Tables.....	xv
List of Symbols and Abbreviations.....	xvii
List of Appendices .....	xx
<b>CHAPTER 1: INTRODUCTION.....</b>	<b>1</b>
1.1 Primary Background Studies .....	1
1.2 Motivation.....	2
1.2.1 Industry Impact.....	3
1.2.2 Academic Impact.....	4
1.3 Problem Statement.....	5
1.4 Aim and Objectives of the Research .....	7
1.5 Research Questions.....	8
1.6 Relationship Between Objectives and Questions .....	8
1.7 Research Methodology and Proposed Approach.....	9
1.8 Scope of the Research.....	10
1.9 Principal Contribution .....	11
1.10 Organization of the Thesis.....	13
1.11 Summary.....	14

<b>CHAPTER 2: LITERATURE REVIEW</b> .....	<b>15</b>
2.1 Introduction.....	15
2.2 Sport Analysis.....	16
2.2.1 Non-football Analysis .....	16
2.2.2 Football Analysis.....	17
2.3 Event Detection and Localization in Videos .....	21
2.3.1 Classical Approaches .....	23
2.3.2 Deep Learning Approaches .....	25
2.4 Review of Machine Learning .....	27
2.4.1 Classical Machine Learning with Handcrafted Features.....	27
2.4.2 Deep Convolutional Neural Networks .....	30
2.4.2.1 Convolutional Neural Networks Fundamental.....	31
2.4.2.2 VGG Network .....	32
2.4.2.3 ResNet Network .....	33
2.4.2.4 Two-Stream Convolutional Neural Network.....	35
2.4.3 Recurrent Neural Networks.....	39
2.4.3.1 Training Difficulties.....	42
2.4.3.2 Different Recurrent Units.....	44
2.4.3.3 Different Recurrent Neural Network Architectures .....	51
2.5 Conclusion.....	53
<b>CHAPTER 3: METHODOLOGY</b> .....	<b>55</b>
3.1 Introduction.....	55
3.2 Approaches of the Research .....	55
3.2.1 Review of Related Literature.....	56
3.3 Football Dataset (SoccerNet).....	58
3.3.1 Video Collection.....	58

3.3.2	Data Preprocessing .....	59
3.3.2.1	Game Synchronization with OCR .....	59
3.3.2.2	Collecting Event annotations .....	59
3.3.2.3	Splitting Dataset for Training, Testing and Validation .....	60
3.3.3	Dataset Comparison .....	61
3.4	Proposed Model .....	62
3.5	Evaluation of Proposed Model .....	66
3.6	Summary .....	70
<b>CHAPTER 4: EVENT SPOTTING AND CLASSIFICATION .....</b>		<b>71</b>
4.1	Introduction .....	71
4.2	Proposed Method .....	71
4.3	Feature Description .....	74
4.3.1	Short-range .....	74
4.3.1.1	Spatial Stream .....	75
4.3.1.2	Temporal Stream .....	75
4.3.2	Mid-range .....	76
4.3.3	Long-range .....	77
4.4	Classification .....	77
4.4.1	Two-stream CNN Training .....	78
4.4.2	Dilated RNN Training .....	81
4.5	Spotting .....	83
4.6	Implementation and Technical Details .....	87
4.6.1	Hardware Description .....	87
4.6.2	Software Description .....	88
4.6.3	Training, Testing and Validation Time .....	88
4.7	Summary .....	89

<b>CHAPTER 5: EXPERIMENTAL RESULTS AND DISCUSSION.....</b>	<b>91</b>
5.1 Introduction.....	91
5.2 Test and Evaluation of the Results for the Proposed Model .....	92
5.2.1 Event Classification.....	92
5.2.1.1 Summary and Discussion of Event Classification .....	94
5.2.2 Event Spotting .....	95
5.2.2.1 Baselines.....	97
5.2.2.2 Variations of Our Approach.....	98
5.2.2.3 Qualitative results.....	105
5.2.2.4 Summary and Discussion of Event Spotting.....	108
5.3 Summary.....	109
<b>CHAPTER 6: CONCLUSION AND FUTURE DIRECTION .....</b>	<b>111</b>
6.1 Achievement of Research Objectives.....	111
6.2 Main Contribution .....	113
6.3 Research findings and Outcome of the Research .....	115
6.4 Limitation and Future Discussion.....	115
References .....	117
List of Publications and Papers Presented .....	129
Appendix A.....	130
Appendix B .....	131
Appendix C .....	132

## LIST OF FIGURES

Figure 1.1: Research methodology framework followed in this thesis.....	10
Figure 2.1: Abstract view of the literature review structure in this research work.....	15
Figure 2.2 : Classical computer vision methods based on handcrafted features.....	25
Figure 2.3: Deep learning-based models outline. ....	27
Figure 2.4 : An example of 2D convolutional filter.....	32
Figure 2.5: The AlexNet network architecture proposed in (Krizhevsky et al., 2012)...	32
Figure 2.6: Architecture of AlexNet vs. VGG-16. Top: Architecture of AlexNet, Bottom: Architecture of VGG-16 (Yu et al., 2016). ....	33
Figure 2.7: Visualization of the skip-connection concept in neural networks.....	34
Figure 2.8: Original Two-Stream CNN proposed in (K. Simonyan & A. Zisserman, 2014). .....	37
Figure 2.9: Unrolled recurrent neural network structure. ....	40
Figure 2.10: Overall architecture of the LSTM unit. ....	44
Figure 2.11: Dilated version of single LSTM.....	45
Figure 2.12: Illustration of forget gate impact in LSTM. ....	46
Figure 2.13: Illustration of input gate impact in LSTM.....	46
Figure 2.14: Illustration of output gate impact in LSTM.....	47
Figure 2.15: Overall architecture of the GRU unit. ....	48
Figure 2.16: Illustration of a single GRU in an unfolded RNN.....	48
Figure 2.17: Illustration of update gate impact. ....	49
Figure 2.18: Illustration of the reset gate impact. ....	49
Figure 2.19: Illustration if the current memory content. ....	50
Figure 2.20: Illustration of the final memory at time “t”. ....	50
Figure 2.21 : Structure of the skipped RNN model (Chang et al., 2017).....	51

Figure 2.22: Overview of the Highway-LSTM-RNNs architecture (Y. Zhang et al., 2016). .....	52
Figure 2.23: Overall architecture of the 3 layer dilated RNN with dilations of one, two, and four (Chang et al., 2017) .....	53
Figure 3.1: Research methodology process followed in this thesis. ....	55
Figure 3.2: Three football events annotated in SoccerNet. ....	60
Figure 3.3 :Visualization of comparison between SoccerNet and available video datasets (Giancola et al., 2018).....	62
Figure 3.4: Overview: Our goal is to use low-level spatiotemporal features and a hierarchical recurrent model with skip connections (DilatedRNN with LSTM units) to improve event classification and event spotting in long football videos.....	63
Figure 3.5: Visual representation of the intersection over union (IoU) .....	67
Figure 3.6: Overall idea of spotting (Candidate X spot the event within a tolerance of $3\delta$ and $4\delta$ ) .....	67
Figure 3.7: High-level overview of various evaluations of event classification and spotting.....	68
Figure 3.8: Illustration of true positive, false positive, false negative and true negative instances. ....	68
Figure 4.1: Detailed illustration of the proposed model. Given the input video frames, we first compute a dense Opticalflow. Then the spatial stream network consumes the first frame, and the temporal stream network consumes the Opticalflow. The results from the two-stream networks are fused to form a single feature vector for future classification components (SoftMax layer or Dilated RNN). .....	72
Figure 4.2: Illustration of the Opticalflow computation. Opticalflow calculates the displacement of a location in two frames in the horizontal and vertical directions.....	76
Figure 4.3 : Various fine-tuning strategies in deep neural networks. ....	79
Figure 4.4: The fine-tuning approach used in this work. ....	79
Figure 4.5 : The training (fine-tuning) process of Two-stream CNN .....	81

Figure 4.6 : Process of generating one-minute training clips for RNN. Given the event annotation, we create a training set using the one-minute clips by subsampling five frames per second. ....	83
Figure 4.7: Different temporal definition of events in videos for a single event (left) and multiple events with overlap (right) (Mleya et al., 2019). ....	84
Figure 4.8: Definition of anchor point in event spotting and the corresponding tolerance thresholds(Giancola et al., 2018).....	85
Figure 4.9: Spotting algorithms used based on the watershed segmentation method.....	86
Figure 4.10: Spotting algorithms used based on the Non-Maximum Suppression (NMS) method. ....	87
Figure 5.1: Average lost per training epoch. (a) training loss, (b) validation loss.....	93
Figure 5.2: Illustration of an anchor point in event spotting. For a given anchor point, we consider multiple error tolerance thresholds. A candidate event spot is correct if it falls in the error tolerance window and it is incorrect otherwise.....	96
Figure 5.3: Spotting results for variants of our single-frame models (mAP vs error tolerance threshold). All models are trained on 60-second videos. ....	98
Figure 5.4: Spotting results for variants of our recurrent-based models (mAP vs error tolerance threshold). All models are trained on 60-second videos. ....	102
Figure 5.5: Spotting results for variants of our recurrent-based models (mAP vs error tolerance threshold). All models are trained on 20-second videos. ....	103
Figure 5.6: Spotting results for variants of our recurrent-based models (mAP vs error tolerance threshold). All models are trained on 5-second videos. ....	103
Figure 5.7: Successful goal spotting example for 5s tolerance (Italy Série A, 2015-2016/2015-09-22, 21:45, Udinese 2 - 3 AC Milan. ....	106
Figure 5.8: Failed substitution spotting example for 20s tolerance (Germany Bundesliga, 2015-2016/2015-10-04, 18:30, Bayern Munich 5 - 1 Dortmund). ....	106
Figure 5.9 : Successful examples of goal, substitution, and card event spotting ((a) presents goals sample, (b) presents substitution sample and (c) presents cards sample). ....	107

## LIST OF TABLES

Table 1.1: Mapping between objectives and questions.....	9
Table 2.1: The table consist of two groups of paper: papers from 2002 to 2014 which use classical computer vision feature and shallow machine learning models, papers from 2014 to 2020 which use deep learning methods. ....	20
Table 2.2: Summary of classical ML methods and hand-crafted features. ....	28
Table 2.3: Comparison between different CNN architectures. ....	38
Table 2.4: Comparison between different feature extraction techniques.....	39
Table 2.5: Example applications of RNN in various domains.....	41
Table 3.1: Details of the collected games in SoccerNet.....	58
Table 3.2: Details of the dataset splits: training, testing and validation splits. ....	61
Table 3.3: Comparison of SoccerNet dataset with available video datasets.....	61
Table 3.4: Definition of the confusion matrix.....	68
Table 4.1 : Architecture of ResNet-50 used in this study (unit are by pixel) .....	72
Table 4.2 : Training and Testing (processing) time .....	88
Table 5.1 : Overview of test and evaluation of the proposed model for event spotting in Football videos .....	91
Table 5.2: Accuracy result of the proposed models for event classification presented as mean average precision (mAP). ....	94
Table 5.3 :Definition of different event spotting baselines.....	97
Table 5.4 : Definition of different variation of our approaches. ....	99
Table 5.5: mAP result for spotting football events trained on 5, 20, and 60 seconds, averaged over three different error thresholds of 5 seconds, 20 seconds, and 60 seconds. ....	101
Table 5.6: Comparison of our proposed event spotting approach compared to state of the. The models are trained on 5, 20, and 60 second videos. The results show the area under curve (AUC) of the mAP plots shown in Figure 5.4. ....	104



Table 5.7: Comparison of our proposed event spotting approach compared to state of the art. The reported mAP results from the state of the art are from published results. Note that the best result reported in (Giancola et al., 2018) is trained on 20-second videos. For the results in (Cioppa et al., 2020) and (Vats et al., 2020) it is not clear what video length the models are trained on. Our best result is obtained from models trained on 60-second videos. This shows that explicit modeling of long-range dependencies improves the accuracy for longer videos. .... 105

Table 5.8 :A summary of improvements based on the results presented in Table 5.5 compare to different baselines and state of the art ..... 109

Universiti Malaya

## LIST OF SYMBOLS AND ABBREVIATIONS

VAR	:	Video Assistant Referee
HUDL	:	HUDL is a product and service of Agile Sports Technologies
FIFA	:	Fédération Internationale de Football Association
CVsports	:	Computer Vision in Sports
AI	:	Artificial Intelligence
AWS	:	Amazon Web Service
CNNs	:	Convolutional Neural Networks
RNNs	:	Recurrent Neural Networks
AP	:	Average Precision
mAP	:	Mean Average Precision
LSTMs	:	Long-Short Term Memories
GRUs	:	Gated Recurrent Units
BNs	:	Bayesian networks
DBNs	:	Dynamic Bayesian networks
HMMs	:	Hidden Markov Models
IoU	:	Intersection over Union
HOG	:	Histogram of Oriented Gradient
HOF	:	Histogram of Optical Flow
SIFT	:	Scale-Invariant Feature Transform
SVMs	:	Supported Vector Machines
KNNs	:	K-Nearest Neighbors
PCA	:	Principle Component Analysis
BoW	:	Bag-of-word
iDT	:	Improved Dense Trajectory

MBH	:	Motion Boundary Histogram
HSV	:	Hue saturation value
FV	:	Feature vectors
DeCaf	:	Deep Convolutional Activation Feature
STIPs	:	Spatiotemporal Interest Points
DTFs	:	Dense Trajectory Features
SSVMs	:	Structured Supported Vector Machines
NBNN	:	Naive Bayes Nearest Neighbor
CRFs	:	Conditional Random Fields
HCRFs	:	Hidden Conditional Random Fields
DCNNs	:	Deep Convolutional Neural Networks
BPTT	:	Backpropagation Through Time
CVPR	:	Computer Vision and Pattern Recognition
ICCV	:	International Conference on Computer Vision
FPS	:	Frame Per Second
SD	:	Standard Definition
HD	:	High Definition
EN	:	England
ES	:	Spain
DE	:	Germany
ET	:	Italia
EU	:	Europe
OCR	:	Optical Character Recognition
TP	:	True Positive
FP	:	False Positive
TN	:	True Negative

FN : False Negative  
AUC : Area Under Curve  
XAI : Explainable AI

Universiti Malaya

## LIST OF APPENDICES

APPENDIX: A	130
APPENDIX: B	131
APPENDIX: C	132

Universiti Malaya

## CHAPTER 1: INTRODUCTION

### 1.1 Primary Background Studies

Sports video analysis has been an active research area in the last few years (Assfalg, Bertini, Del Bimbo, Nunziati, & Pala, 2002; Brendel, Fern, & Todorovic, 2011; S. Chen, Fern, Mahasseni, & Todorovic, 2013; S. Chen, Fern, & Todorovic, 2014; Cioppa et al., 2020; D’Orazio & Leo, 2010; Ekin, Tekalp, & Mehrotra, 2003; C.-L. Huang, Shih, & Chao, 2006; Ibrahim, Muralidharan, Deng, Vahdat, & Mori, 2016; Jiang, Lu, & Xue, 2016; Kautz et al., 2017; Lan, Sigal, & Mori, 2012b; Ramanathan et al., 2016; Tavassolipour, Karimian, & Kasaei, 2014; Todorovic & Mahasseni, 2016; Tovinkere & Qian, 2001; Vats, Fani, Walters, Clausi, & Zelek, 2020; Z. Wang, Yu, & He, 2016; P. Xu et al., 2001; Y. Yang, Lin, Zhang, & Tang, 2007). Thanks to high-speed internet on portable platforms, streaming technology has advanced rapidly in the past decade. As a result, there has been a great demand for the use of video streaming and sharing services, and annotation platforms of sports videos. Unfortunately, majority of the cloud-based services provide a limited set of capabilities to their customers for rapid access to particular video highlights or functionalities to reduce the manual effort for a content search in videos.

Two important challenges in video analysis are the localization of the key moments in a video and the classification of the localized key moments into certain event categories. While the former considers the temporal segmentation of a given video, the latter focuses on classifying the content of a short segment of the video. Both of these problems are even more complicated in scenarios with very high dynamics such as sports videos.

Among all sports, football (soccer) is unquestionably one of the most, if not the most popular sport. Just in Europe, revenue from football is more than 25 billion dollars. In a revolutionary decision, FIFA decided to use the Video Assistant Referee system, known

as VAR, in world cup 2018 games. This was football's first use of computer vision technology. This shows the importance of intelligent football video analysis systems, especially for real-time events. Note that in this manuscript, “soccer” and “football” are used interchangeably if not stated otherwise.

While multiple solutions have been proposed to automatically analyze other sports telecasts such as hockey and basketball, understanding football videos is much more complicated. This is mainly because of the event sparsity in long football videos. Unlike similar outdoor sports such as “American Football” and “Baseball” which are episodic, football is a non-episodic game. A football game is at least a continuous 45-minutes long sport (each half of the game). As a result, manual search for contents and highlights in a large number of long football videos is not plausible.

In this work, we propose a new architecture for temporal segmentation and classification of important football events. The proposed architecture considers the mid-range and long-range correlation between frames in addition to the local spatiotemporal information to efficiently segment major events in a video. More specifically, we benefit from a combination of models at different granularity levels to consider local spatiotemporal clues, short-range temporal information, and mid to long-range temporal correlation.

## **1.2 Motivation**

This research focuses on identifying a solution for football event classification and spotting. We argue that this is a very important research problem with a huge impact on academic and industry. In the next two subsections, we briefly provide more information to highlight the industry and academic impact of this work.

### 1.2.1 Industry Impact

As stated in (Giancola, Amine, Dghaily, & Ghanem, 2018), "the global sports market is estimated to generate an annual revenue of \$91 billion, whereby the European football market contributes about \$28.7 billion (i.e., more than 30%). After merchandising, Television broadcast rights are the second major revenue stream for a football club". While entertainment serves as the main purpose of the football broadcast, recently main broadcasting platforms (e.g., Wyscout<sup>1</sup>, Reely<sup>2</sup>, and Stats SPORTVU<sup>3</sup>) offer sports analytics as part of their core marketing to help professionals to produce statistics, analyze plans, and scout new players.

Also, football is the most popular sport in the world with more than 250 million players in more than 200 countries. With the rapid development in video streaming technology, users watch, share and annotate more football videos than before. Most of the available web services, however, do not provide enhanced functionalities to their users that would enable faster access to certain video moments, or would reduce manual labor in video annotation. This means having more intelligent highlights and event detection algorithms is helpful for users to save their time and money. In addition to online viewers and fans, professional athletes and coaches need to access important events to review their games or to analyze the opponent's. It is ideal to achieve this without a need to browse the entire video.

---

<sup>1</sup> Wyscout is an Italian company that supports football scouting, match analysis and transfer dynamic

<sup>2</sup> Reely is a computer vision, AI, deep learning platform specializing in transforming sports video content into actionable data.

<sup>3</sup> SportVU is a camera system hung from the rafters that collects data 25 times per second and follows the ball and every player on the court.



Given the growing sport analysis industry (e.g., SportsLogiq<sup>4</sup>, HUDL<sup>5</sup>, and Reely), results from this study are valuable to the industry practitioners as well as related software providers in developing better practices and tools for football video analysis. We believe since this study is evaluated on a large soccer dataset (i.e., SoccerNet), the results will generalize well in real setting and the aforementioned industry solutions will benefit from this research work.

To summarize, more accurate event localization helps various potential stakeholders by enabling quicker and more effortless access to important events in videos.

### **1.2.2 Academic Impact**

Sport Analysis has attracted a large number of researchers in academia in the past two decades. We believe this is mainly due to certain challenges in sport videos such as high-speed events and complex correlation between events. Football in particular has certain unique characteristics. It is an outdoor game with varying lighting conditions which makes the processing of raw images challenging. In addition, football field is exceptionally large compared to other team sports such as volleyball or basketball with more players spread out in the field. More importantly, football is a highly dynamic game with “sparse interesting events in long videos”. In other words, there are few interesting events in a long football video, which is very different from volleyball or basketball with high rate of interesting events (e.g., goal). These challenges make sports video analysis and in particular analysis of football videos an interesting problem domain for academia. This has resulted in multiple workshops and challenges in top academic conferences such as International Workshop on Computer Vision in Sports (CVSports), ActivityNet

---

<sup>4</sup> SportLogiq is an AI powered sports analytics company.

<sup>5</sup> Hudl is a product and service of Agile Sports Technologies, Inc.

challenge<sup>6</sup>. We believe that this research provides a new baseline for other researchers in this field.

### 1.3 Problem Statement

The main goal of Artificial Intelligence (AI) is to train intelligent systems which are able to behave in a similar way as humans. While humans, perform a large class of actions without carefully analyzing the action itself, training an intelligent system with similar behavior is a challenging task. One of the main abilities of humans is the ability to recognize (to differentiate) different events upon observation. For example, given a clip of a football video, humans can easily identify if the clip contains a penalty shot or not. Thanks to high-speed internet on portable devices, streaming technology has advanced rapidly in the past decade. This has resulted in an exponential demand increase of cloud-based storage, sharing, and annotation platforms as well as the streaming services for sports videos. Unfortunately, most of these cloud-based services do not provide enhanced functionalities to their users that would enable faster access to certain video highlights or reduce manual effort in video annotation. Sports videos analysis became a very important research topic in the last few years (D’Orazio & Leo, 2010; Jiang et al., 2016; Tovinkere & Qian, 2001; Z. Wang et al., 2016; P. Xu et al., 2001) and several high-rank conferences have assigned certain workshops (e.g., CVSports) for this area of research. One of the key challenges in any video analysis is to localize the key moments in the video and to classify these key moments into certain categories. This is even harder in sports videos which usually cover high dynamic content. Undoubtedly, football is one of the most (if not the most) attracting sports in the world. Just in Europe, its revenue is more than 25 billion dollars. Given the long duration of a football game, intelligent methods for understanding football videos help the viewers with the localization of the salient

---

<sup>6</sup> <http://activity-net.org/>

moments of a game. While several companies such as HUDL, Reely, SportLogiq has built semi-automatic approaches to analyze sports telecasts such as “hockey” videos, identifying football events is still a challenge and most of the current commercial software still rely heavily on human annotations.

In a revolutionary decision, FIFA decided to use the Video Assistant Referee system, known as VAR, in world cup 2018 games. While using computer vision is still at its infancy in football, this shows the importance of intelligent football video analysis systems with high accuracy and real-time capability.

While researchers have applied preliminary recurrent based architectures to classify videos captured in the wild (Donahue et al., 2015; Todorovic & Mahasseni, 2016), these models have not yet been fully applied to the problem of event detection in football videos.

As mentioned in the introduction section, our goal is to localize the events in football videos. While similar research work has addressed event localization on other sports and even football videos, we identified the following challenges:

- **One key observation** is that unlike most of the events in generic videos, events in sport videos occur in a glimpse. This means that the interval that the event happens is very short. Also, since the sport is very fast pace the boundaries of the event are very subjective.
- **Another important observation** is that there are *short-range, mid-range, and long-range* dependencies between frames in football videos. To the best of our knowledge, the prior work has only addressed one of the above correlations at a time. We believe, in order to effectively localize events in football videos, a model should consider long-range, mid-range, and short-range correlation.

- **Last but not least**, the majority of the prior work has only considered short football videos for evaluation. To show the strength of an algorithm and performance on actual football videos it is important to evaluate the models on a large-scale video dataset which is representative of the diversity in long football videos.

To summarize, our goal is to recognize the significant moments in a football video. Certain events such as "goal event" or "card event" are of great importance and localizing them in a video helps with faster retrieval of key highlights. Localization of the significant moments in a video and classification of the localized moments into pre-defined event categories are the two important challenges in video analysis.

The problem of temporal segmentation of a given video is one of the challenging problems in computer vision. This problem is even more complicated in scenarios where 1) there are highly dynamic scenes such as sports videos, 2) events happen in a glimpse, 3) there are dependencies between the distant video frames.

In our opinion, the main reason for less accurate results in football is the fact that football is a highly dynamic game. Different events in a football video are highly correlated, and the correlation in time varies from short-range to mid-range and long-range. Football is an outdoor game, which means that scene appearance and lighting conditions highly vary between games. Finally, while 22 players spread over a large area on the football field, the game focus can instantaneously change in a matter of seconds.

#### **1.4 Aim and Objectives of the Research**

The main goal of this research is to provide an approach to analyze football videos. In this work, we define the analysis of football videos as the process of localizing and classifying three key events in a football game. We argue that these three events are

amongst the most important events in any football game, and accurate estimation would improve the quality of the automatic video analysis in sports videos. We identify the following research objectives:

1. To investigate the state-of-the-art feature extraction models in videos in order to improve the event classification and spotting in football videos.
2. To design and implement a neural network model to improve event localization in long football videos.
3. To evaluate the accuracy of the proposed algorithm for event localization and classification in long football videos.

## **1.5 Research Questions**

The following research questions are sought while conducting various significance of this research:

1. What type of feature descriptors should be considered to improve the event localization accuracy in football videos?
2. How to combine modern neural networks architecture to address the limitation of previous event classification and detection approaches in football videos?
3. What is the event classification and spotting performance of our method compared to the baselines and existing approaches?

## **1.6 Relationship Between Objectives and Questions**

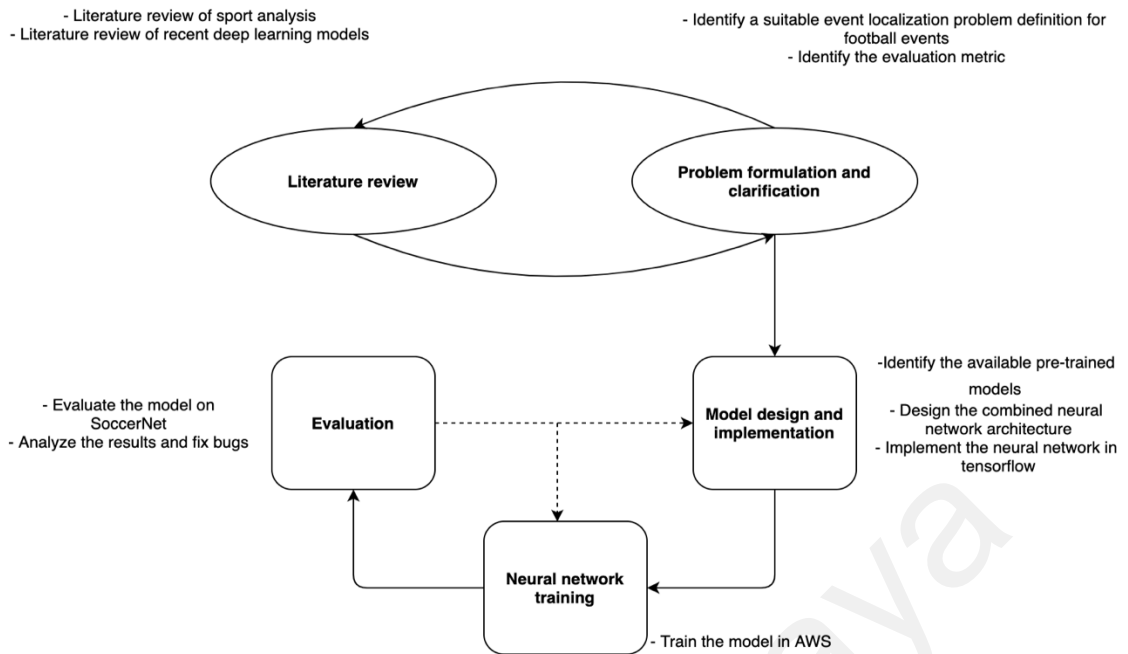
To demonstrate how research questions and objectives are connected, Table 1.1 provides the mapping between the research questions and the objectives of the research.

**Table 1.1: Mapping between objectives and questions**

<b>Research question</b>	<b>Research objective</b>
What type of feature descriptors should be considered to improve the event localization accuracy in football videos?	To investigate the state-of-the-art feature extraction models in videos to improve the event classification and spotting in football videos.
How to combine modern neural networks architecture to address the limitation of previous event classification and detection approaches in football videos?	To design and implement a neural network model to classify and localize three events in long football videos.
What is the event classification and spotting performance of our method compared to the baselines and existing approaches?	To evaluate the accuracy of the proposed algorithm for event localization and classification in long football videos.

## **1.7 Research Methodology and Proposed Approach**

To clarify the research problem, answer the research questions and achieve the research objectives, we followed the following quantitative research method presented in Figure 1.1. First, in order to explain the problem definition and properly formulate the problem statement we did an extensive research literature review. Second, based on the literature review, we refined the problem formulation and identified the current issues and challenges. Note that this itself resulted in more literature review. After formalizing the problem, we designed our event classification and spotting model and implemented its combined neural network components, in the TensorFlow framework. We then trained the network on Amazon Web Service (AWS) machines and evaluated the results on the SoccerNet dataset.



**Figure 1.1: Research methodology framework followed in this thesis.**

## 1.8 Scope of the Research

The scope of this study focuses on proposing a unified neural network model that combines the most recent successful neural network architectures to model short-range, mid-range, and long-range dependencies for event spotting in football videos. For evaluation, the scope is limited to the largest publicly available football dataset (SoccerNet). The proposed architecture considers the long-range correlation between frames in addition to the local spatiotemporal information to efficiently segment major events in a video. More specifically, we benefit from a combination of models at different granularity levels to consider local spatiotemporal clues (i.e., short-range temporal information) as well as mid-range and long-range temporal correlation.

In addition, two neural network architecture are combined for event spotting task in football videos. These are 1) Two-stream Convolutional Neural Networks (Two-stream CNN), and 2) Dilated Recurrent Neural Network (Dilated-RNN). The evaluation results are limited to three neural network architectures which include RNN and Two-stream CNN and ResNet-CNN using the SoccerNet dataset.

As part of the evaluation of our methods, presented in Chapter 5 we perform ablation studies to show the effectiveness of each component of our proposed neural network model. Performance evaluation is limited to the mean average precision (mAP) metric which is compared with the baseline.

## 1.9 Principal Contribution

A more accurate event detection model in football will help online users to access the key moments of a football matches. On the other hand, it would be a great advantage for the professional athletes and coaches to access the important events in a football video without a need to browse the entire video. Also, as a result of successful applications of the state-of-the-art methods in video analysis, there are multiple activity detection competitions in top tier conferences such as “International Conference on Computer Vision (ICCV)” and “Computer Vision and Pattern Recognition (CVPR)”. Activity detection in sports is even a more challenging problem because of the **high complexity** and **complex dynamics**. Recently, most of these top tier conferences, have a separate “sports analysis workshop” (e.g., CVSports).

Every sport has its unique characteristics which will lead to certain assumptions. These assumptions result in different video analysis models for different sports. Unlike some sports like basketball and volleyball, football is outdoor. As a result, scene appearance and lighting conditions highly vary between games. Also, unlike basketball and volleyball, football is a non-episodic game with high dynamics (e.g., both basketball and volleyball all episodic games). As a result, we argue that events in a football video are highly correlated and the correlation in time could be short-range, mid-range, or long-range. Considering the aforementioned impact of this research in academia and industry in the first chapter, and the fact that previous work on football video analysis has not considered the short-range, mid-range, and long-range correlation between frames, has



motivated me to undertake this research. My goal is to achieve the research objectives of this work and to propose a more sophisticated method for event localization.

To the best of our knowledge, none of the previous works in the area of football video analysis has modeled the complex correlation among frames together in one single model. We believe there are two reasons for this:

1. Lack of large-scale football datasets which makes it hard to train deep learning models (Giancola et al., 2018).
2. It is harder to model complex frame relations in long football videos using standard RNN and CNN models (Jiang et al., 2016).

As far as we are aware, our work is the first to consider various ranges of dependencies among video frames for event localization in sports videos. Our main contribution is the development of a new unified model for video event localization in football videos. The following summarizes our contributions:

- We used the two-stream CNN network for extracting local spatiotemporal features in long football videos.
- We explicitly model mid-term correlation between frames using LSTM network.
- We used dilated RNNs to capture the long-range dependencies between video frames.
- We evaluate our approach on the **largest publicly available** football video dataset which has shown up to **(5.4% - 6.9%)** accuracy improvement in event spotting compared to the state of the art and up to **(7.2% - 11.4%)** accuracy improvement compared to our simple baseline.
- We ran an extensive ablation study to analyze the contribution of each component in our proposed unified neural network model.

## **1.10 Organization of the Thesis**

This thesis is organized in seven chapters, including the introduction chapter. The rest of the thesis is structured as follows:

CHAPTER 2, “Literature review”, presents a comprehensive review of the most recent relevant published literatures related to sport video analysis. This chapter also provides an overview of the state-of-the-art approaches and techniques used in video understanding. In addition, it covers the successful techniques and approaches for classical feature extraction and machine learning methods as well as modern deep learning-based techniques in computer vision.

CHAPTER 3, “Methodology”, presents the research methodology used for this research work. This chapter describes the data collection sources and presents the proposed model for event classification and localization. Finally, the chapter describes the evaluation approach and metric used in this study to analyze and test the proposed model.

CHAPTER 4, “Football Event Spotting and Classification”, provides a detailed description of the problem statement and proposed model in our research as well as the methods and concepts used for feature extraction and classification. Finally, this chapter provides a complete overview of the implementation details and, training/testing pipelines.

CHAPTER 5, “Experimental Results and Discussion”, presents the evaluation results and accuracy of the proposed models for event classification and localization. It also provides explanations and analysis of the provided results.

CHAPTER 6, “Conclusion and Future Direction”, discusses the research findings and summarizes the significance of the study and the research contributions made in this work. It also explains a set of limitations and proposes potential future works.

### **1.11 Summary**

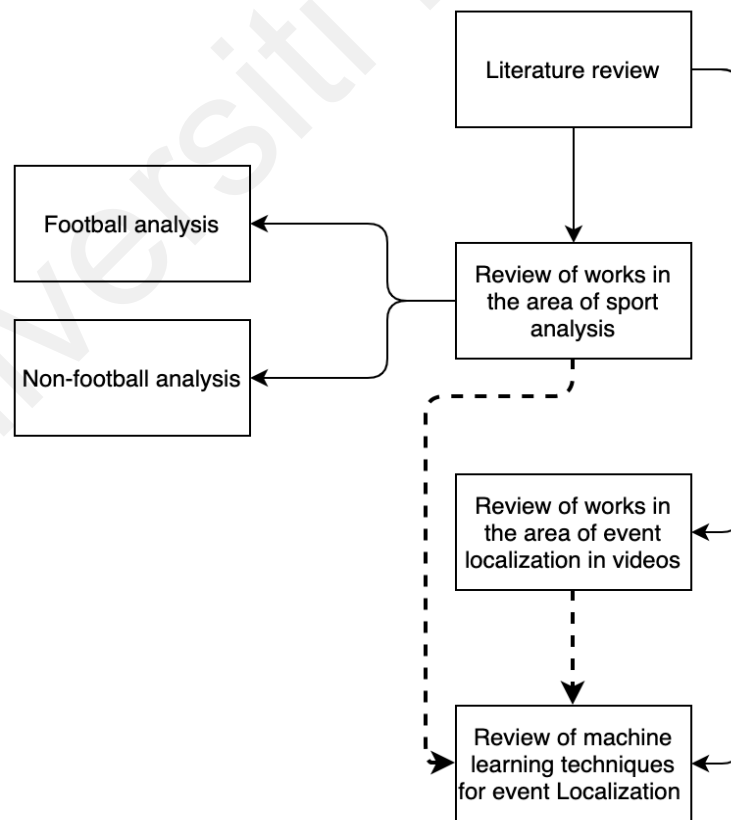
In summary, this chapter presented an introduction to this research and presented an overall view of the event localization and spotting in football videos. The motivation which inspired this research work, and the research objectives and questions were also presented in this chapter. The mapping between the objectives and research questions and how we planned to achieve those is also provided.

Universiti Malaysia

## CHAPTER 2: LITERATURE REVIEW

### 2.1 Introduction

The following two sections provide a review of the prior work on sports video analysis (football and non-football) and a review of the state-of-the-art machine learning models applied on image and video analysis. We first provide a general overview of prior work on sport analysis in computer vision. We then provide a detailed overview of prior work on analyzing football videos, including different problems addressed in each work. Second, we provide a brief overview of relevant recent approaches for event detection and localization in videos (not necessarily sport-related). Finally, we provide a comprehensive overview of the building blocks of the most recent models we plan to use in our work. Figure 2.1 illustrates the semantic view of the literature review structure. In addition, APPENDIX A presents the knowledge map of our literature review.



**Figure 2.1: Abstract view of the literature review structure in this research work.**

## **2.2 Sport Analysis**

Sports analysis gained recognition in the research community in the past decade due to rapid growth in video streaming on mobile platforms and the high demand for broadcasting sports videos. Multiple research and industry projects focused on sport analysis applications provided promising results. While it is not possible to review all prior work in this paper, we provide a high-level categorization of the related work by focusing on different sports each work has addressed.

### **2.2.1 Non-football Analysis**

Basketball has been the focus of multiple prior works (Brendel et al., 2011; L. Chen, Zhai, & Mori, 2017; S. Chen, A. Fern, et al., 2014; Ramanathan et al., 2016). While authors in (S. Chen, A. Fern, et al., 2014) use classical machine learning algorithms grounded on hand-crafted features to track the basketball players and to classify the videos, more recent approaches exploit deep learning-based techniques which learn the features and model parameters in an end-to-end manner (Ramanathan et al., 2016). Volleyball is another highly dynamic sport addressed in (Donahue et al., 2015; Ibrahim et al., 2016; Kautz et al., 2017). Authors in (Kautz et al., 2017) propose a method based on convolutional neural networks to track the players and classify the activities in beach volleyball. In (Ibrahim et al., 2016), deep recurrent models (multi-layer LSTMs) are used to identify the volleyball players in the field and to perform high-level structured activity detection. In a completely different line of research, hierarchical graphical models have been applied to perform group activity recognition in hockey videos (Lan et al., 2012b).

While the prior work has shown promising results in analyzing videos from basketball, volleyball, and hockey games, similar approaches are either not applicable to football videos or do not provide accurate results.

We believe the main reason for these less accurate results in football is the fact that football is a non-episodic and highly dynamic game. As a result, different events in a football video are highly correlated and the correlation in time could be short-range, mid-range, or long-range. Also, unlike most of these sports, football is an outdoor game, which means that scene appearance and lighting conditions highly vary between games. Another major difference is the higher number of players in a football game. In football players are scattered in the football field and the arrangement of the players is extended over a large area.

‘American football’ is probably the most comparable sport to football (soccer) in terms of the field appearance and number of players. Similar to football, it is also played outdoors. Several prior works (Atmosukarto, Ghanem, Saadalla, & Ahuja, 2014; S. Chen, Z. Feng, et al., 2014; S. Chen et al., 2013) have addressed ‘American football’ video analysis. Most of these approaches use classical machine learning methods to model the interaction between players and the motion in the scene. The main difference between football (soccer) and the ‘American football’ is the fact that unlike ‘American football’ where the game is episodic (and is divided to multiple plays) and well-structured (S. Chen et al., 2013), football (soccer) is a non-episodic game and the game spreads over almost the entire field in a very short (sometimes even seconds) period of time.

### **2.2.2 Football Analysis**

Analyzing football videos have attracted researchers in the computer vision field for more than a decade now. Detecting certain events, highlighting an event, tracking football players, and providing game statistics are among the few examples of applications addressed in the literature.

Analyzing football videos has been addressed in (Assfalg, Bertini, Colombo, Del Bimbo, & Nunziati, 2003; Assfalg et al., 2002; Ekin et al., 2003; Yu Huang, Llach, &

Bhagavathy, 2007; Kolekar & Sengupta, 2015; Pallavi, Mukherjee, Majumdar, & Sural, 2008a, 2008b; X Qian, Hou, Tang, Wang, & Li, 2012; Xueming Qian, Liu, Wang, Li, & Wang, 2010; Tavassolipour et al., 2014; Z. Wang et al., 2016; Wickramaratna, Chen, Chen, & Shyu, 2005). Authors in (Z. Wang et al., 2016) propose a video annotation platform based on semantic matching using coarse time constraints. More specifically, video events and external text information (match reports) are synchronized using their semantic correspondence in the temporal sequence. Authors in (Ekin et al., 2003) propose to use cinematic and object-based features to summarize and analyze football videos. Their method includes novel low-level football video processing algorithms, such as dominant color region detection, robust shot boundary detection, and shot classification, as well as some higher-level algorithms for goal detection, referee detection, and penalty-box detection. A recent work (Jiang et al., 2016) focuses on combining CNN (LeCun, Bottou, Bengio, & Haffner, 1998) and RNNs (Hochreiter & Schmidhuber, 1997). CNN features are fed to RNN layers to solve the task of football event detection. This is perhaps the closest work to ours. The main difference is that we use dilated RNNs with LSTM cells grounded on top of two-stream neural networks which enables us to perform well on long videos.

Video summarization and highlight detection is another important application area. Authors in (Tavassolipour et al., 2014) use Bayesian approaches to summarize videos and to detect events using semantic analysis through Bayesian inference. Similarly, Bayesian networks (BN) and Dynamic Bayesian networks (DBNs) grounded on low-level image features are used alongside algorithms which use a high-level knowledge encoded in abstract non-geometric representations to perform video summarization. Authors in (Assfalg et al., 2002) propose Hidden Markov Models (HMMs) to classify and recognize highlight football clips. Finally, authors in (Y. Yang et al., 2007) proposed a more generic approach for high-light extraction in football videos based on the goal-mouth detection.

More recently, authors in (Giancola et al., 2018) proposed to use the anchors of soccer events in football videos for the task of events spotting. They also published the SoccerNet dataset for event classification and spotting in football (soccer) videos. Follow-up research presented in (Cioppa et al., 2020) proposed a new loss function specifically designed for event spotting in football videos. Authors in (Vats et al., 2020) proposed to use a CNN architecture with multi-layer 1D convolutions for event spotting, which improved the accuracy result on SoccerNet (Giancola et al., 2018).

Our approach is different from all prior work in the following:

- We propose to use a hierarchical RNN architecture with LSTM units to find the dependency between sub-events. This enables us to identify long-range correlation between different events in a long video.
- We use two-stream networks for feature detection in the CNN layers which have resulted in better accuracy in other action recognition tasks. This allows us to describe short-range correlation between frames in a more expressive way.
- We evaluate our approach and the baselines on the recently released video dataset which is currently the largest football video dataset in the research community- SoccerNet (Giancola et al., 2018). Our proposed model improved the accuracy of the state-of-the-art method on SoccerNet (Giancola et al., 2018) by introducing a combination of two neural networks and improved the action spotting baseline.

To summarize, our main contribution is the new approach for modeling long-range correlations between frames. Unlike these prior works, we explicitly represent the short-range dependencies using local spatiotemporal features and long-range dependencies using a hierarchical recurrent neural network with skip connections.



Table 2.1 summarizes the relevant work in soccer video analysis.

**Table 2.1: The table consist of two groups of paper: papers from 2002 to 2014 which use classical computer vision feature and shallow machine learning models, papers from 2014 to 2020 which use deep learning methods.**

Author	Method	Problem
(Assfalg et al., 2002)	HMM	Highlight detection
(Ekin et al., 2003)	Bayesian network	Soccer video analysis and summarization,
(Assfalg et al., 2003)	Finite state machines	Highlight detection
(Wickramaratna et al., 2005)	Feed-forward neural network	Goal event detection
(C.-L. Huang et al., 2006)	Bayesian network	Semantic analysis
(Y. Yang et al., 2007)	Top-Hat Transform	Highlight extraction
(Yu Huang et al., 2007)	Color segmentation	Player and ball detection
(Pallavi et al., 2008b)	Graph-based	Player tracking
)Pallavi et al., 2008a(	Hough Transform and trajectory estimation	Ball detection
Baccouche, Mamalet, ( Wolf, Garcia, & Baskurt, )2010	KNN, SVM, LSTM-RNN	Action Classification
(Xueming Qian et al., 2010)	Hidden conditional random field	Highlight events detection
(Zawbaa, El-Bendary, Hassanien, & Abraham, 2011)	SVM	Summarization
(X Qian et al., 2012)	Hidden conditional random field	Events detection,
(Tavassolipour et al., 2014)	Bayesian network and copula	Event detection and summarization
(Kolekar & Sengupta, 2015)	Bayesian network-based	Highlight generation
(Z. Wang et al., 2016)	Bayesian network-based	Event annotation
(Jiang et al., 2016)	CNN + RNN	Event detection
(T. Liu et al., 2017)	CNN+LSTM	Event detection
(Hong, Ling, & Ye, 2018)	CNN	Event classification
(Giancola et al., 2018)	CNN + NetRVLAD	Event spotting
(Cioppa et al., 2020)	CNN + NetRVLAD	Event spotting
(Vats et al., 2020)	A multi-tower temporal convolutional network	Event detection

### 2.3 Event Detection and Localization in Videos

Event understanding and activity recognition research focuses on analyzing events and activities in a video (Kautz et al., 2017; Kumar & John, 2016; Lu, Shi, & Jia, 2013; Ma et al., 2013; Xueming Qian et al., 2010; Ramanathan et al., 2016; K. Tang, Fei-Fei, & Koller, 2012; Ullah, Ahmad, Muhammad, Sajjad, & Baik, 2017; C.-h. Wang, Wang, & Guan, 2011; Yeung, Russakovsky, Mori, & Fei-Fei, 2016). It is either defined as localizing an event in an untrimmed video or classifying video segments according to a set of predefined event classes, mostly referred to as action/activity recognition. Event localization and activity recognition have been applied in different domains such as sports videos, movie clips, and surveillance videos. Other than finding a solution for event localization in videos, it is also important to define an event or activity in terms of the boundaries and characteristics.

One important difficulty is the subjective nature of the event definition. Different people may define an event or an action in different ways. Also, the definition could be domain dependent. One common process is to define temporal segments that will be classified into different event classes using predefined labels (Buch, Escorcia, Shen, Ghanem, & Carlos Niebles, 2017; Caba Heilbron, Carlos Niebles, & Ghanem, 2016; Gao, Yang, Chen, Sun, & Nevatia, 2017; Shou, Wang, & Chang, 2016; Z. Wang et al., 2016).

In addition to the above, the duration of an event can also vary subject to subject. For example, in a football video, while one person can define the goal event to be precisely the time the ball passed the goal line, another user might consider an interval where the ball is heading to the goal line till it hits the net. Recently, (Sigurdsson, Russakovsky, & Gupta, 2017) conduct a semi-large experiment on different algorithms applied on different datasets with different data annotations. Based on the results, authors argue that defining temporal boundaries for an action is an ambiguous task. In another similar

evaluation, (W. Chen, Xiong, Xu, & Corso, 2014) questioned the concept of action, action boundaries, and what are the differences between an action and a motion. In their paper, they define four aspects for an action: 1) Movement that an agent can do, 2) It requires an intention, 3) It requires a bodily movement and 4) An action has side effects on the environment it is being performed. Others, such as (X. Dai, Singh, Zhang, Davis, & Qiu Chen, 2017) simply define actions as set of defined clips with a start and end position.

The concept of an event is even more vague in multimedia communities. In (Awad et al., 2016) , the authors define an event for multimedia event detection system as a kit which consists of 1) a unique title for the event, 2) a textual description of the event, 3) an expression of some event knowledge needed for a human to understand and perform the event, and 4) a set of video examples which demonstrates the event. In addition, a specific event may be described with a specific rule for start and end.

In the context of live sports video broadcasts, defining an action boundary is a complicated task. For example, as mentioned above, the beginning and the end of events, such as scoring a goal (e.g., two-point or three-point field goal in basketball) is subjective. Similarly, temporal boundaries of a slam dunk in basketball and scoring a point/serving in volleyball is not clearly defined.

Due to this ambiguous definition, different work might have different annotations which results in a different definition of an action. For example, authors in (Ramanathan et al., 2016) define a basketball shot as a three seconds action, where the action starts three seconds before the exact instant that a ball crosses the hoop. Similar to the other sports broadcasts, in football, it is also not clear how to objectively define the start and end boundaries of events such as scoring a goal, issuing a card, or substitution. We agree with the authors in (Giancola et al., 2018) that the anchor of each of these events in sports are well defined as a single time instance. Still, as it is mentioned by (Ramanathan et al.,

2016) defining a boundary around the anchor of events with a fixed duration would be subjective.

Some recent datasets such as THUMOS14 (Y.-G. Jiang & Sukthankar, 2014), ActivityNet (Caba Heilbron, Escorcia, Ghanem, & Carlos Niebles, 2015), and Charades (Sigurdsson et al., 2016), try to address the ambiguity problem of temporal boundaries by asking multiple annotators to annotate the same video. Also, AVA (Gu et al., 2018) attempts to resolve the atomic characteristic of actions by providing well-defined annotations within a 3 seconds duration. While the results of these annotations are aggregated to a single annotation, it still does not resolve the main core issue which is the resulting event boundary is not agreed upon.

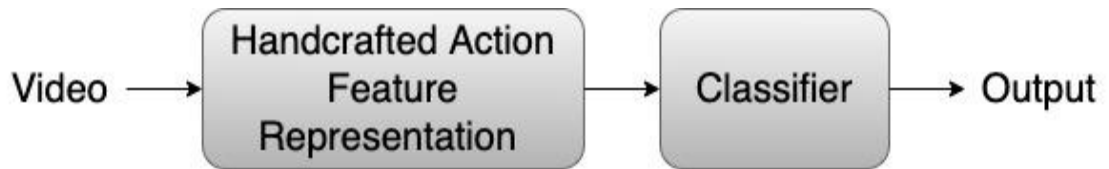
Unlike previous approaches, authors in (Giancola et al., 2018), provided an alternative definition for event localization. In their definition, an event is anchored around a single time instance and the detections will be validated based on a pre-specified error tolerance. This is referred to as event spotting. In our work, we follow a similar concept of event spotting proposed by (Giancola et al., 2018). In other words, rather than identifying the boundaries of an action within a video and looking for Intersection-over-Union (IoU) between temporal windows, spotting identifies the moment that an event occurs. A candidate spot is positive if it lands within a tolerance window around the anchor of an event. Otherwise, it is considered as negative. Note that, we do not claim this is the best definition for event localization for all types of events in videos. Instead, we believe this is a more reasonable definition for sports videos, where the actions/events happen at a glimpse.

### **2.3.1 Classical Approaches**

Video analysis (such as object detection in videos, activity recognition and detection, sports video analysis, etc.) has been an active area of research since the formation of the

computer vision field. Several successful classical approaches have proposed to use a combination of handcrafted feature extraction methods (Figure 2.2) and shallow classification models. The handcrafted features include SIFT-Scale-Invariant Feature Transform (Lowe, 2004), SIFT 3D (Scovanner, Ali, & Shah, 2007), Histograms of Oriented Gradients-HOG (Dalal & Triggs, 2005), HOG3D (Klaser, Marszałek, & Schmid, 2008), Histogram of Optical Flow-HOF (Chaudhry, Ravichandran, Hager, & Vidal, 2009) which are all based on histogram descriptors, and Spatiotemporal Interest Points- STIP (Rapantzikos, Avrithis, & Kollias, 2009) and the classifier is usually a shallow model such as support vector machines –SVM (X. Yang, Zhang, & Tian, 2012), K-nearest neighbor - KNN (Efros, Berg, Mori, & Malik, 2003), logistic regression and Hidden Markov Models (C.-h. Wang et al., 2011) which consumes the handcrafted features to predict or detect the events in videos. (Kumar & John, 2016; X. Li, 2007) addressed human activity detection problem in videos. (X. Li, 2007) adopted the concept of the histogram of oriented gradients in images and proposed oriented histograms of optical flow for feature extraction. Using these feature vectors and HMM, the authors proposed human motion descriptors in videos. In a similar work, (Efros et al., 2003; Kumar & John, 2016) use the HOF proposed by (X. Li, 2007). While (Kumar & John, 2016) uses a multi-class SVM classifier to classify/recognize human-human interactions, (Efros et al., 2003) benefits from KNN for classifying human actions in videos. Using an alternative approach which is based on appearance features, authors in (X. Yang et al., 2012) used SVM classifier on top of HOG features, extracted from video frames, for action recognition in broadcast tennis and golf videos. Using a similar approach, (Oreifej & Liu, 2013) applied SVM classifier on top of the histogram of normal orientation for faster human activity recognition. (X. Yang & Tian, 2012) proposed a different approach by using an action recognition system based on the Eigen representation of human joints and NBNN classifier. In (Efros et al., 2003) authors use optical flow histograms and KNN

to detect the motion of a player in football games. Note that in almost all of these approaches, a dimensionality reduction technique (such as principal component analysis – PCA) is applied on the original features to improve the classification performance.



**Figure 2.2 : Classical computer vision methods based on handcrafted features.**

### 2.3.2 Deep Learning Approaches

In recent years, the computer vision community benefitted from deep learning-based approaches (Figure 2.3) significantly. Specifically, in video analysis, researchers have applied CNN and RNN architectures to a wide range of problems including action recognition and detection (Baccouche et al., 2010; L. Chen et al., 2017; Ibrahim et al., 2016; Jiang et al., 2016; Karpathy et al., 2014; Ramanathan et al., 2016). Most of these models, notably the CNN-based models, are inspired by successful applications of similar architectures in image analysis domain. The number of prior works which have used deep learning for video analysis is more than a few thousand papers, and it is out of the scope of this work to provide a comprehensive review of all deep learning-based models in this area. Instead, we present the most relevant work which are relevant to ours with respect to the model and architecture.

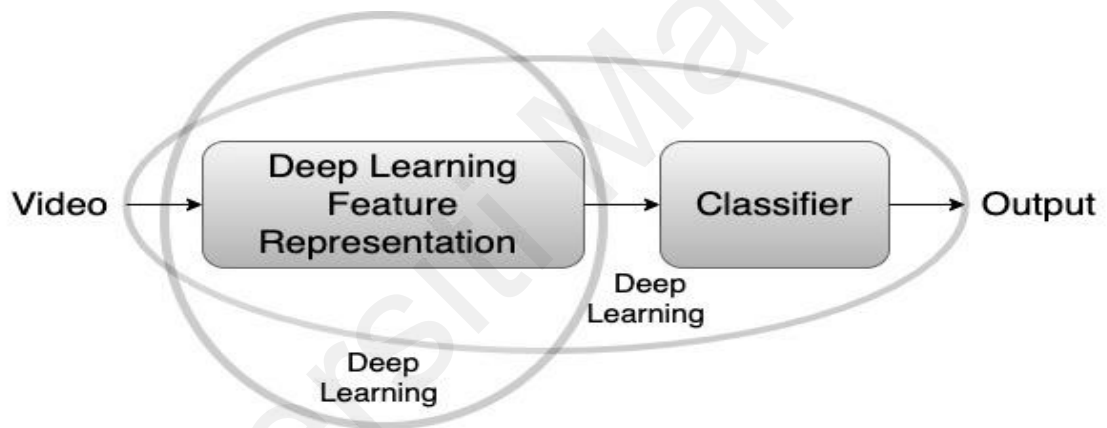
In the context of sports analysis, various approaches have been proposed to address different problems in sports videos. In basketball, (L. Chen et al., 2017) used a combination of CNN and SVM for utilizing weakly supervised data for action localization in basketball videos, and (Ramanathan et al., 2016) utilized the same architecture for events and key actor detection in multi-person basketball videos. (Ibrahim et al., 2016) build a deep model based on RNN-LSTM for activity recognition in volleyball videos.

Authors in (Karpathy et al., 2014) introduced a novel approach which uses two stream multiresolution CNN architecture to capture the important spatiotemporal features and to classify videos. In another effort, (Baccouche et al., 2010) used an RNN build on top of handcrafted features (BoW and SIFT) to perform action classification in football videos. Similarly, authors in (Jiang et al., 2016) proposed to ground RNN on top of CNN features for event detection in short football videos instead.

Deep neural network-based models have been extensively applied in other domains in non-sport video analysis. While a large number of approaches have relied on non-temporal models using different variants of CNNs (e.g., two-stream CNNs (W. Dai, Chen, Huang, Gao, & Zhang, 2019; Feichtenhofer, Pinz, & Zisserman, 2016; C. Li, Wu, Zhao, Cao, & Tang, 2018; K. Simonyan & A. Zisserman, 2014; Limin Wang, Xiong, Wang, & Qiao, 2015; H. Xu, Das, & Saenko, 2019), others have built temporal models using RNN based architectures for temporal analysis of videos (Q. Dai et al., 2015; Fan, Lu, Li, & Liu, 2016; Ullah et al., 2017; Zhao, Ali, & Van der Smagt, 2017). In (H. Xu et al., 2019) the authors addressed the problem of activity detection in 3D videos by using a two-stream CNN architecture. For violent scene detection and affective impact prediction in videos, authors in (Q. Dai et al., 2015) used a combination of two-stream CNN and RNN to capture the short and long-term dependencies and SVM for scene classification. In similar approaches, (Ullah et al., 2017), (Fan et al., 2016) and (Zhao et al., 2017) utilized the combination of two-stream and RNN for action recognition in videos, video-based emotion recognition and action recognition in 3D videos respectively. Authors in (Lin Wang, Zhou, Li, Zuo, & Tan, 2018) proposed a hybrid autoencoder architecture based on the LSTM Encoder-Decoder and the convolutional Autoencoder for anomaly event detection. Their experimental results show better spatiotemporal feature extraction. They improved the extrapolate capability of the decoder. In a similar problem, Authors in (Yan, Smith, Lu, & Zhang, 2018) proposed a two-stream recurrent variational autoencoder for

anomaly event detection by capturing the spatiotemporal and optical flow features. Authors in (Nguyen & Meunier, 2019) proposed a CNN to address the problem of anomaly detection in surveillance videos by learning a correspondence between common object appearances. A generic deep one-class is used in (P. Wu, Liu, & Shen, 2019) to develop a framework for event detection.

To summarize, we identify two groups of approaches and methods. On one side, there are approaches that build classifiers using RNNs which is grounded on top of hand-crafted features. On the other side, there are approaches that build the classifiers and underlying feature extraction layers together using RNNs and CNNs.



**Figure 2.3: Deep learning-based models outline.**

## 2.4 Review of Machine Learning

In this section, we will provide a review of the most relevant machine learning approaches used in computer vision. Not surprisingly, we categorize these models to two broad spectrums of models: “Classical machine learning methods with handcrafted features” and “Deep-learning based methods”.

### 2.4.1 Classical Machine Learning with Handcrafted Features

This group of methods have been studied for a long time in the computer vision community. The common theme behind these methods is that the process of extracting



features from the raw input data is separated from the classification/reasoning process. As for the feature extraction, the design of the features are done by experts in the field and is usually formulated as a predefined mathematical formulation without any learnable component. Table 2.2 presents a summary of relevant work which have used classical features and machine learning methods in computer vision research. Note that while these models have showed successful results in a subset of problems, they usually suffered from generalization to real-world setting and slow computation. We reviewed the relevant applications of classical methods in activity detection and recognition in Section 2.3.1. Since our methods are not based on hand-crafted features nor classical machine learning, the comprehensive review of classical machine learning is out of the scope of this thesis. Instead, for completeness, we provide a high-level review of the combination of these approaches in the following table.

**Table 2.2: Summary of classical ML methods and hand-crafted features.**

<b>Feature</b> <b>Method</b>	<b>HOG</b>	<b>HOF</b>	<b>MBH</b>	<b>SIFT</b>	<b>STIP</b>	<b>Others</b>
<b>KNNs</b>	(Yuanyuan Huang, Yang, & Huang, 2012; Serpush & Rezaei, 2020; Shri & Jothilakshmi, 2018)	(Efros et al., 2003)	N/A	N/A	(Rapantzi et al., 2009)	(Maheswari & Ramakrishnan, 2015; Zhan, Liu, Gou, & Wang, 2016)

**Table 2.2 Continued.**

<b>Feature</b> <b>Method</b>	<b>HOG</b>	<b>HOF</b>	<b>MBH</b>	<b>SIFT</b>	<b>STIP</b>	<b>Others</b>
<b>HCRFs/ CRFs</b>	N/A	N/A	N/A	N/A	N/A	(X Qian et al., 2012; Jin Wang, Liu, She, & Liu, 2011; T. Wang et al., 2006)
<b>HMMs</b>	(C.-h. Wang et al., 2011)	(X. Li, 2007)	(Sun & Nevatia, 2013)	N/A	N/A	(Assfalg et al., 2002; Xie, Chang, Divakaran, & Sun, 2002)
<b>SVM</b>	(Dalal & Triggs, 2005; C.-P. Huang, Hsieh, Lai, & Huang, 2011; X. Yang et al., 2012)	(Kumar & John, 2016)	(Dalal, Triggs, & Schmid, 2006; H. Wang, Kläser, Schmid, & Liu, 2011, 2013)	(Scovanner et al., 2007; J.-T. Zhang, Tsoi, & Lo, 2014; Zhou et al., 2008)	(Thi, Zhang, Cheng, Wang, & Satoh, 2010)	(Jinjun Wang, Xu, Chng, Wah, & Tian, 2004)
<b>SSVM</b>	(Lan, Sigal, & Mori, 2012a)	N/A	N/A	N/A	N/A	(Soomro, Idrees, & Shah, 2018; Todorovic & Mahasseni, 2013)

**Table 2.2 Continued.**

<b>Feature</b> <b>Method</b>	<b>HOG</b>	<b>HOF</b>	<b>MBH</b>	<b>SIFT</b>	<b>STIP</b>	<b>Others</b>
<b>Others</b>	N/A	(Chaudhry et al., 2009; Lertniphonphan, Aramvith, & Chalidabhongse, 2011; Raptis & Sigal, 2013)	N/A	N/A	(P. Liu, Wang, She, & Liu, 2011)	N/A

To summarize, as it is presented in Table 2.1 and Table 2.2, while **prior to 2015** classical machine learning approaches dominated the research community, recent advances in deep learning techniques resulted in a paradigm shift in the community. The state-of-the-art computer vision techniques heavily benefit from these advances and improve the accuracy of various problems including the activity recognition and detection. Following this paradigm shift, our approach is mainly built on top of the most recent successful deep learning models.

#### **2.4.2 Deep Convolutional Neural Networks**

Deep neural networks refer to a class of machine learning methods that are inspired by our understanding of the human brain. These methods are particularly powerful in automatic feature detection. The main theme behind these methods is that the network architecture itself also learns how to extract the features from the raw data. Two categories of deep neural networks have been extensively used in the past few years: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). As we

heavily rely on these two categorize, the following two subsections provide a detailed summary of both.

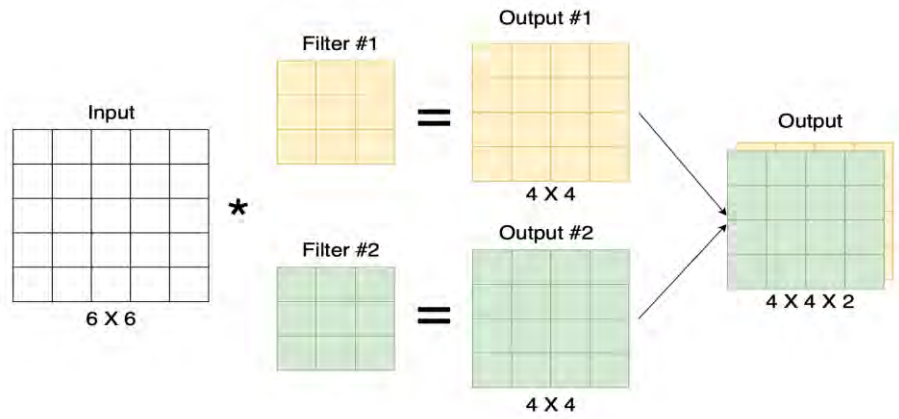
#### **2.4.2.1 Convolutional Neural Networks Fundamental**

A convolutional neural network (CNN, or ConvNet) is a type of deep neural network proposed by (LeCun et al., 1998) in 1998. It is inspired by the human visual system and has been mostly applied in visual image analysis.

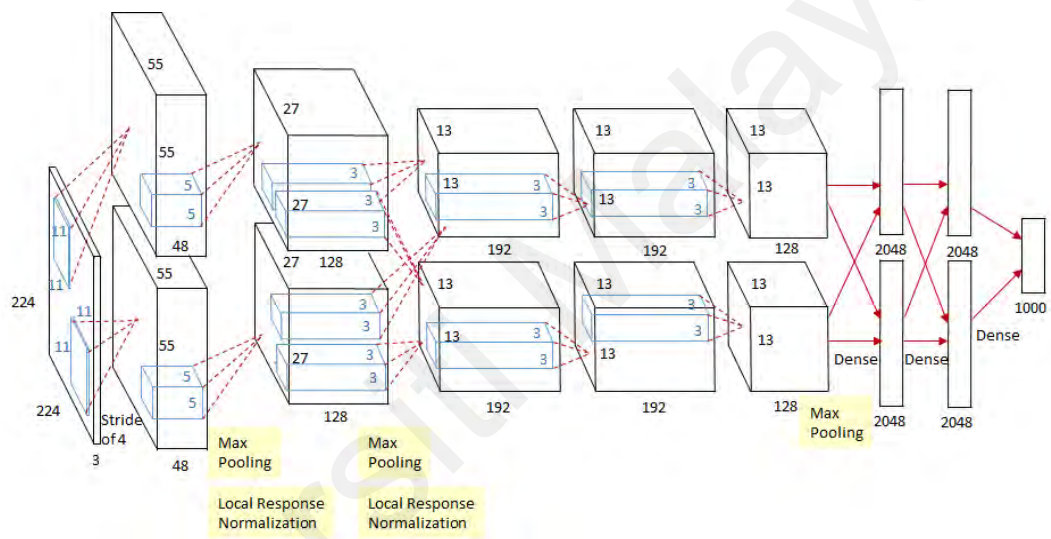
While the underlying theory is similar to standard feed-forward neural networks such as multi-layer perceptron, a convolutional neural network avoids the dense connections to prevent overfitting. To the contrary, convolutional neural networks limit the connectivity of the neurons between two consecutive layers to a small local neighborhood of each neuron. This reduces the number of network parameters which effectively results in better generalizability and less overfitting. Figure 2.4 illustrates the overall idea behind the convolutional filters.

A deep convolutional neural network (DCNN) is designed by concatenating multiple layers of convolutional filters and features. The first modern Deep CNN model, 'AlexNet', was proposed in (Krizhevsky, Sutskever, & Hinton, 2012). Figure 2.5 shows the AlexNet network architecture. After the success of 'AlexNet', multiple other DCNN architectures are purposed. The most famous ones are the VGG (K. Simonyan & A. J. a. p. a. Zisserman, 2014) and ResNet (He, Zhang, Ren, & Sun, 2016).

While CNN was originally designed for 2D image analysis, extensions of 2D convolutions to 3D convolutions, were proposed for three-dimensional data including videos and 3D point clouds. Unlike the 2D convolution, in addition to spatial convolution operation, the filters are convolved in the temporal domain as well.



**Figure 2.4 : An example of 2D convolutional filter**



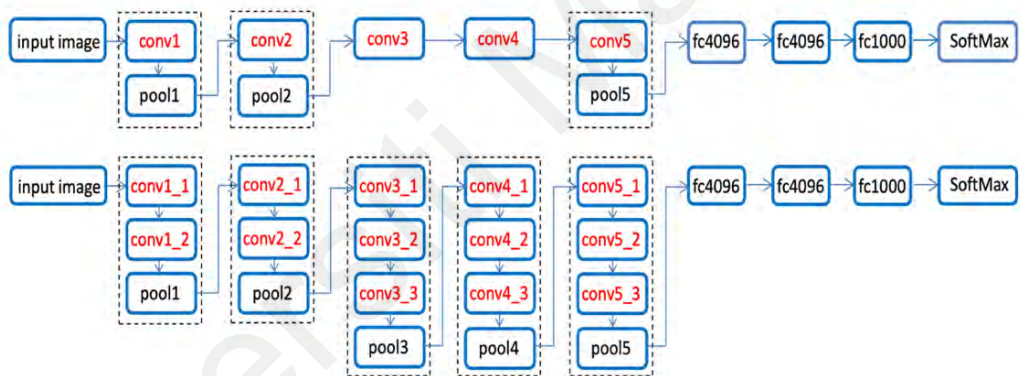
**Figure 2.5: The AlexNet network architecture proposed in (Krizhevsky et al., 2012).**

### 2.4.2.2 VGG Network

VGG is a CNN architecture for object recognition and classification tasks. It refers to “Visual Geometry Group” and developed in 2014 by (K. Simonyan & A. J. a. p. a. Zisserman, 2014) at Oxford Robotics Institute and submitted to the large-scale image recognition challenge (ILSVRC2014<sup>7</sup>). At the time, the proposed model scored the first and the second places in the localization and classification tracks respectively. Originally,

<sup>7</sup> <http://image-net.org/challenges/LSVRC/2014/>

it was introduced after the success of AlexNet (Krizhevsky et al., 2012). The main important difference is the use of 3X3 kernel-size filters in all layers which allows the model to use a fewer number of parameters at each layer. This enabled the authors to design a deeper network with 16 layers. More number of layers help the network to learn better and more abstract feature representations at higher layers. Also, deeper networks have larger receptive fields<sup>8</sup> which improves the classification accuracy. On ImageNet (J. a. D. Deng, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L., 2009) dataset, VGG achieved 92.3% in the top-5 accuracy metric. Figure 2.6 illustrates compares VGG and AlexNet structures. Following the original paper, another variation was proposed with 19 layers which was called VGG-19 (vs the original VGG-16).



**Figure 2.6: Architecture of AlexNet vs. VGG-16. Top: Architecture of AlexNet, Bottom: Architecture of VGG-16 (Yu et al., 2016).**

### 2.4.2.3 ResNet Network

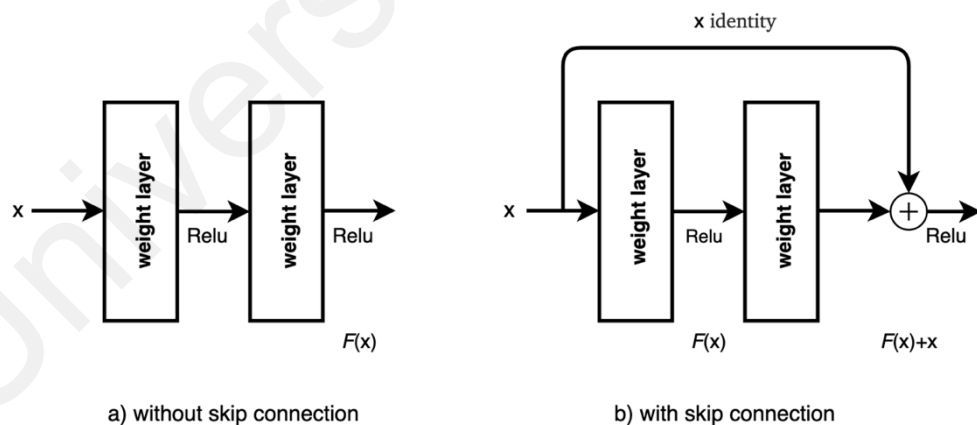
Since AlexNet, the CNN architectures such as VGG proposed to use more layers which result in deeper networks. For example, VGG-19 has 19 layers while AlexNet had only five convolutional layers. However, increasing network depth negatively affects the training and convergence of the learning mostly due to the vanishing gradient problem.

<sup>8</sup> Effective area of the input image visible by the hidden/output layers

A residual neural network (ResNet) is a neural network architecture proposed by (He et al., 2016). After AlexNet winning at the LSVRC2012<sup>9</sup> ResNet became the most novel network architecture in the computer vision community by winning the 1st place of ImageNet challenge on the ILSVRC2015<sup>10</sup> for the classification task.

The fundamental difference of ResNet is that it allows successful training of extremely deep neural networks (more 150 layers) by introducing skip connections. These skip connections add the original input to the output of the convolution block. Figure 2.7 illustrates the skip connection concept. To summarize, skip connections have the following benefits:

1. Overcome the prior difficulties due to the vanishing gradients problem by allowing an alternate shortcut path for the gradient to flow through.
2. Allow the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer, and not worse.



**Figure 2.7: Visualization of the skip-connection concept in neural networks.**

<sup>9</sup> <http://www.image-net.org/challenges/LSVRC/2012/>

<sup>10</sup> <http://www.image-net.org/challenges/LSVRC/2015/>

As it is shown in Equation 2.1, the identity shortcut ( $x$ ) can be directly used when the input and output are of the same dimensions.

$$y = F(x, \{W_i\}) + x \quad (2.1)$$

If the input and output dimensions are different, the shortcut still performs identity mapping, and a projection shortcut is used to match the dimension using the Equation 2.2.

$$y = F(x, \{W_i\}) + W_s x \quad (2.2)$$

#### 2.4.2.4 Two-Stream Convolutional Neural Network

Compare to images, in addition to spatial information, videos provide another important clue referred to as the temporal component based on the motion. To benefit from the motion information, inspired by a similar approach or technique in the domain of image analysis, a large number of video analysis methods are proposed. One classical example is the HOF which is an extension of HOG to “spatiotemporal” domain.

Deep learning is not an exception to this phenomenon. (Karpathy et al., 2014) introduced a novel approach by proposing two networks for capturing the important spatiotemporal features present in videos. For the low-resolution frames, they used a “context stream” to capture the important features and for the high-resolution middle region of the frame, they used a “fovea stream” to capture the important detailed features. The information from these two streams is subsequently fused to provide a more descriptive feature of the video content. Despite improved results, motion related features from temporal axis were not considered for learning. Authors in (K. Simonyan & A. Zisserman, 2014) extended Karpathy’s ideas by explicitly designing one network to capture **spatial features** (only using static images) and another network to extract **temporal features** (using opticalflow streams). Both networks are trained simultaneously in an end-to-end learning fashion.

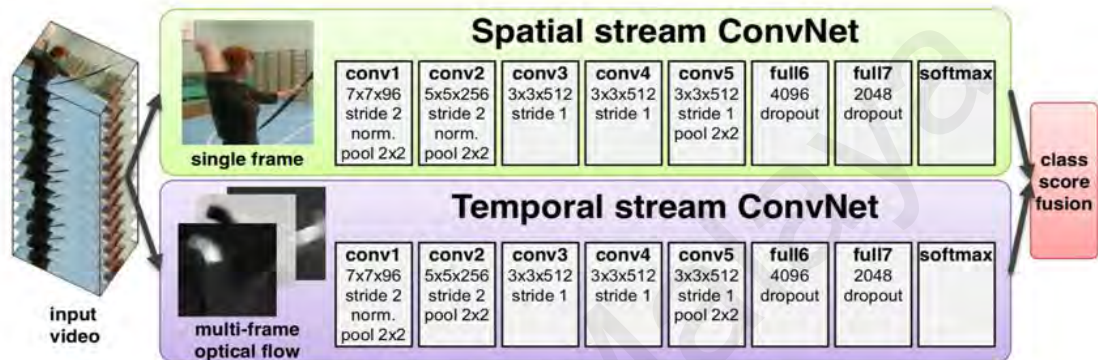


As it is stated in (K. Simonyan & A. Zisserman, 2014), breaking up the spatial and temporal subnetworks has two benefits:

1. It closely resembles the human visual system. Based on the studies in neuroscience, our visual system processes what we observe through: a) the ventral stream, which is responsible for processing spatial information (such as shape and color), and b) the dorsal stream which process the motion information (Goodale & Milner, 1992; Kruger et al., 2012). Inspired by the human visual system, we can decompose a video into the spatial and temporal components in a similar way. The spatial-stream is responsible for the image contents such as objects, colors, and shapes while the motion information across video frames is processed in the temporal-stream. For better clarification, consider the differences between the “card event” and “substitution event”. While the specific hand movement is captured by the temporal stream, the object in hand (card or the substitution board) differences are captured in the spatial stream. This provides more evidence for the reasoning system to distinguish between these two events.
2. From a practical perspective, separating these two streams, enables us to leverage a large amount of training data for image analysis task to train the spatial network (e.g., ImageNet). This reduces the training time and a need for large scale video datasets for appearance generalization.

Given a video ( $v_i$ ), the standard two-stream CNN (K. Simonyan & A. Zisserman, 2014) is applied to extract features from video frames. The two-stream network introduces an architecture which consists of two subnetworks operating on spatial and temporal streams. The extracted features from each subnetwork are then combined by late fusion. The spatial stream subnetwork extracts feature for action recognition from still

video frames. On the other hand, the temporal stream computes the features from a dense opticalflow map. While in the original paper, the late fusion combines the results of the last SoftMax layer, in this work we propose to combine the features of the last fully connected layer before the SoftMax layer using an additional fully connected layer. This is shown in Figure 2.8.



**Figure 2.8: Original Two-Stream CNN proposed in (K. Simonyan & A. Zisserman, 2014).**

Following the success of the two-stream networks (K. Simonyan & A. Zisserman, 2014), multiple research work applied a similar network architecture and have shown state-of-the-art results in their problem domain (W. Dai et al., 2019; Feichtenhofer, Pinz, & Wildes, 2017; H. Xu et al., 2019; Zhao et al., 2017).

We want to emphasize that inspired by the original CNN introduction (LeCun et al., 1998), multiple modern CNN architectures are proposed to improve the efficiency and accuracy of the CNN models. It is not possible to go over the details of each of these networks. Instead, we summarize the most successful CNN architectures in Table 2.3.

**Table 2.3: Comparison between different CNN architectures.**

<b>CNNs Architectures</b>	<b>Summary</b>	<b>Advantage</b>	<b>Disadvantage</b>
AlexNet (Krizhevsky et al., 2012)	First successful CNN to use GPU, Rectified Linear Units, and Dropout	Simple	Large number of parameters, overfitting
GoogleNet, Inception (Szegedy et al., 2015)	Introduced multiple kernel sizes in each layer called Inspection Layer	Flexible kernel sizes	Average number of parameters
ResNet (He et al., 2016)	Introduced the skip connection and the residual learning	Best results, Fewer number of parameters	Hard to train from scratch
VGG (19, 100) (K. Simonyan & A. J. a. p. a. Zisserman, 2014)	The first attempt to have very deep neural networks with more than 16 layers of 3X3 convolutions	Small number of parameters dues to using 3x3 filters only	Hard to train when the network is deep
Two-Stream CNN (K. Simonyan & A. Zisserman, 2014)	The first 2D CNN, designed specifically for videos	Uses the opticalflow information, Suitable for video analysis	More parameters due to two parallel networks
Conv3D (Ji, Xu, Yang, & Yu, 2012)	A successful extension of 2D CNNs for videos.	Uses 3D convolutional filters, Suitable for video analysis	Average number of parameters, slower compared to 2D convolutions

As it is presented in Table 2.4, Two-stream CNN and Conv3D based models are both specifically designed for video analysis and they have shown promising results. One important practical concern regarding the Conv3D approaches is that in a setting where the length of the video is long, Conv3D approaches usually suffer from slow training and inference. This is specifically a concern for us working with long football videos. In addition, as we rely on pre-trained models and fine-tuning for our CNN based models.

**Table 2.4: Comparison between different feature extraction techniques.**

Approach \ Data set	NTSEL (%)	NDRDB (%)	UCF101 mAP (%)	HMDB mAP (%)
<b>Handcrafted based</b>				
iDT (HOG)	70.18	50.43	-	-
iDT (HOF)	64.76	52.05	-	-
iDT (MBH)	65.38	49.12	-	-
iDT+FV	-	-	85.9	57.2
iDT+HSV	-	-	87.9	61.1
<b>CNN based</b>				
DeCAF (ImageNet with VGG-16)	53.63	<b>50.54</b>	-	-
Two-stream ConvNet (Spatial)	69.04	48.47	-	-
Two-stream ConvNet (Temporal)	64.05	45.93	-	-
Two-stream ConvNet	<b>85.44</b>	<b>50.50</b>	-	-
TDD(ZFNetNet)	-	-	90.30	63.20
Two-stream (CNN-M)	-	-	88	59.40
Improved Two-stream (VGG-16)	-	-	92.50	65.40
ActionVLAD (VGG-16)	-	-	92.70	<b>66.90</b>
LSTM/ConvPooling (GoogleNet)	-	-	88.60	-
Residual Two-stream (ResNet-50)	-	-	<b>93.40</b>	<b>66.40</b>
CNN-based action recognition (3D CNN)	-	-	90.80	63.60

Note that while prior work (Gavrilyuk, Ghodrati, Li, & Snoek, 2018; Sultani, Chen, & Shah, 2018; Tran, Wang, Torresani, & Feiszli, 2019; Tran et al., 2018; H. Xu, Das, & Saenko, 2017) use networks with 3D convolutions for some video analysis problems, authors in (Giancola et al., 2018) and (Cioppa et al., 2020) have shown that using ResNet (He et al., 2016) features outperform the C3D (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015) and I3D (Carreira & Zisserman, 2017) features for event spotting.

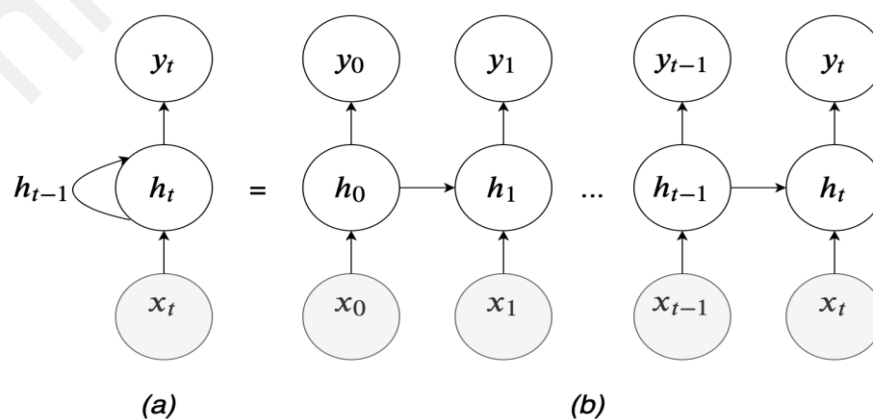
### 2.4.3 Recurrent Neural Networks

Unlike images, videos contain temporal information which can be utilized for video analysis problems. Feed-forward neural networks are not capable of consuming the available temporal information, mainly because they assume independence between observations at different time steps. This is a significant shortcoming which limits the application of these models to non-sequential data. Consider the problem of classifying

words in a sentence. Inherently, there is a correlation between words in a sentence. One approach is to use all the words in a sentence as an input to a feed-forward network. Unfortunately, this is not possible for two reasons. First, the number of words in a sentence vary from sentence to sentence. Second, even if we fix the number of inputs to a large constant value and assume that the network can handle it, this approach is for sure not extendable for a larger sequence of words such as paragraphs or documents.

Recurrent Neural Networks (RNNs) (Hochreiter & Schmidhuber, 1997) is a different type of neural network which is designed to address this issue. An RNN contains feedback loops, which allows the network to memorize previous observations as abstract information. This is achieved through a feedback signal from the previous state  $h_{t-1}$ . At each time step, the network consumes the current input  $x_t$  and the previous abstract information  $h_{t-1}$  and generates the current abstract information  $h_t$ . Figure 2.9(a), shows the overall architecture of an RNN.

In practice, an unfolded RNNs through time and an extension of the backpropagation algorithm called backpropagation through time is used to train the network parameters. This is shown in Figure 2.9(b).



**Figure 2.9: Unrolled recurrent neural network structure.**

More specifically, the Equation 2.3 and 2.4 provide the detail mathematical formulations of a vanilla RNN (Hochreiter & Schmidhuber, 1997)




$$h_i = \sigma_h(W_{hh}h_{i-1} + W_{hx}x_i + b_h). \quad (2.3)$$

$$\hat{y}_i = \sigma_y(W_{yh}h_i + b_y). \quad (2.4)$$

where  $W_{hx}$  is input to hidden layer weight,  $W_{hh}$  is hidden to hidden layer weight, and  $W_{yh}$  is hidden to output weight and ( $\sigma_h$  and  $\sigma_y$ ) are activation functions.

Sequential data analysis, which includes classification and prediction, has received a significant amount of attention in the machine learning and artificial intelligence research community. In recent years, RNNs have been applied to multiple sequential data problems. Table 2.5, provides a brief overview of these applications.

**Table 2.5: Example applications of RNN in various domains.**

Problem domain	X	→	Y	Reference
Speech recognition			“My dog was running around the house”	(Graves, Mohamed, & Hinton, 2013)
Music generation	∅			(J. Wu, Hu, Wang, Hu, & Zhu, 2019)
Sentiment classification	There is nothing to like about this music.		★☆☆☆☆	(D. Tang, Qin, & Liu, 2015)
DNA sequence analysis	AGCCCCTGTGAG GAACTAG		AGCCCCTGTGAG GAACTAG	(Quang & Xie, 2016)
Machine translation	bonne après-midi		Good afternoon	(Kalchbrenner & Blunsom, 2013)
Video activity recognition			Walking/Waiting	(Z. Deng, Vahdat, Hu, & Mori, 2016)
Name entity recognition	Yesterday Sara met George		Yesterday Sara met George	(W. Wang, Bao, & Gao, 2016)

While RNNs have been successfully applied to a variety of problems, training RNNs is challenging and difficult. This results in various limitations in applying RNNs in certain settings and problem domains. In the next chapters, we will explain the RNN limitations and will discuss the possible solutions for these difficulties.

#### **2.4.3.1 Training Difficulties**

While recurrent models have been successfully applied to activity recognition and detection problems (Donahue et al., 2015), researchers have identified multiple issues with training standard RNNs (Pascanu, Mikolov, & Bengio, 2013). The most important observation is presented by (Bengio, Simard, & Frasconi, 1994). The authors have identified two important technical issues with backpropagation through time (BPTT) algorithm, referred to as the “vanishing” and “exploding” gradient problems. The high-level explanation is that, during the backpropagation, due to numerical instability, the gradients backpropagated to deep layers (further away time steps in RNN) are either too small (vanish) or too large (explode). While the vanishing gradient problem causes the early layers not to learn anything, the exploding gradient problem causes inconsistent learning which prohibits convergence. In summary, this results in gradient-based optimization methods fail to capture hidden long-term dependencies, and mostly effected by short-term dependencies. This makes it hard to apply RNNs for very long data sequences.

Another important challenge with standard vanilla RNN is that memorizing very long dependencies is hard to achieve while keeping track of the mid-range and short-range dependencies. Last but not least, due to the sequential nature of the training, training RNNs takes longer than standard feedforward networks (Pascanu et al., 2013) to converge.

To address the above-mentioned problems, three main research directions have been explored by the researchers. The first category of the research is focused on finding better optimization algorithms (Bengio, Boulanger-Lewandowski, & Pascanu, 2013; Martens & Sutskever, 2011; Pascanu et al., 2013). Most of these approaches are considered as extensions of stochastic gradient descent. In addition, a large number of practical heuristics have been also employed during training. These efforts include 1) Clipped gradient approach, by which the norm of the gradient vector is clipped, 2) Using different activation functions compared to standard (tanh, sigmoid) activation functions with more stable gradients, 3) Using moment based gradient descent methods which may be less sensitive to learning rate. Techniques such as dropout have also been applied to recurrent connections or input connections in RNNs to improve generalization.

The second category of approaches are focused on designing more sophisticated hidden units for the RNNs. The pioneer method in this direction resulted in a successful recurrent unit called Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997). LSTM introduces the concept of memory and gates to regularize the behavior of the recurrent unit. Recently, a similar hidden unit is also proposed which simplifies the original LSTM unit by removing the memory and reducing the number of gates. It is referred to as a gated recurrent unit (GRU) which was proposed by Cho (Chung, Gulcehre, Cho, & Bengio, 2014). Overall, these types of hidden units have shown superior performance to vanilla RNNs in a variety of machine learning applications such as computer vision (Donahue et al., 2015), speech recognition (Graves, Jaitly, & Mohamed, 2013) or natural language processing (Cho et al., 2014).

Recently, the third category of approaches have been proposed which are based on more sophisticated recurrent connections. This includes (Chang et al., 2017; Y. Zhang et al., 2016) and (Campos, 2018). The core idea is to use skip connections to allow additional



feedback signals both in the forward and backward pass. Using skip connection can mitigate the issue of slow training. Combined with hierarchical RNNs, it also helps with memorizing longer dependencies.

To summarize, while in theory, the RNNs are capable of exploiting the information observed in the past, in practice, due to training difficulties, the final trained RNN capacity in modeling the observed history is usually limited in length.

### 2.4.3.2 Different Recurrent Units

In this section, we provide a detailed summary of the two most successful extensions to recurrent units used in the research literature.

**Long Short-Term Memories (LSTMs):** Long Short-Term Memory (LSTM), introduced in (Hochreiter & Schmidhuber, 1997), proposes a designated memory for RNN which allows the model to explicitly memorize the observed content in the past and choose when to use, modify or clear the memory. Figure 2.10 shows the architecture of an LSTM unit where  $x$ ,  $h$  and  $c$  stand for input state, hidden state and cell state respectively.

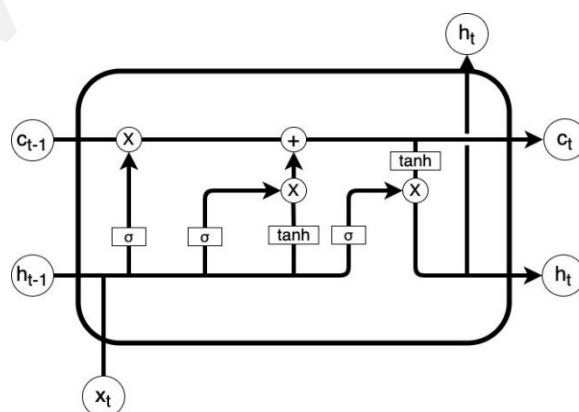
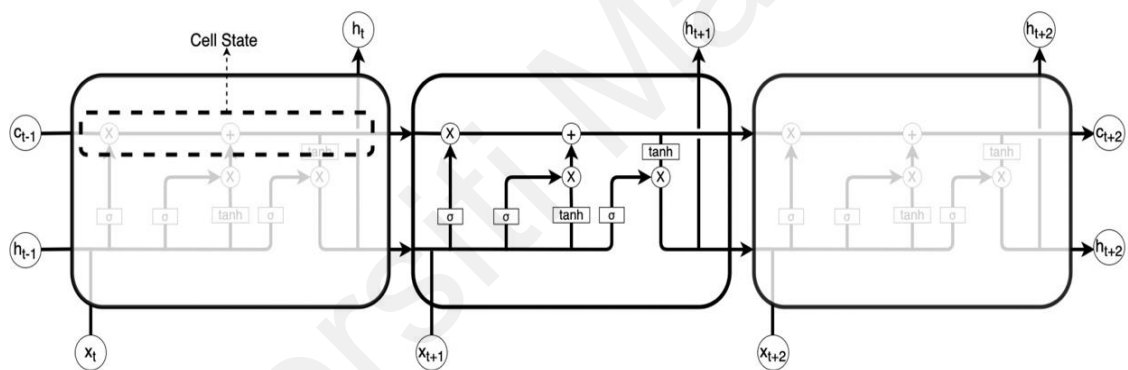


Figure 2.10: Overall architecture of the LSTM unit.

More specifically, an LSTM unit uses an explicit memory buffer and three gates (forget gate, input gate, and output gate) which control the information flow to and from the memory buffer.

In the following paragraphs, we provide a comprehensive explanation of each of the components and present the mathematical equations.

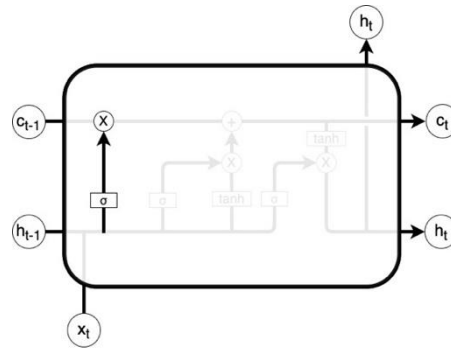
Figure 2.11, illustrates LSTM unit and its relation to the original recurrent neural network. To better understand the information flow, we provide a more detail explanation of the different gates and their goal as well as mathematical formulation that supports the ideas behind them in the following.



**Figure 2.11: Dilated version of single LSTM**

**1. Forget gate** — The value of the forget gate Figure 2.12 indicates how much information from past should be discarded or kept. Information from the previous state ( $h_{t-1}$ ) and the input content ( $x_t$ ) pass through a sigmoid function, which is a number between 0 and 1. If the outputs value is closer to one, the network preserves most of the previous memory content. If it is closer to zero, it replaces the current value and forgets most of the previous memory content. The mathematical equation is shown in Equation 2.5.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2.5)$$

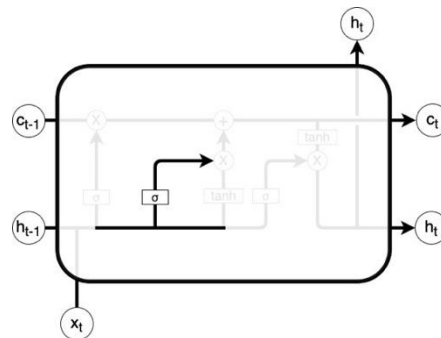


**Figure 2.12: Illustration of forget gate impact in LSTM.**

2. **Input gate** — The value of the input gate, shown in (Figure 2.13), identifies how much of the input content will be added to the memory. In other words, it indicates the contribution of the input value in the memory. To do this, the content of the input is multiplied by the output of the input gate before being added to the memory. If the output of the input gate is closer to one, it indicates the network should try to memorize most of the input, otherwise, the input will be suppressed before adding to the memory. The mathematical equations are shown in Equations 2.6 and 2.7.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2.6)$$

$$\bar{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (2.7)$$

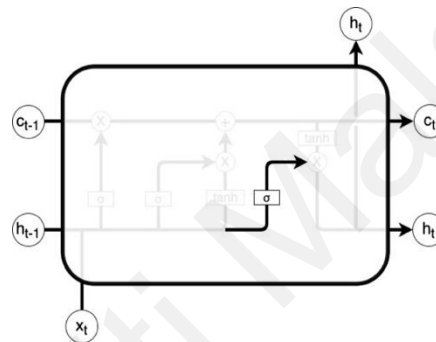


**Figure 2.13: Illustration of input gate impact in LSTM.**

**3. Output gate** — The output gate (Figure 2.14) identifies the amount of information in memory which is allowed to be visible outside of the LSTM unit. This is controlled by a sigmoid function where one indicates all the memory content should be visible and zero means no memory content is visible. The mathematical equations are shown in Equation 2.8 and 2.9.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (2.8)$$

$$h_t = o_t * \tanh(C_t) \quad (2.9)$$



**Figure 2.14: Illustration of output gate impact in LSTM.**

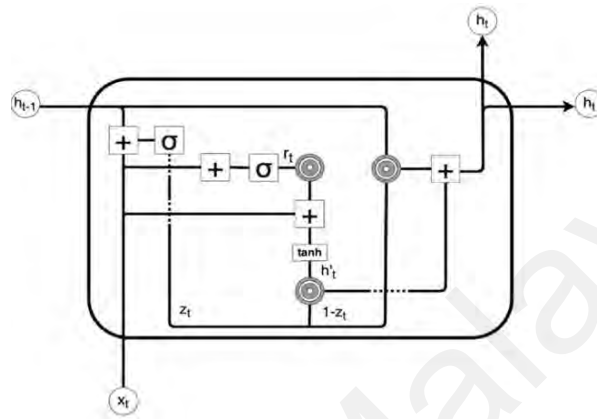
To summarize, the forget gate allows the model to clear the memory from previous information. The input gate allows the model to change the content of the memory based on the most recent processed observation. Finally, the output gate allows the model to choose how the memory impacts the current latent state which is visible to others.

Since its introduction, multiple improvements were suggested to the original LSTM cells. In this work we follow the implementation of LSTM as used in (Graves, Jaitly, et al., 2013).

**Gated Recurrent Units (GRUs):** This type of recurrent unit shares a lot of similarity with LSTM units and was introduced by (Cho et al., 2014). The most obvious difference is that the memory cell is removed. Instead, the information is assumed to be encoded in

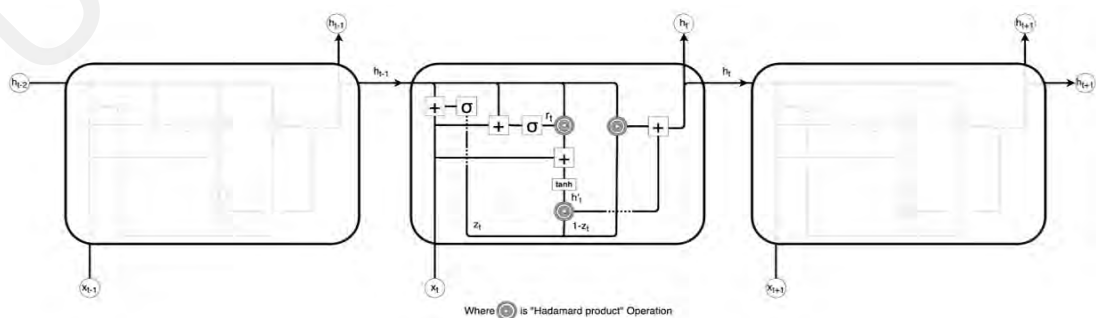
the hidden state itself. As a result of memory cell removal, the GRU does not have the output gate and the content of the hidden state is always visible.

Figure 2.15 shows the architecture of a recurrent neural network with gated recurrent unit where  $x, h$  stand for input state, hidden state respectively.



**Figure 2.15: Overall architecture of the GRU unit.**

More specifically, GRU is designed based on the LSTM's concept and has a similar structure. Similar to LSTM, the additional gate components support the information flow in RNN hidden states. GRU unit decides when a hidden state should be updated or when it should reset. This is pretty much like a simplified version of the LSTM and in some application domains GRU can even produce equally excellent results as LSTM units. Figure 2.16 illustrates a recurrent network with GRU units.

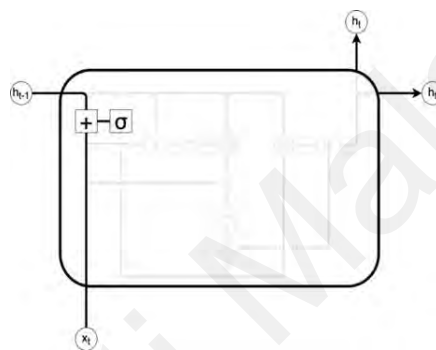


**Figure 2.16: Illustration of a single GRU in an unfolded RNN.**

In the following, we provide a more detail regarding each gate and its purpose.

**1. Update gate** - The update gate  $z_t$  (Figure 2.17) decides how much information from previous time steps need to be kept for the next time step. This is shown in the following equation where  $x_t$  and  $h_t$  are the input and hidden state and  $z_t$  is the update gate value. We will show the usage of the update gate later in the "final memory at the current time step" section. The mathematical equation is shown in Equation 2.10.

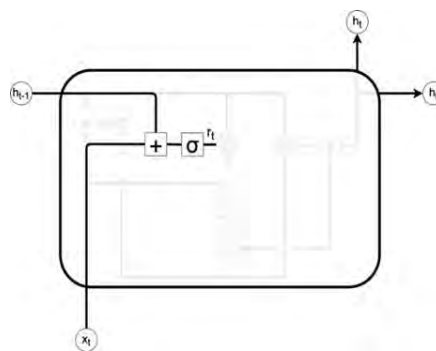
$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \quad (2.10)$$



**Figure 2.17: Illustration of update gate impact.**

**2. Reset gate** - The reset gate (Figure 2.18) decides how much of the past information needs to be forgotten. Basically, it is very similar to the forget gate in LSTM. This is shown in the Equation 2.11.

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \quad (2.11)$$

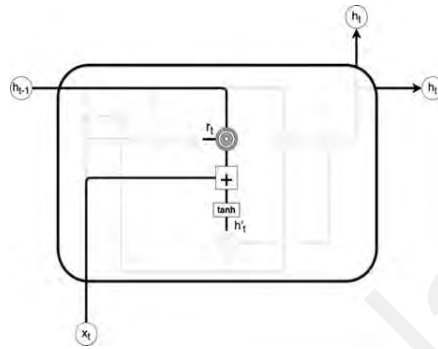


**Figure 2.18: Illustration of the reset gate impact.**

3. **Current hidden state content-** is show in the Equation 2.12 and Figure 2.19.

Basically, this is very similar to the original vanilla RNN formulation with an additional reset gate influencing the content of the previous hidden state.

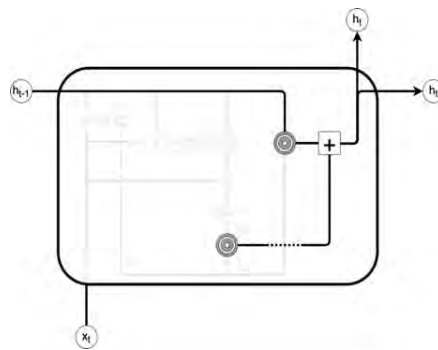
$$h'_t = \tanh(Wx_t + r_t \odot U h_{t-1}) \quad (2.12)$$



**Figure 2.19: Illustration if the current memory content.**

4. **Final memory at current time step** – shown in Figure 2.20 , is very similar to the update equation from LSTM. Basically, the Equation 2.13 merges the content of the hidden state in the previous time step and the content in the current time step to produce the final memory content.

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \quad (2.13)$$



**Figure 2.20: Illustration of the final memory at time “t”.**

### 2.4.3.3 Different Recurrent Neural Network Architectures

RNNs have shown great success in sequence modeling tasks, training RNNs for long sequences is challenging. Among the challenges presented in the literature the following are the most important ones:

1. Long-range complex dependencies
2. Vanishing and exploding gradient
3. Efficient parallelization

Multiple recurrent structures have been proposed to address the above issues. In this section, we review the most successful architectures.

**Skip RNN model** - is an extension of the original recurrent connection in an RNN model introduced by (Campos, 2018). The goal is to improve the RNN models by limiting the size of the computational graph. To do this, the authors have proposed to skip the state updates. Their experimental results have shown that skip RNNs with vanilla recurrent units can match or in some cases even outperform the RNN models with GRU or LSTM units. In addition, they were also able to decrease the computational requirements. Having skip connections help gradients flow being backpropagated through fewer time steps. This makes it easier for the gradient based optimization approaches to learn faster and more efficient when considering long sequences models. This is shown in Figure 2.21.

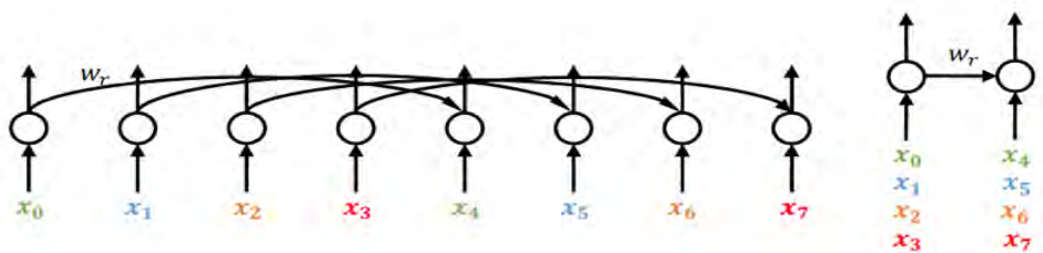
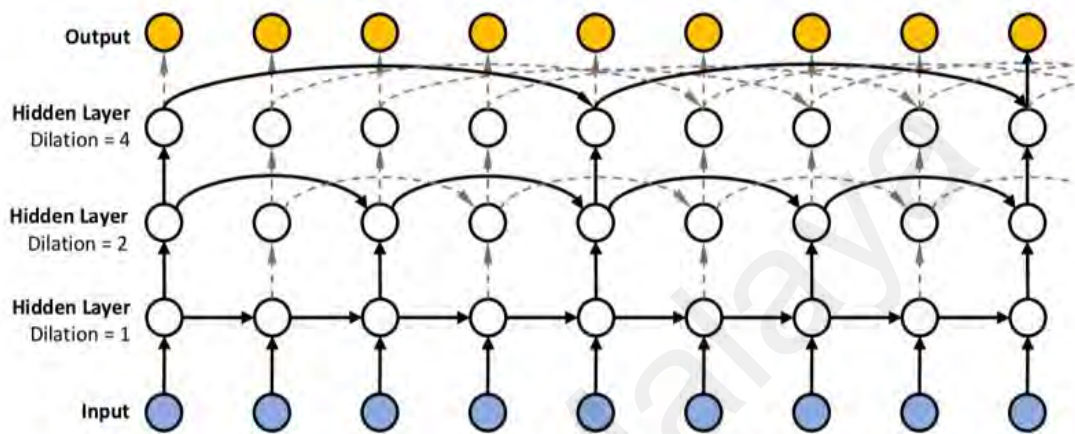


Figure 2.21 :Structure of the skipped RNN model (Chang et al., 2017)





information in a coherent unified manner. Effectively, adding these structured recurrent skip connections helps with the (vanishing /exploding) gradient problem. It also makes it easier to model long-range dependencies. One important characteristic of dilated RNN is that it is compatible with all recurrent cell types (e.g., LSTM and GRU).



**Figure 2.23: Overall architecture of the 3 layer dilated RNN with dilations of one, two, and four (Chang et al., 2017)**

More specifically, let  $c_t(l)$  denote a cell in layer  $l$  at time  $t$ . The skip connection is mathematically defined in Equation 2.14.

$$c_t(l) = f(x_t(l), c_{t-s(l)}^{(l)}) \quad (2.14)$$

, where  $x$  the input to layer  $l$  at time  $t$  and  $f$  is the RNN cell (e.g., LSTM).

## 2.5 Conclusion

To summarize, as it is shown in recent years, deep learning-based approaches have significantly improved the accuracy and efficiency compared to classical machine learning models which were mostly based on hand-crafted features in computer vision. Building on top of the success of these approaches, we intend to use the relevant and successful neural network architectures for the purpose of long sports video analysis. While there are commonalities between sports, given the large variation in scenes,

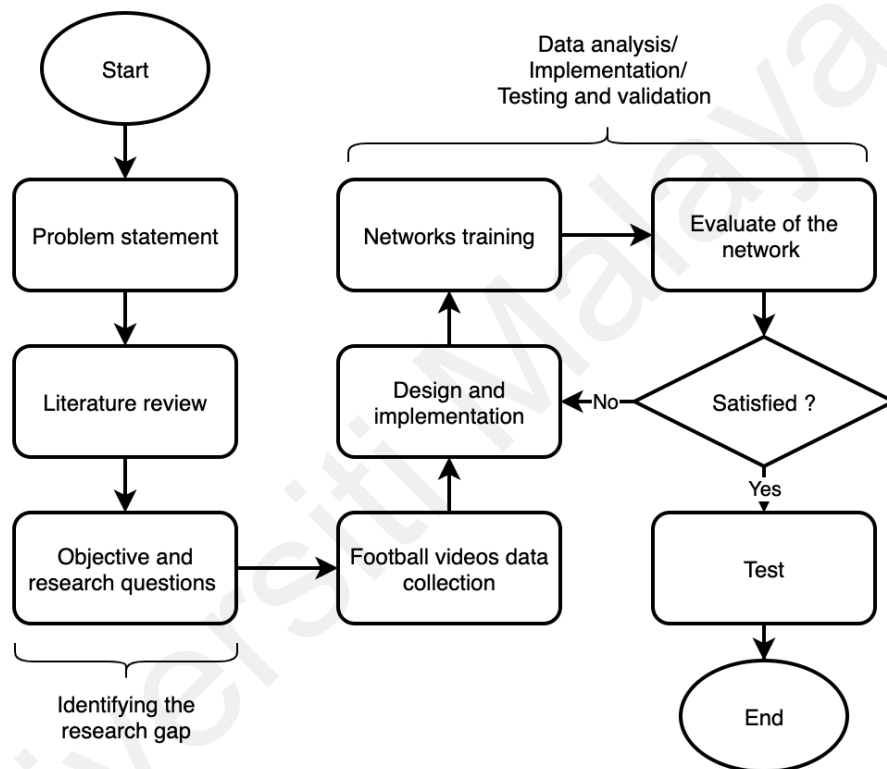
crowds, player formation and spread in the field as well as different dynamics in sports, sports analysis model designed for non-football sports (e.g., Basketball) does not necessarily address the challenges in football videos. Unlike so many other sports (e.g., American football, basketball), football is a non-episodic sport. As a result, events in a football video are highly correlated and the correlation in time could be short-range, mid-range, or long-range. It is an outdoor sport with more than 20 players in the field which makes the appearance features very important to capture. We believe using specific CNN models designed for video analysis is important to capture fine-grained local spatiotemporal features. As mentioned earlier, authors in (Giancola et al., 2018) and (Cioppa et al., 2020) have shown that using ResNet (He et al., 2016) features outperform the various 3D convolution features. Also, in addition, as discussed above authors in (K. Simonyan & A. Zisserman, 2014), demonstrated the importance of using local temporal data. To this end, we plan to use two-stream CNNs with ResNet backbones instead of Conv3D architecture. Due to limited hardware resources which limits our training capabilities, we choose to use a pre-trained two stream-CNN with ResNet-50 architecture as a backbone. Using a pre-trained model is even more important when there is a large variety of scenes and their lighting conditions as well as a large number of players in the field.

RNNs have shown to be effective when applied in problems which require temporal reasoning over frames (Jiang et al., 2016). Events in football videos are highly correlated and exhibit short-range, mid-range, and long-range dependencies. Additionally, football is a highly dynamic sport. As a result, we believe it is important to use RNN based models to capture temporal content. In this work we planned to use Dilated RNN with LSTM units to capture various correlation between frames. Note that Dilated RNN addresses the training difficulties for standard RNNs (Hochreiter & Schmidhuber, 1997).

## CHAPTER 3: METHODOLOGY

### 3.1 Introduction

In this chapter, we present the research design and the methodology used in this study to answer the research questions and to achieve the objectives of the research. We used quantitative research methods in this work. Our research methodology is implemented in the following steps presented in Figure 3.1.



**Figure 3.1: Research methodology process followed in this thesis.**

In the following sections we will provide more details for each of the above steps.

### 3.2 Approaches of the Research

We followed the standard practice for academic research. We first reviewed the relevant literature and group them based on their relevance. This enables us to refine our research questions and objectives and helps us to clarify our problem statement. We then focused on identifying the dataset requirements and data processing needs. Finally, we propose a model to support our hypothesis and studied the potential evaluation metrics

for fair quantitative comparison of our method with baselines and state of the art. In the following subsections, we provide a detailed explanation of each of the above processes.

### 3.2.1 Review of Related Literature

In the early stage of this research work, we conduct a critical review of sport analysis techniques applied to different sports. In summary, our investigation shows that:

- First, each sport has its own unique characteristics which results in certain models to be only appropriate for a specific sport. For example, while both (Ramanathan et al., 2016) and (Ibrahim et al., 2016) have addressed player tracking in basketball and volleyball, each approach relies on certain assumptions which results in certain model to be only applicable for that specific sport.
- Second, while before 2015 most of the proposed models used classical computer vision features and machine learning approaches, almost all of the recent work benefit from a rich set of deep learning-based models which mostly use CNNs for feature extraction from frames and RNNs for temporal reasoning.

Since the main problem we focused on in this work is event localization in football videos, we specifically perform a thorough review of the most recent studies that conducted research on football video analysis. The range of problems varies from automatic event annotation using text and video data to highlight detection and video summarization. Our findings show that:

- First, the lack of large-scale football datasets makes it hard to train deep learning models. In other words, since there is a limited number of datasets with very few short videos, it is hard to apply modern deep learning techniques.
- Second, football is non-episodic, there is a large variation in the appearance of the football field, the field itself is large, there are more players in the field

compared to other group sports. All of these make it harder to model complex frame relations in football videos using standard temporal models such as vanilla recurrent neural networks.

Based on our extensive literature review, we realized that deep learning and machine learning approaches are among the most successful techniques applied in sports analysis, specifically event understanding and localization. One key observation is that, to the best of our knowledge, the prior works that have focused on machine learning techniques, did not model the impact of various range of frame correlation in feature extraction. In other words, they did not specifically model short-range, mid-range and long-range dependencies in football videos.

Since machine learning and deep learning methods are general mathematical and statistical approaches for data driven learning, it is possible to benefit from findings in other application areas. To achieve this and to make sure we are also aware of studies in other application areas, we studied the application of most recent deep learning techniques which have been applied on a variety of times series problems such as video analysis and natural language processing.

We collected published studies by searching for relevant articles in the English language published between 1990 and 2019 from the Institute of Electrical and Electronics Engineers (IEEE), Elsevier, Springer, ScienceDirect, PubMed and arXiv databases. Key search terms were a combination of "artificial intelligence", "machine learning", "relevant machine learning methods", "deep learning", "relevant deep learning methods", "object detection in videos", "activity/action recognition/detection in videos", "object tracking in videos", "video summarization", "video(sport/non-sport) event detection", "video (sport/non-sport) event classification". Additional studies were conducted by searching the reference lists of the retrieved articles and manually searching

in relevant journals and conferences in the computer vision field such as CVPR and ICCV.

### 3.3 Football Dataset (SoccerNet)

As stated in (Giancola et al., 2018), the dataset is collected from online sources. The game and video times are synchronized by the game clock. A semi-automatic approach is used to generate rough temporal annotations. Based on the available game report, which is parsed and temporally aligned with the video, annotation labels for the events are created in a semi-automatic fashion.

#### 3.3.1 Video Collection

Videos are obtained and collected from the six main European Championships which are collected from 2015, 2016, and 2017 seasons. The details of the games are provided in Table 3.1. As videos are collected from various online providers with different video encodings (e.g., H264 and MPEG). Different videos might have used different containers (e.g., MJPEG and MKV). Frame rate varies from 25 up to 50 frames per second (FPS). In addition, the image resolution ranges from SD (Standard Definition) to Full HD (High Definition). Overall, the dataset contains 764 hours of football videos which is almost 4TB of data.

**Table 3.1: Details of the collected games in SoccerNet.**

League	14/15	15/16	16/17	Total
EN-EPL	6	49	40	95
ES-LaLiga	18	36	63	117
FR-Ligue1	1	3	34	38
DE-BundisLiga	8	18	27	53
ET-Serie A	11	19	76	96
EU - Champions	37	45	19	101
Total	81	160	259	500

### **3.3.2 Data Preprocessing**

In the following subsections, we provide details of the pre-processing steps applied to videos for semi-automatic annotation.

#### **3.3.2.1 Game Synchronization with OCR**

The original videos contain recordings from before and after the game. We refer to this as untrimmed videos. Authors in (Giancola et al., 2018) used an Optical Character Recognition (OCR) based approach to identify the exact time presented in the video frame (usually in the top right or the bottom of the frame). As it is argued, this is a more robust approach compared to previous methods which rely on the appearance of the center of the frame and the referee’s whistle sound proposed in (Z. Wang et al., 2016). It is important to mention that this is possible because of the fact that football is a non-episodic game which ends when the time is up. Since using OCR from a single game could be noisy, multiple randomly sampled frames are used to identify the region of interest for game information and the RANSAC algorithm is used to remove the outliers. More details and a complete explanation are provided in (Fischler & Bolles, 1981).

#### **3.3.2.2 Collecting Event annotations**

Event annotations are collected from the available game reports provided for free at league’s websites. These reports provide a summary of the main events that happened during a game within a one-minute window. A total of 171,778 annotations of three main event categories, (i.e., “goals”, “cards” and “substitutions”), are collected from 13,489 games. While the games are from the Champions League of five main European leagues from 2010 to 2017, due to storage limits, only 500 matches with a total number of 6,637 events are practically used. To summarize, the dataset contains a total number of 6,637 temporal annotations which are automatically parsed from online match reports at a one-

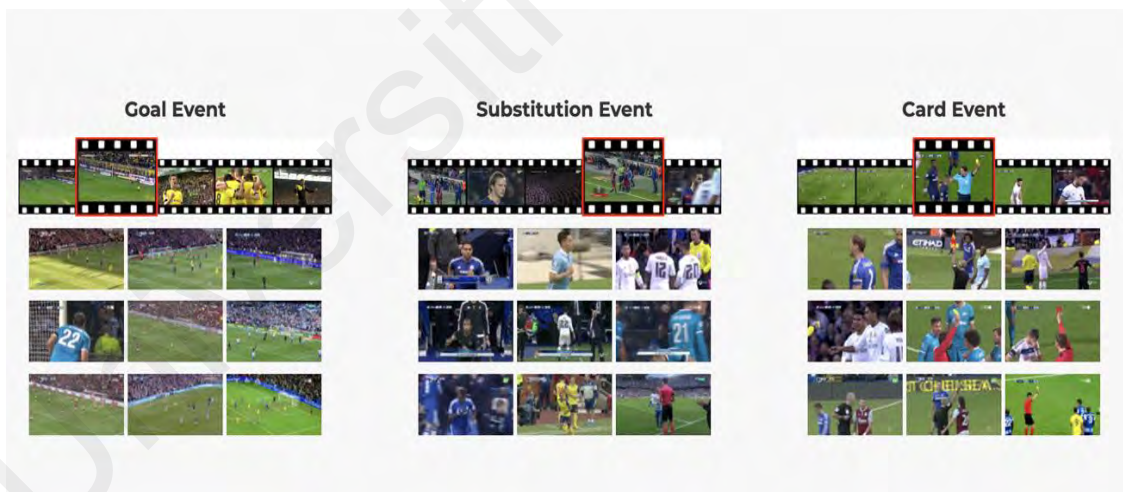


minute resolution window. Three main classes of events (Goal, Yellow/Red Card, and Substitution) are annotated. APPENDIX B presents a sample of game events annotation.

A manual annotation process is then used to convert the “one-minute” resolution annotations to the “second-level” resolution annotations. To define second-level granularity, the events are defined as the following:

1. A card event is an instant that a referee shows a yellow or a red card to a player.
2. A goal event is an instant that the ball crosses the goal line.
3. A substitution event is an instant that a new player enters the football field (Note that substitutions that occur during half time break are not included).

These event definitions are demonstrated in (Figure 3.2) and are used to identify temporal anchors for each event during the annotation process.



**Figure 3.2: Three football events annotated in SoccerNet.**

### 3.3.2.3 Splitting Dataset for Training, Testing and Validation

Table 3.2 shows the training, test, and validation splits for the collected dataset. While the number of videos for each event class are not exactly the same, the authors in (Giancola et al., 2018) have tried to make it reasonably distributed. The videos are split to one-minute annotated chunks where one of the three events occur in this one-minute

window. To summarize, the games are randomly split into 300, 100, and 100 games for training, validation, and testing which contains 3965 event instances for training, 1314 for testing, and 1358 for validation.

**Table 3.2: Details of the dataset splits: training, testing and validation splits.**

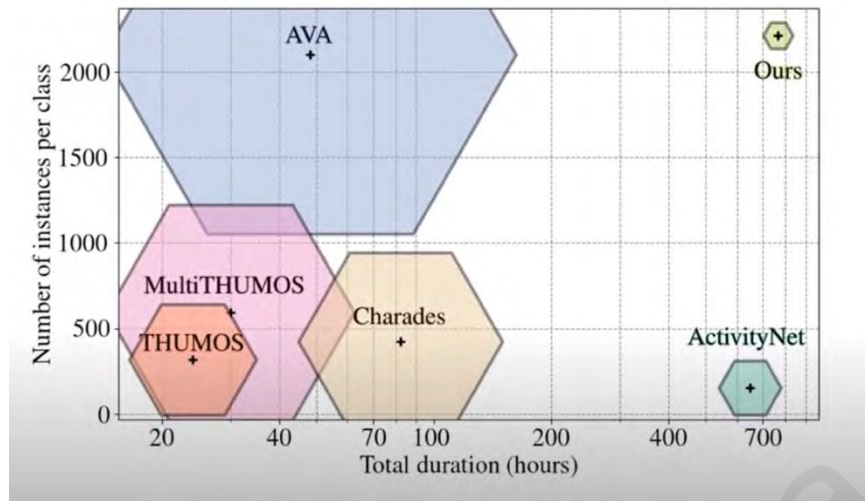
Split	Event			Total
	Goals	Cards	Substitution	
<b>Train</b>	961	1296	1708	3965
<b>Valid</b>	365	396	562	1314
<b>Test</b>	326	453	579	1358
<b>Total</b>	1643	2145	2849	6637

### 3.3.3 Dataset Comparison

Considering the total duration and number of instances per class, the SoccerNet dataset is the largest localization dataset available. Table 3.3 and Figure 3.3 compares various relevant action localization datasets in terms of the number of instances per class, and the total duration. In Figure 3.3 the size of the hexagon shows the density of the event within the video.

**Table 3.3: Comparison of SoccerNet dataset with available video datasets.**

Dataset	Content	Video	Instance	Duration (hrs.)	Sparsity (event/h)	Classes	Instance per classes
THUNUS'14	General	413	6363	24	260.4	20	318
MultiTHUMOS	General	400	38690	30	1289.7	65	595
Activitynet	General	19994	30791	648	47.5	200	154
Charades	General	9848	66500	82	811.0	157	424
AVA	Movies	57600	210000	48	4375.0	100	2100
<b>Ours (Succernet)</b>	<b>Football</b>	<b>1000</b>	<b>6637</b>	<b>764</b>	<b>8.7</b>	<b>3</b>	<b>2212</b>

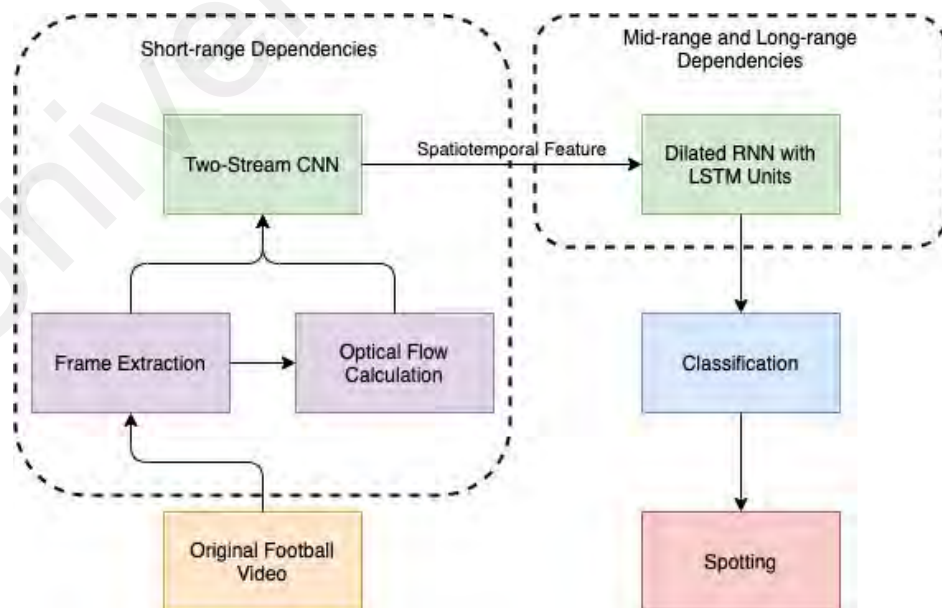


**Figure 3.3 :Visualization of comparison between SoccerNet and available video datasets (Giancola et al., 2018)**

### 3.4 Proposed Model

We hypothesize that the main difficulty in event localization is directly associated with various correlations among video frames at different granularity levels. Our key idea is to address these complex dependencies in football videos, by modeling short-range, mid-range, and long-range correlations between video frames. As mentioned in Chapter 2, certain network architectures are more suitable for sequential data such as videos. Different network architectures can be used to understand and model various correlations among frames. While short-range correlation should consider both spatial features in every frame and local temporal features between a limited number of frames, mid-range and long-range features should consider longer correlations between frames. As we mentioned in Chapter 2, different CNN architectures have been proposed to benefit from spatial and local temporal features and capture the short-range correlations. In the mostly used CNN architecture for videos (K. Simonyan & A. Zisserman, 2014), which is cited more than 5263 time, the author used two-stream CNN to capture spatial and local temporal features. For the mid-range and long-range correlation, as we discussed in Chapter 2 most of the recent works benefit from different RNN architectures (Campos, 2018; Chang et al., 2017; Chung, Ahn, & Bengio, 2016; Koutnik, Greff, Gomez, &

Schmidhuber, 2014; Neil, Pfeiffer, & Liu, 2016; Y. Zhang et al., 2016). Dilated RNN is one of the latest architectures which implemented for longer sequential dependencies and resolve some training difficulties in standard RNNs (Chang et al., 2017). In addition, as we mentioned in Chapter 2 author in (Jiang et al., 2016) showed that using a combination of CNN and RNN can improve the accuracy of football events classification. But as discussed earlier training RNNs is challenging for long sequences which results in poor performance in long videos. As a result, In this work, to achieve our hypothesis, we propose a network with two main components: A two-stream convolutional neural network (K. Simonyan & A. Zisserman, 2014) for short-range spatiotemporal feature extraction, and a dilated RNN with LSTM units (Chang et al., 2017) to model mid-range and long-range correlations. This is illustrated in Figure 3.4. While the two-stream neural network provides the local spatiotemporal description of video frames, the hierarchical nature of the dilated recurrent network combined with the structured skip connections and sophisticated design of LSTM units enables our proposed model to link the local and global information in a coherent unified manner.



**Figure 3.4: Overview: Our goal is to use low-level spatiotemporal features and a hierarchical recurrent model with skip connections (DilatedRNN with LSTM units) to improve event classification and event spotting in long football videos.**

We can formally define the problem of classifying events in football videos as follow. Given a football video as a set of frames  $\{v_i\}$ ,  $\{x_i\}$  is an ordered set of extracted features from video frames. While in classical computer vision, handcrafted features such as SIFT, HOG, or color histogram were used to define  $x_i$ , given the recent success of convolutional neural networks (CNN), it is common to define  $x_i$  as one of the hidden layers of a CNN architecture. We define a temporal segment  $t_j$  to be a subset of frames with the same assigned event. Considering an input video, the event detection problem is formally defined as estimating temporal segments,  $\{t_j\}$ , of the pre-known event categories  $c \in C$ . In activity detection research, the goal is to propose a model that correctly estimates the entire temporal segments associated with a video and avoids estimating false/incorrect segments. While this definition of an event segment is useful for certain activities (e.g., sliding from the slide), for high-speed and fast pace activities such as scoring a goal this definition is vague and it is not suitable. In other words, for events that occur spontaneously, it is hard to clearly define the temporal segment. As a result, the event localization in this research refers to “spotting” of an event. Authors in (Giancola et al., 2018) has defined spotting as: ‘finding the anchor time (or spot) that identifies an event. Intuitively, the closer the candidate spot is from the target, the better the spotting would be, and its quality is measured by its distance from the target’.

Given a set of videos of football games, which we call them clips, we aim to learn a model which is capable of identifying certain predefined events. More specifically, the goal is to identify if a certain clip is of a certain event class (i.e., Goal event, Substitution event, Card event). One interesting application of such a system would be to generate highlights from a long football video. To formalize the problem statement, we introduce few notations in this section. Assume that we have access to a set of video clips from a set of football games. We call this set  $D = \{(v_i, y_i): 1 \leq i \leq n\}$ , where  $v_i$  is the video clips,  $y_i$  is the actual event class of  $v_i$ , and  $n$  is the number of given video clips. Note that

each video clip  $v_i$  itself if described as  $v_i = \{x_j: 1 \leq j \leq m\}$ , where  $m$  is the length of the video clip. Given a video  $v_i$ , the goal is to use the learned classification model to classify the video to one of the following three event classes: 1) *Goal event*, 2) *Substitution event*, and 3) *Card event*.

This is shown in following equations:

$$o_i = \text{Opticalflow}(v_i, v_{i+1}) \quad (3.1)$$

$$x = \text{TwoStreamCNN}(v_i, o_i) \quad (3.2)$$

$$y = \text{DilatedRNN}(x) \quad (3.3)$$

Where  $o_i$  is the amount of displacement of each pixel from a frame “ $v_{i+1}$ ” at time “ $t + 1$ ” relative to a frame “ $v_i$ ” at time “ $t$ ”,  $x$  is set of extracted features by  $\text{TwoStreamCNN}(v_i, o_i)$  from video frame “ $v_i$ ” and opticalflow “ $o_i$ ” at time “ $t$ ”. Finally, the class score  $y$  is calculated by  $\text{DilatedRNN}(x)$ . A detailed explanation of each network’s equation is discussed in Chapter 2.

As mentioned in Chapter 1, classifying the events in a football video is challenging due to their enormous variability in appearance and motion features. In particular, the videos are shot by people of different skills under varying weather and lighting conditions. The videos are usually recorded from multiple different viewpoints. In addition, each football field has its own specific compositions and markings. The appearance of the football scenes varies a lot from team to team. Additionally, as mentioned in Chapter 1 and 2, football is an outdoor sport that adds additional complexity for different lighting conditions. The football field is large, the sport itself is fast-paced and has very high dynamics. The above phenomena result in fast motions in football videos which directly contributes to motion complexity. That is why we believe that our

two-stream CNN is a sophisticated local spatiotemporal feature extractor. We train our spatial stream network on ImageNet, which allows a reasonable generalization of the appearance features. Our spatial stream is pre-trained on the UCF-101 activity dataset, which includes multiple outdoor sports. This allows our temporal network to learn from the dynamics in sports videos.

We will provide a detailed explanation of each component in Chapter 4. In sections 4.3 and 4.4.1 we explain the details of our two-stream CNN component and in sections 4.3 and 4.4.2 we present our new network architecture based on dilated recurrent neural network and LSTMs.

### **3.5 Evaluation of Proposed Model**

To achieve our last objective, the classification and spotting quality of our proposed model was assessed and compared with baselines and reported results in (Giancola et al., 2018). Various evaluation metric has been used by the research community to report the evaluation results for detection tasks. The most common metric used in computer vision for object/event detection is the mean average precision (mAP) (Giancola et al., 2018). We use mAP to compare both our classification and spotting results as proposed by (Giancola et al., 2018). While for the classification task, mAP is well-defined, it is inherently difficult and ambiguous to define it for Spotting. As a result, in (Giancola et al., 2018), a tolerance threshold is introduced to define the correct and incorrect spotted event within that threshold. This means that rather than identifying the boundaries of an action within a video and looking for the intersection over union (IoU) between temporal windows (Figure 3.5), spotting identifies the moment that an event occurs. A candidate spot is positive if it lands within a tolerance window around the anchor of an event. Otherwise, it is considered as negative. This is illustrated in Figure 3.6.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Figure 3.5: Visual representation of the intersection over union (IoU)

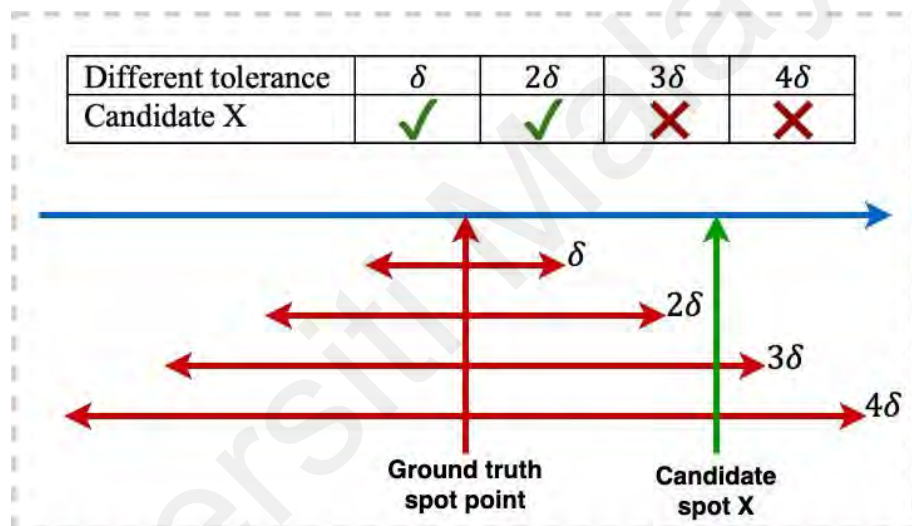
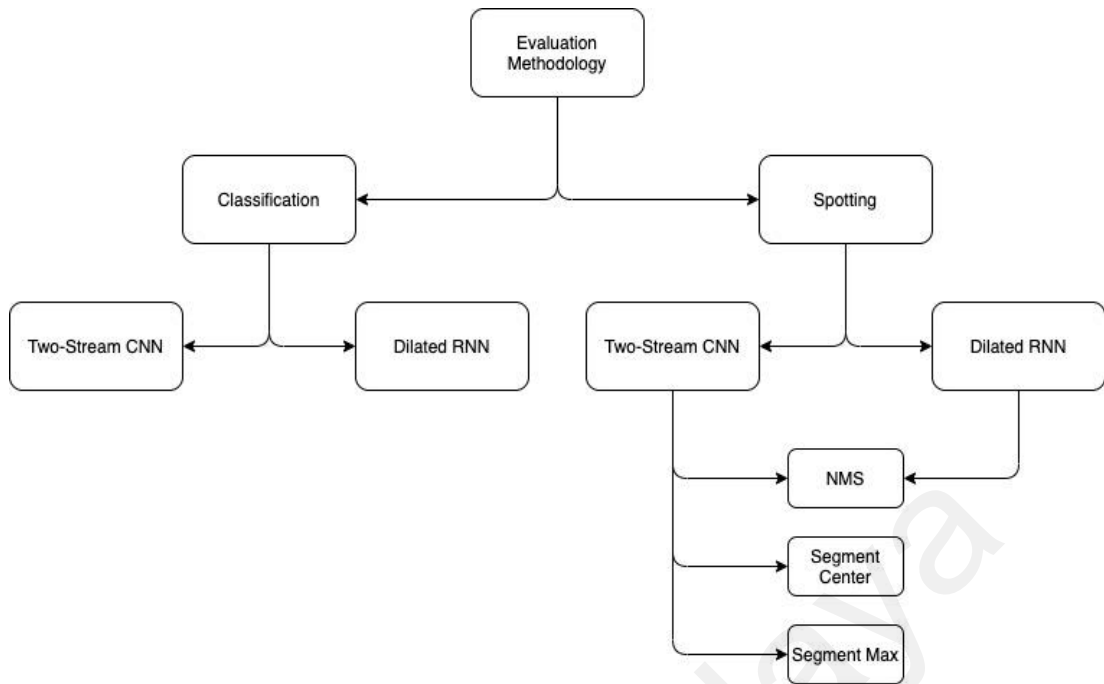


Figure 3.6: Overall idea of spotting (Candidate X spot the event within a tolerance of  $3\delta$  and  $4\delta$ )

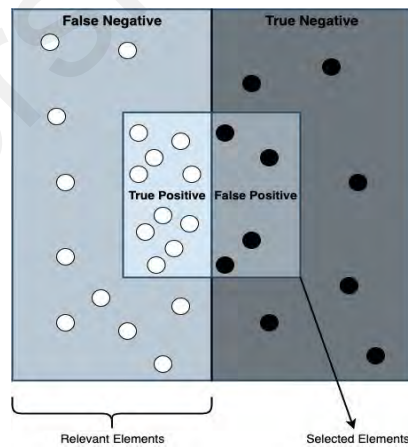
This makes spotting a detection problem which allows us to define False positive, False negative, True positive, and True negative instances for this classification problem. The most common evaluation metric for such a detection problem is mAP. We used mAP for both event classification and spotting problems. Figure 3.7 summarizes the evaluation methodology carried on in this work.





**Figure 3.7: High-level overview of various evaluations of event classification and spotting**

To better understand the mAP, we provide two illustration as in Figure 3.8 and Table 3.4.



**Figure 3.8: Illustration of true positive, false positive, false negative and true negative instances.**

**Table 3.4: Definition of the confusion matrix**

True class	Predicted class	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Given the classifier's outputs during test time, one can define four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). FP is the number of predictions that are classified incorrectly as the target output, TP is the number of predictions which are classified correctly as the target output. On the other hand, TN is the number of predictions which are classified correctly as non-target output. Finally, FN is the number of predictions which are incorrectly classified as non-target. During training, the goal is to approach a zero FP and FN.

While TP, FP, TN, and FN are useful and essential, they usually do not provide a high-level overview of the quality of a classifier. As a result, it is common to compute other metrics based on the above four.

It is important to know how many times the estimated output (i.e., prediction) is correct. This is formulated as the percentage of correct estimates or  $\frac{TP}{(TP+FP)}$ . In addition, it is also important to know, how many of the ground truth instances have been correctly identified. This is formulated as the percentage of the correct estimates or  $\frac{TP}{(TP+FN)}$ . Based on the above definition, in practice, there is usually an inverse relation between precision and recall. If precision goes up, recall usually comes down and vice versa. As a result, researchers usually provide a plot referred to as precision/recall curve, which basically provides both precision and recall for certain choices of a trade-off parameter and report the "Area Under Curve" (AUC) (Zhu, 2004). This is shown in Equation 3.4, where  $R$  is a set of recall values, and for a given recall value of  $r$ ,  $p(r)$  represents the corresponding precision values

$$AP = \int_0^1 p(r)dr = \frac{1}{R} \sum_{r_i} p(r_i) \quad (3.4)$$

Another metric which has been used more often recently, is the mAP. Basically, for one output class, average precision (AP) is defined as the average precision for different recall values, and it can be seen as the AUC of recall/precision plot for that class. Since precision is always in the range of (0,1), AP is also in the range of (0,1). The mAP metric basically computes the mean AP among different output classes. This is shown in Equation 3.5.

$$mAP = \frac{\sum_q^Q AveP(q)}{Q} \quad (3.5)$$

### 3.6 Summary

This chapter provided a detailed explanation of the methodology process of this research. We provided more information regarding the details of the football video dataset which includes the video acquisition process, data processing and comparison with other datasets. We also provided the high-level components of our proposed model for event classification and spotting. Our proposed model consists of two state of the art neural network architectures. While one of the proposed subnetworks captures the short-range spatiotemporal features, the other modules model mid-range, and long-range dependencies between frames in football videos. In Section 3.5, we explained the evaluation metric used to assess the accuracy of both classification and spotting problems. This evaluation is used to report the accuracy of classification and spotting models.

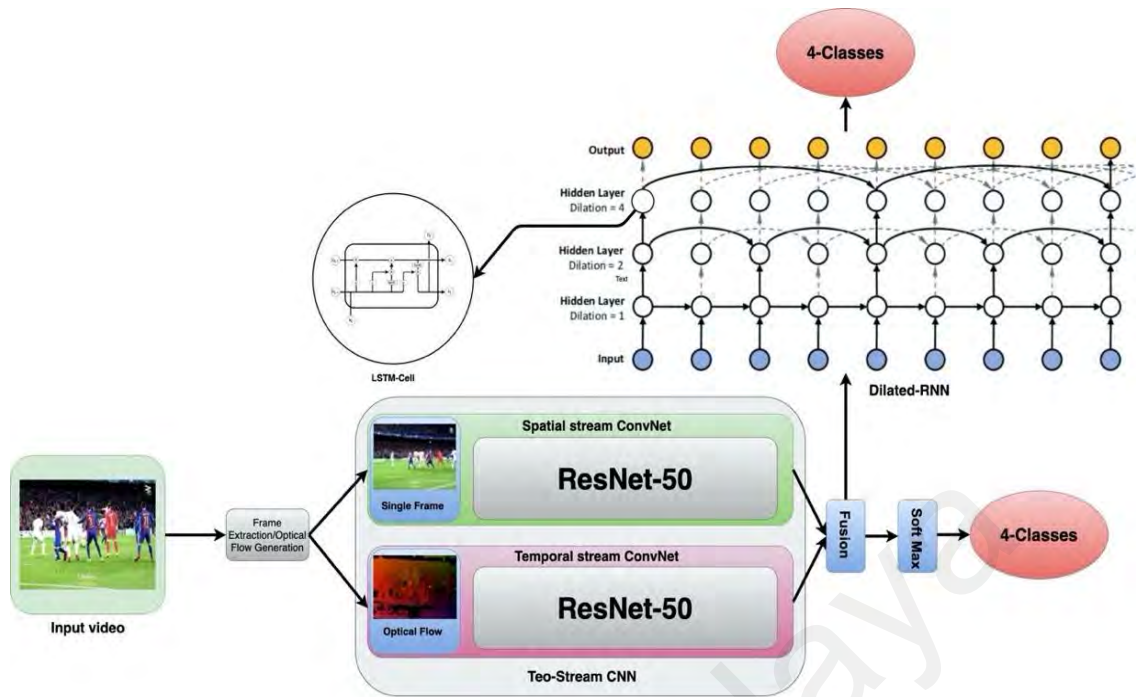
## CHAPTER 4: EVENT SPOTTING AND CLASSIFICATION

### 4.1 Introduction

Due to the rapid advancements in machine learning techniques over the past few years, there has been a momentous increase of interest in video analysis, event detection in videos, and human recognition in videos. In this work, our aim is to develop a machine learning algorithm for event spotting and classification in long football videos. As mentioned in the literature review, presented in Chapter 2, multiple prior research work have addressed the event localization problem. However, the literature acknowledged that several improvements are required to improve the accuracy of event classification and spotting, specifically in long videos with a high dynamic content such as sports videos. In this chapter, we present the proposed model and a complete description of the implementation steps which were applied during the implementation and network training.

### 4.2 Proposed Method

As stated in Section 3.4, given a video, our model uses two-stream CNN to generate local spatiotemporal features from two consecutive frames. These local features capture the short-range correlation between frames in a short temporal window. Building on top of these local spatiotemporal features, we use a dilated recurrent neural network with LSTM cells to capture mid-range and long-range correlations between video frames in long videos. This is illustrated in Figure 4.1. Note that due to limited hardware resources our two-stream CNN implementation uses pre-trained ResNet-50 networks for each stream. Table 4.1 provides details of the convolutional layers, normalization layers, and number of feature maps in each layer for the ResNet-50 model used in the above two-stream CNN.



**Figure 4.1: Detailed illustration of the proposed model. Given the input video frames, we first compute a dense Opticalflow. Then the spatial stream network consumes the first frame, and the temporal stream network consumes the Opticalflow. The results from the two-stream networks are fused to form a single feature vector for future classification components (SoftMax layer or Dilated RNN).**

**Table 4.1 : Architecture of ResNet-50 used in this study (unit are by pixel)**

Layer name		Filter size	Stride	Padding	Number of filters	Output feature map size		
Input layer						224 × 224 × 3		
Conv1	Conv.	7 × 7 × 3	2	3	64	112 × 112 × 64		
	Max Pooling	3 × 3	2	1		56 × 56 × 64		
Conv2	Res2a	Conv.	1 × 1 × 64	1	0	64	56 × 56 × 256	
		Conv.	3 × 3 × 64	1	1	64		
		Conv.	1 × 1 × 64	1	0	256		
		Conv.	1 × 1 × 64	1	0	256		
	Res2b-c	(Shortcut)	Conv.	1 × 1 × 256	1	0	64	56 × 56 × 256
			Conv.	3 × 3 × 64	1	1	64	
			Conv.	1 × 1 × 64	1	0	256	

**Table 4.1 Continued.**

Layer name			Filter size	Stride	Padding	Number of filters	Output feature map size
Conv3	Res3a	Conv.	$1 \times 1 \times 256$	2	0	128	$28 \times 28 \times 512$
		Conv.	$3 \times 3 \times 128$	1	1	128	
		Conv.	$1 \times 1 \times 128$	1	0	512	
		Conv. (Shortcut)	$1 \times 1 \times 256$	2	0	512	
	Res3b-d	Conv.	$1 \times 1 \times 512$	1	0	128	$28 \times 28 \times 512$
		Conv.	$3 \times 3 \times 128$	1	1	128	
		Conv.	$1 \times 1 \times 128$	1	0	512	
Conv4	Res4a	Conv.	$1 \times 1 \times 512$	2	0	256	$14 \times 14 \times 1024$
		Conv.	$3 \times 3 \times 256$	1	1	256	
		Conv.	$1 \times 1 \times 256$	1	0	1024	
		Conv. (Shortcut)	$1 \times 1 \times 512$	2	0	1024	
	Res4b-f	Conv.	$1 \times 1 \times 1024$	1	0	256	$14 \times 14 \times 1024$
		Conv.	$3 \times 3 \times 256$	1	1	256	
		Conv.	$1 \times 1 \times 256$	1	0	1024	
Conv5	Res5a	Conv.	$1 \times 1 \times 1024$	2	0	512	$7 \times 7 \times 2048$
		Conv.	$3 \times 3 \times 512$	1	1	512	
		Conv.	$1 \times 1 \times 512$	1	0	2048	
		Conv. (Shortcut)	$1 \times 1 \times 1024$	2	0	2048	
	Res5b-c	Conv.	$1 \times 1 \times 2048$	1	0	512	$7 \times 7 \times 2048$
		Conv.	$1 \times 1 \times 512$	1	1	512	
		Conv.	$1 \times 1 \times 512$	1	0	2048	
Average Pooling			$4 \times 8$	11	0		$1 \times 1 \times 2048$
Fully Connected Layers	FC						101
Soft Max							

### **4.3 Feature Description**

In this section, we provide more details on our feature extraction approach. While we are using deep neural networks, where the intermediate layers (hidden layers) of the neural networks are responsible for automatic feature extraction from the original raw input, it is shown that that different parts of the networks are responsible for different types of features abstractions. As a result, in the following subsections, we explain the reasoning behind using different neural network architectures and what type of features are computed and modeled using each of them.

#### **4.3.1 Short-range**

One can breakdown the data in a video to spatial and temporal elements. While the spatial element provides information about the appearance of the scene and the objects in the scene, the temporal element provides information on how the appearance changes with time. As a result, (K. Simonyan & A. Zisserman, 2014) investigated the separation of these two information streams. We believe that modeling the temporal stream using Opticalflow provides a rich set of local spatiotemporal features. As a result, building on top of the original model in (K. Simonyan & A. Zisserman, 2014), in this work we split the CNN architecture into a spatial stream for object detection and a temporal stream for the motion detection for the task of event localization. Each network is implemented using CNN architecture and combined by late fusion in the last layer. Note that, while the two-stream CNN network generates sophisticated local spatiotemporal features by modeling the short-term correlation between frames, it fails to capture longer temporal correlations between frames. In the next two subsections, we provide a detailed explanation of the spatial and temporal neural networks used in this work.

#### 4.3.1.1 Spatial Stream

Certain events and actions are highly correlated with objects and people in the scene. This kind of correlation can be captured by appearance features from a single frame. The core idea behind the spatial stream is to use convolutional layers to extract appearance features from a single static frame. One important benefit of using a spatial stream is that we can leverage from the sophisticated network architectures designed for image analysis problems and pre-train the models on large image datasets. Note that, for better accuracy and practical reasons, we have used a pre-trained model trained on ImageNet and only fine-tuned the last layer.

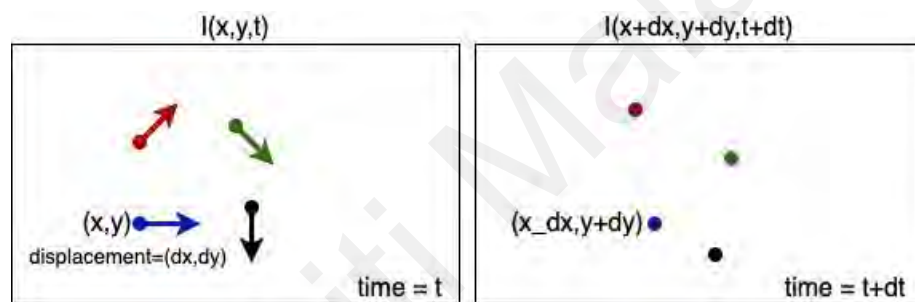
#### 4.3.1.2 Temporal Stream

While the spatial stream provides a rich set of appearance features from a single frame, for particular events and actions, additional temporal features are needed for classification and recognition. The core responsibility of the temporal stream is to obtain such local temporal information from adjacent video frames. Rather than using raw frames, it has been shown that using low-level opticalflow features are more informative. Training on video datasets is harder and more expensive. As a result, we use a pre-trained temporal stream network, with a ResNet-50 backbone, which is trained on UCF101(Soomro, Zamir, & Shah, 2012) in our implementation. Note that, unlike the spatial stream ConvNet, which can be pre-trained on a large still image classification dataset (such as ImageNet), the temporal ConvNet needs to be trained on video data – and the available datasets for video action classification are still rather small. The details of the opticalflow computation is presented in the following paragraph.

**OpticalFlow Computation-** As demonstrated in Figure 4.2, opticalflow simply computes the amount of displacement of each pixel from a frame at time “ $t + 1$ ” relative to a frame at time “ $t$ ”. In other words, opticalflow defines a set of vector fields based on



changes in the pixel values in two consecutive frames, “ $t$ ” and “ $t + 1$ ” The defined vector field has two components: a) A horizontal component, and b) a vertical component. The horizontal component,  $d_{xt}$ , represents the amount of displacement in the  $x$  direction. The vertical component,  $d_{yt}$ , represents the amount of displacement in the  $y$  direction. These two vector fields can be represented as two input channels with exactly the same dimensions as the original raw data. The input to the temporal stream is a stacked opticalflow channels of pairs of video frames. In our implementation, we only use two frames. This is mainly due to small performance gain using the additional channels and also practical computational efficiency and limitations in the computational resources.



**Figure 4.2: Illustration of the Opticalflow computation. Opticalflow calculates the displacement of a location in two frames in the horizontal and vertical directions.**

#### 4.3.2 Mid-range

While local spatiotemporal features have shown significant improvement in event classification and localization, it has been shown in the research community that more complex events and actions would benefit from analyzing longer correlations between video frames. RNNs have shown a great capacity in summarizing and categorizing the correlation between video frames which are further apart.

As most of the available datasets only consider relatively short videos (with length up to a few minutes at most), in practice, most of the models only consider mid-range correlation between frames rather than long-range correlation. One of the most successful RNN architectures uses LSTM cells and is usually referred to LSTM networks. As a

result, we also use LSTMs to capture the mid-range correlation between video frames. The core components of LSTM are shown in Figure 2.11.

### 4.3.3 Long-range

For long videos with complex event patterns, it is important to better model the correlation between frames that are far away from each other. While LSTMs is capable of capturing such a dependency for mid-range correlations, they fail to do that for long videos with longer correlations. This is mainly due to the fact that while in theory the information flow is not constrained by time, in practice the vanilla recurrent network architectures fail to learn these longer dependencies due to various learning challenges discussed before. One successful architecture which enables the model to access the information from earlier frames is the “Dilated Recurrent Neural Network”. The core components of the architecture are shown in Figure 2.23. Two important features in dilated RNN helps with modeling longer dependencies: 1) Hierarchy of recurrent units, and 2) Skip connections between recurrent units. In other words, Dilated-RNN is a multi-layer architecture designed by multi-resolution dilated recurrent skip connections. By using dilated connections, different layers of the network can focus on different temporal resolutions. In addition, and more importantly, the dilation reduces the average path's length between multiple nodes at different timestamps. This improves the ability of normal RNNs to capture long-term dependencies while prevents vanishing and exploding gradients. We want to emphasize that we improve the original dilated RNN by adding LSTM units to provide an additional memory capacity to the dilated RNN which helps it with better mid-range memory.

## 4.4 Classification

In this section, we provide the details of training our classification models. As we have multiple components, we provide the details of each training separately. We first

split the dataset into training, validation, and testing. The same training set is used to train the RNN and fine-tune the Two-stream CNN nets. We used the validation set for hyperparameter tuning and also to identify the overfitting. We choose the model before the overfitting occurs while training and fine-tuning our networks.

First, we fine-tune the two-stream CNN component of our model. This is important because the training of the RNN depends on the extracted features from the fine-tuned two-stream CNN. Later, we trained our dilated RNN with LSTM cells on top of the features computed using the fine-tuned two-stream CNN. This has two benefits: 1) It is computationally more efficient, 2) It allows us to extract the features prior to training which makes it easier and faster to train the dilated RNN network.

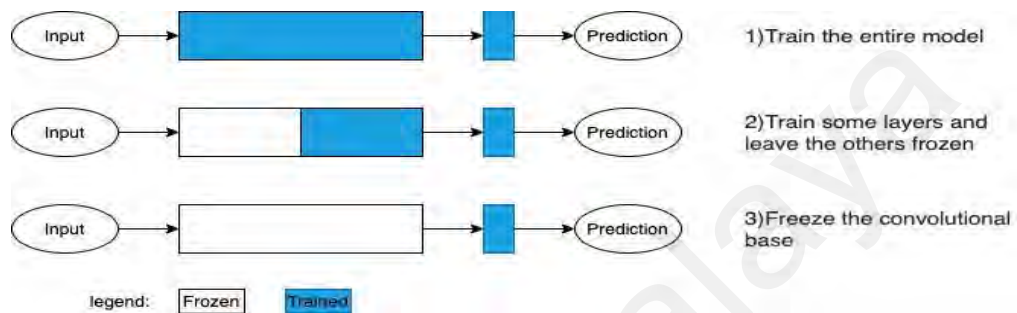
#### **4.4.1 Two-stream CNN Training**

While one can always train the entire model from scratch, using a specific dataset, it is a common approach to use pre-trained models to help with better results and faster training. In addition, using pre-trained models also helps with limited computational resources. In this work, we used a pre-trained two-stream CNN trained on large benchmark datasets of ImageNet, and UCF101. The common practice in fine-tuning the neural network models is to remove the original classification layer and to replace it with a new classification layer that is suitable for the problem of interest.

Various training strategies have been proposed for proper fine-tuning in the literature. The common approach is to freeze lower layers and train higher layers. In this approach, we need to decide which layers will be frozen and which layers will be (re) trained. In the case of smaller datasets, it is preferred to freeze more layers to reduce the network complexity during training and to avoid overfitting. The common understanding among the researchers is that the lower layers of a neural network learn to extract generic features

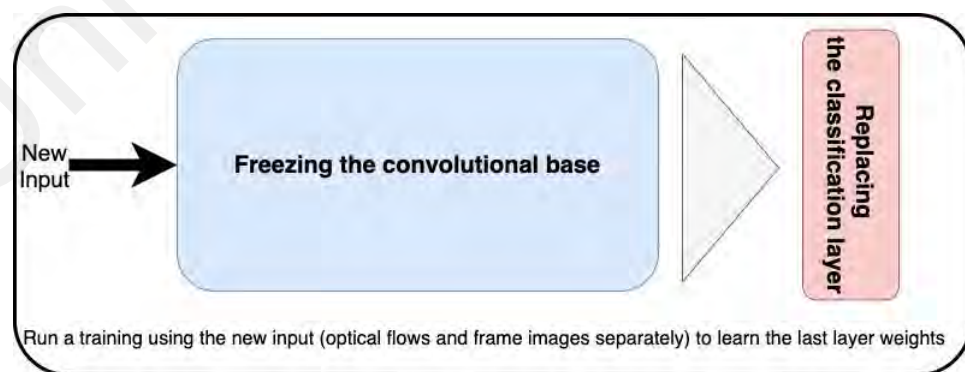
and usually, they are independent of the problem. As a result, in these approaches it is recommended to freeze almost all layers and fine-tune the last few layers in the hierarchy.

Figure 4.3 demonstrates various strategies used in fine-tuning a neural network. Note that one extreme scenario is to use the pre-trained model as a feature extraction component and do not train any new layers.



**Figure 4.3 : Various fine-tuning strategies in deep neural networks.**

In this research, due to the resource limitations we will face if train from scratch, we keep all but the last convolutional layers of the pre-trained models and only replace the last SoftMax classification layer with a four-dimensional classification layer. This is shown in Figure 4.4. The following provides the details of fine-tuning the spatial and temporal streams respectively.



**Figure 4.4: The fine-tuning approach used in this work.**

**Fine-tuning the spatial stream network-** Since the spatial network operates on single static frames, we fine-tune the classification layer using individual frames sampled from

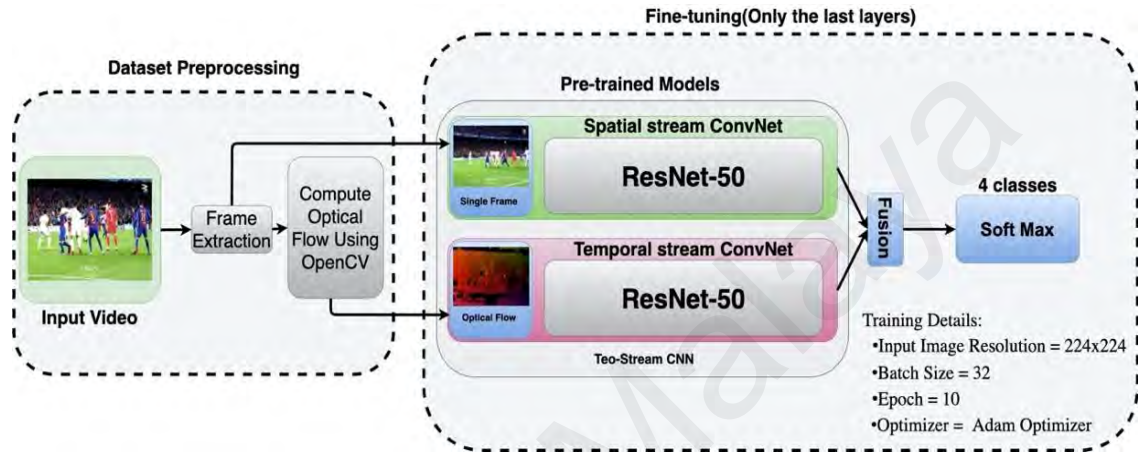
the videos. Using these frames as input, fine-tuning the spatial network is basically an image classification task. Our pretrained model is based on the ResNet-50 implementation and is trained on ImageNet.

During fine-tuning, the frames are randomly sampled from a “one-minute” video chunks provided as part of the training set. We deliberately sampled randomly to make sure the fine-tuning does not have a bias towards any specific game or event class and to uniformly cover the entire training dataset.

**Fine-tuning the temporal stream network-**Unlike the spatial stream CNN, the input of the temporal stream network is the computed opticalflow channels from two frames rather than a static image. As a result, the temporal-stream CNN needs to be trained on a video dataset for video event classification rather than image analysis. Training a temporal stream network from scratch requires a large video dataset and a powerful computation platform. To overcome this, one approach is to use a pre-trained image model and only fine-tune the very first layer and the last classification layer. In this work, we used a pre-trained ResNet-50 model which is fine-tuned on the UCF101 action recognition and activity detection dataset. Similar to the spatial stream network, we added a four-dimensional classification layer and fine-tuned these layers during training.

For fine-tuning, we similarly used the one-minute annotated video chunks. While it is possible to use all frames in a video, it is a common practice to subsample frames in the temporal dimension. We subsampled 5 frames per second (a total of 300 frames per minute). Using the previous frames of each of these frames, we compute the dense opticalflow. In other words, 10 frames were processed to compute the opticalflow channels per second. The output of the opticalflow computation is then used to fine-tune the classification layer of the temporal stream. Note that for faster training, we pre-compute Opticalflow for the sub-sampled frames. In other words, for fine-tuning, we first

computed the Opticalflow values for the training videos. These Opticalflow values are then stored as NumPy arrays to be used during the fine-tuning process. For fine-tuning, we used images of size (224 X 224), and batches of size 32. We only fine-tune for ten epochs and used Adam optimizer with an initial learning rate set to 1.0e-4. This is shown in Figure 4.5



**Figure 4.5 : The training (fine-tuning) process of Two-stream CNN**

For fusing the results from the spatial and temporal streams, similar to the original two-stream CNN(K. Simonyan & A. Zisserman, 2014) paper, we use Max-Pooling. Another reason to choose the Max-Pooling operation is that we believe average pooling might result in less confident scores due to ambiguity in prediction.

#### 4.4.2 Dilated RNN Training

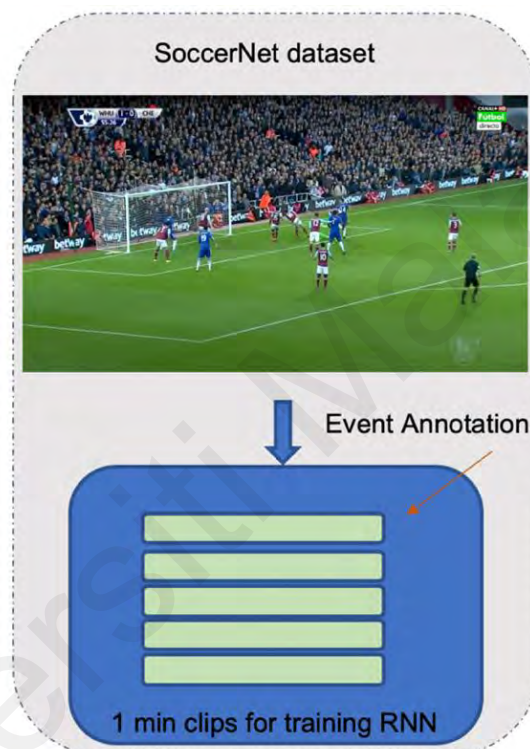
As mentioned in Section 4.2, the input to the dilated RNN is the output of the fusion layer from two-stream CNN. More specifically, we use the 101-dimensional output of the two-stream CNN which we fine-tuned in Section 4.4.1, at the input layer of the first layer of the dilated RNN. Our proposed dilated RNN has three layers where each layer is 128 dimensional. In other words, we have a dilated RNN with [128, 128, 128] dimensional hidden layers.

We use the pre-computed frame features using the underlying two-stream CNN model. For the two-stream CNN, we first identify the frames based on the temporal subsampling, followed by the opticalflow computation. In this work, we temporally subsampled five frames per second and apply the trained two-stream CNN on the sub-sampled frames. When training the dilated RNN, we load the computed features from the already computed and saved feature arrays.

It is shown that adding the dropout layer for the input layer in RNNs improves the generalization quality of the network (Gal & Ghahramani, 2016). In this work, we use dropout layers with the dropout-rate set to 0.25. The main reason we use the dropout layer only in the first layer is that for the second and third layers, the network uses skip connections and we do not want to add more uncertainty by removing parts of the inputs from the underlying layers. During training, the dropout layer randomly drops %25 of the input from the 101-dimensional input. During test time, the input is kept as it is, but it is re-scaled to compensate for the input magnitude during training. In addition, we use a batch normalization layer after the activation function of the first layer of recurrent units to ensure a standardized latent space with zero mean and one variance.

It is shown that weight-decay, improves the quality of generalization when training deep learning. We followed the best practice in the field and applied a weight decay with the coefficient set to “0.01”. In addition to this regularization, we used “Xavier normal” initializer (Glorot & Bengio, 2010) which is recently referred to “Glorot normal” initializer. For training, we used “Adam” optimizer with an initial learning rate set to 0.001. Adam optimizer combines the benefits the two previously known extensions of gradient descent optimizer RMSProp and Adagrad (Ruder, 2016) and adapts the learning rate based on the moment information at each learning steps.

Since our dilated RNN is specifically designed for event classification in sports videos, we trained the model from scratch without using any pre-trained model. We use a mini-batch size of 32 in our training. For fair comparison, all models are trained for 25 epochs. The input data is shuffled before starting each epoch. To train the dilated RNN model, we create a training dataset based on the original dataset which contains “one-minute” clips. This is shown in Figure 4.6.



**Figure 4.6 : Process of generating one-minute training clips for RNN. Given the event annotation, we create a training set using the one-minute clips by subsampling five frames per second.**

#### **4.5 Spotting**

Event localization is defined in different ways in the research community. The most common definition is the event detection. This is shown in Figure 4.7. Basically, event detection is defined as finding the boundary (start and end) of a certain event with a correct event class associated with it. This is a reasonable and useful definition when the event has a long duration, and it has a clear start and end.



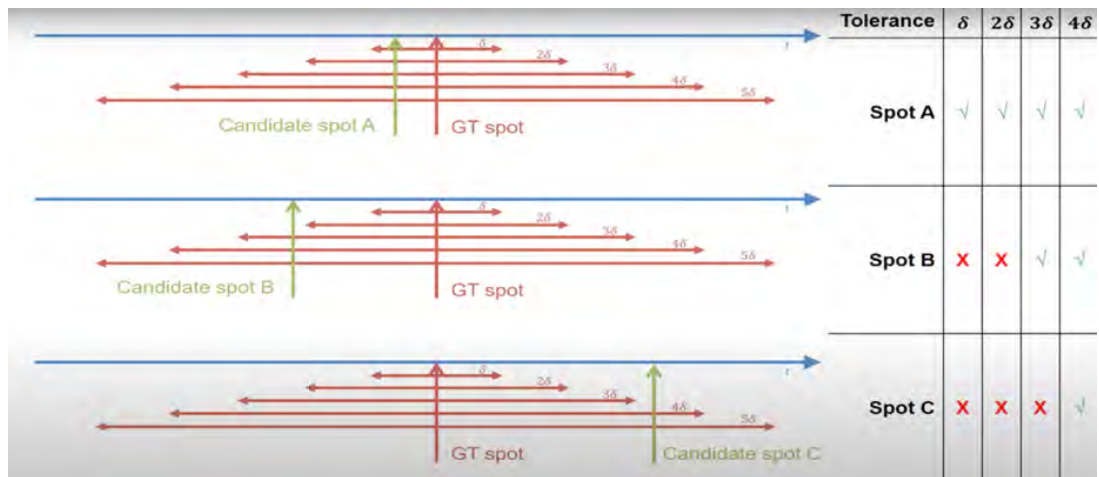


**Figure 4.7: Different temporal definition of events in videos for a single event (left) and multiple events with overlap (right) (Mleya et al., 2019).**

Unfortunately, defining an action or event boundary (start and end time) is an ambiguous task. This is also discussed in detail in (Giancola et al., 2018) where the following reasons are identified:

1. It is not clear how to define the event boundary for an event that occurs in a glimpse. For example, the goal event happens in a very short and unclear period of time.
2. Defining start and end boundary is not well-defined for continuous event within a video. For example, it is subjective to define measurable quantities for the sun rise (some have defined different light illumination condition for it).
3. For events that have an overlap (i.e., concurrent), it is hard to define the boundaries in a clear manner. For example, it is unclear how to separate call event from walking event, when someone receives a call while he is walking.

An alternative event localization definition is defined in (Giancola et al., 2018) which is referred to as “event spotting”. Event spotting is clearly defined when the events happen in a very short period of time and when the boundary (start and end time) is very hard and ambiguous to define. Spotting identifies the moment that an event occurs. A candidate spot is positive if it lands within a tolerance window around the anchor of an event. Otherwise, it is considered as negative. This is shown in Figure 4.8.



**Figure 4.8: Definition of anchor point in event spotting and the corresponding tolerance thresholds(Giancola et al., 2018).**

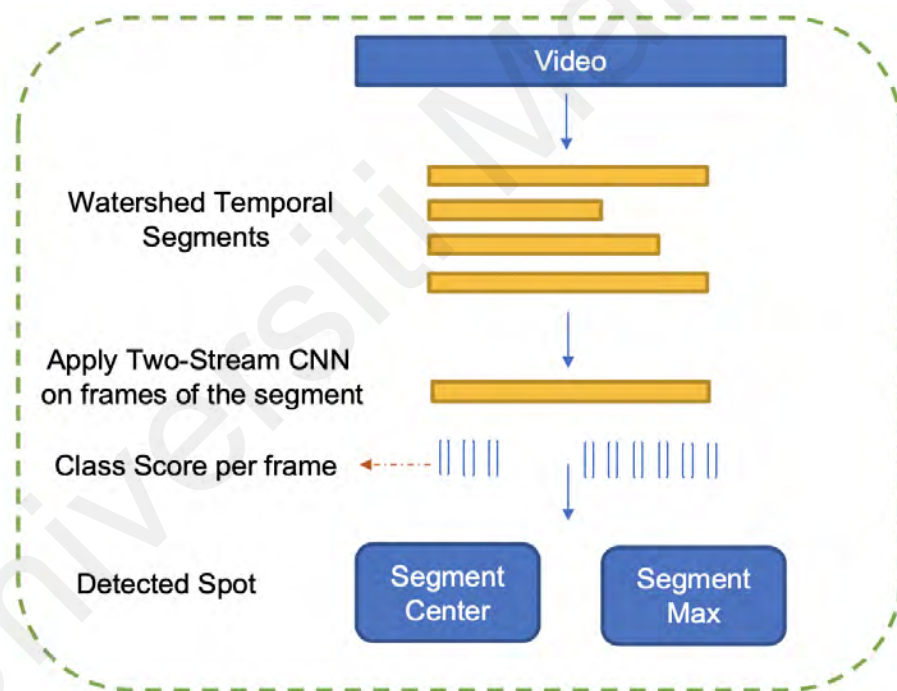
Given the fact that events in sport usually happen in a glimpse, it is reasonable to choose the second definition as event localization. As a result, in this work, we are interested in the event spotting problem. This means that, rather than identifying the boundaries of an action within a video and looking for IoU between temporal windows, we use spotting to identify the moment that an event occurs.

To identify the spots in a long video, we used the following three approaches.

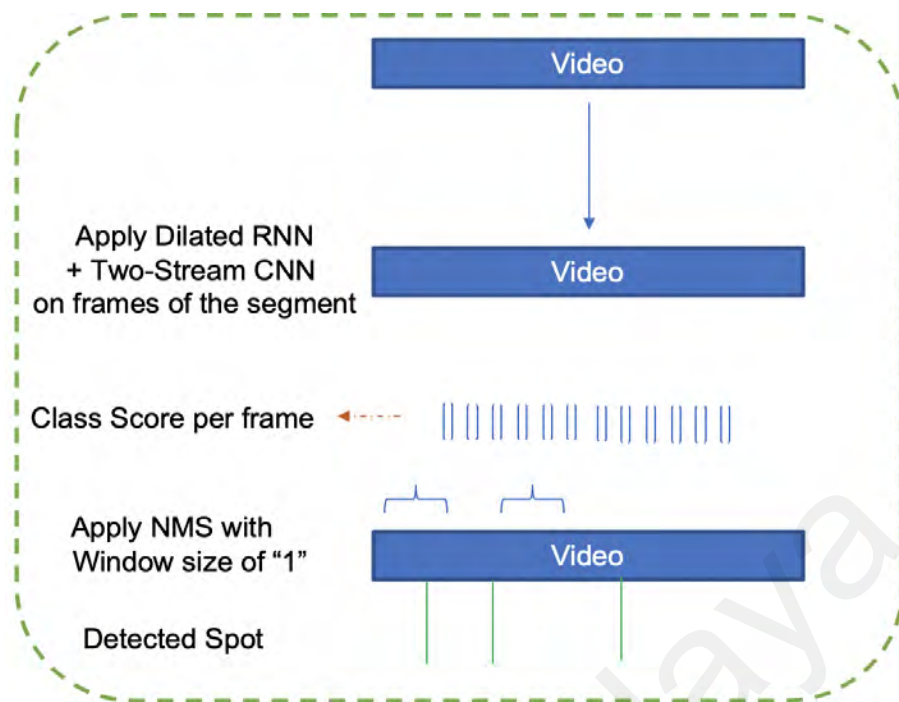
- **Segment Center:** We first use an unsupervised segmentation algorithm (Watershed) to temporally segment a given video to shorter clips. The center frames in short clips are identified as the event spots and are classified using our frame-based model trained for classification (ResNet, Two-stream CNN). This is shown in Figure 4.9.
- **Segment Max:** Similar to the above, we used the watershed algorithm to identify the shorter clips from a longer input video. For frames in each clip, we use our frame-based model trained for classification (ResNet, Two-stream CNN) to classify the event associated with each frame. The frame with the maximum classification score is selected as the event spot. This is shown in Figure 4.9.

- **Non-Maximum Suppression (NMS):** Given a football video, we classify the video frames using one of our classification models (1- frame-based model trained for classification (ResNet, Two-stream CNN), 2- temporal models (LSTM, Dilated RNN)). We then apply a one-minute non-overlapping non-maximum suppression window on the entire video. The NMS will choose the frame with a higher classification confident as the anchor of the event for that one-minute window. This is demonstrated in Figure 4.10.

Note that the first two approaches are only used for our baseline models which only operate at a frame level. Our recurrent temporal models only use the third approach.



**Figure 4.9: Spotting algorithms used based on the watershed segmentation method.**



**Figure 4.10: Spotting algorithms used based on the Non-Maximum Suppression (NMS) method.**

#### 4.6 Implementation and Technical Details

Training neural networks could be challenging. Details of hardware and software can affect the performance and accuracy of the final model. In the following sections, we provide a complete overview of the hardware used in this work as well as the development platform and third-party libraries.

##### 4.6.1 Hardware Description

All experiments are performed on an AWS EC2 Machine. We used the AWS instances with the following spec:

- CPU: Intel quad core-i7
- RAM: 61GB
- Graphic card: Tesla-K80 Nvidia card
- Memory: 11Gb

Using the AWS machines has multiple benefits. They are usually upgraded with the latest CUDA libraries and offer an easy-to-use interface to install necessarily third-party libraries. Also, there are specific nodes suitable for deep-learning training.

#### 4.6.2 Software Description

We implemented our code using Python. For classical computer vision operations such as opticalflow or some of the data processing parts, we use Opencv2.4.1. For all the deep learning modules, we use TensorFlow 1.14 (Abadi et al., 2016) in our experiments. We chose TensorFlow for the following two reasons:

1. Extensive support of pre-trained models in TensorFlow.
2. Great visualization of the network and loss during training. This is important because it allows us to debug the training behavior.

#### 4.6.3 Training, Testing and Validation Time

Table 4.2 shows the training and test time for our proposed models. Training time for a mini batch is usually close to twice the test time for the same batch. This is mainly due to the fact that during training, we process both forward and backward passes while in test time we only perform the forward pass. Note that the training time in Table 4.2 is computed for 25 epochs.

**Table 4.2 : Training and Testing (processing) time**

Model	Training Time	Testing Time
LSTM-Res	173.5H	7.2H
LSTM-2S-CNN	184.2H	7.3H
D-LSTM-Res	154.4H	7.1H
D_LSTM-2S-CNN	153.8H	6.9H

## 4.7 Summary

In this chapter, we provided more details regarding our proposed model and the neural network components. We explained how each component addresses various correlations (short-range, mid-range, and long-range correlations) between frames. We presented additional details of the proposed model and defined the steps towards the implementation and training of the proposed neural network modules.

We used the largest publicly available dataset for football video analysis. The existing approaches either use the classical machine learning models or simply use an RNN model, grounded on CNN features, for event classification. Unlike these approaches, our proposed model consists of two states of the art neural network models: a) Two-stream CNN, and b) Dilated RNN with LSTM units. Each component is carefully designed to consider certain correlation patterns between frames. Two-stream CNN is designed to learn the local spatiotemporal features which models short-range correlation among frames. The dilated RNN with LSTM cells, use the skip-connections and the memory cells to address both mid-range and long-range dependencies.

For practical reasons we used a pre-trained two-stream CNN network and trained the dilated RNN from scratch. A total number of 3965, 1314, 1358 were used for training, validation, and testing datasets respectively for both event classification and spotting. For training and testing, we used video chunks of one minute as an input of our networks. As for event spotting, we compared three different spotting approaches based on the classification results of both two-Stream CNN model and its combination with Dilated RNN. Note that the validation set is used for hyper-parameter tuning (e.g., learning rate, batch sizes) as well as identifying the epoch where the overfitting happens.

To summarize, the main contribution of our model is the proposed combination of neural network components to capture the short-range, mid-range, and long-range

dependencies to improve classification and spotting accuracy. The next chapter will present the evaluation results and discussions of the proposed model.

Universiti Malaya

## CHAPTER 5: EXPERIMENTAL RESULTS AND DISCUSSION

### 5.1 Introduction

In this chapter, we provide the present evaluation results of our proposed models for event classification and spotting in long football videos. In our discussion, we focus on explaining our understanding of the results and why do we believe some models provide better results.

For non-recurrent models, the models trained for action classification are applied in a sliding window fashion on a testing video. For a fair comparison, a stride of one is used as in (Giancola et al., 2018). For the recurrent models, a temporal sub-sampling rate of five frames per second is applied. Table 5.1 summarizes the evaluation categories we present in the rest of this chapter.

**Table 5.1 : Overview of test and evaluation of the proposed model for event spotting in Football videos**

<b>Test and evaluation result of the proposed model for analyzing long football videos</b>	
<b>Test and evaluation of the classification result</b>	<b>Test and evaluation of the spotting result</b>
<ul style="list-style-type: none"><li>• Evaluation of the classification results related to proposed two-Stream CNN and dilated RNN.</li><li>• Discussion of the results related to the proposed two-Stream CNN and dilated RNN.</li><li>• Comparison between proposed model with the base line.</li></ul>	<ul style="list-style-type: none"><li>• Evaluation of events spotting results using segment center, segment max, NMS methods and two-Stream CNN.</li><li>• Evaluation of events spotting results using NMS method and dilated RNN.</li><li>• Discussion of the results related to the proposed spotting method</li><li>• Comparison between proposed model with the base line.</li></ul>



## 5.2 Test and Evaluation of the Results for the Proposed Model

As it has been briefly discussed in the introduction, event localization (either spotting or detection) strongly depends on the accuracy of event classification. The main difference between classification and localization is the fact that in the classification problem we are given a trimmed football video clip (a video which solely contains a single activity), and the goal is to classify the activity class in that video clip. For the localization problem, given an untrimmed football video, the goal is to identify the events in that video. This means to temporally localize the events in the video and to classify the event category in the video clip. As a result, the quality of the event classification results directly impacts the quality of the localization results.

For a comprehensive evaluation, similar to (Gu et al., 2018), we also report the classification results on one-minute video chunks for our proposed model and the baselines. The test and evaluation results of the proposed event classification model include two parts:

1. The single frame classification results which compare our proposed two-stream CNN with the baselines and the state of the art.
2. The novel proposed holistic model in this work which combines the two-stream CNN and the dilated RNN with LSTM cells.

In the following, Section 5.2.1 provides the evaluation results for event classification and Section 5.2.2 provides the evaluation results for event spotting.

### 5.2.1 Event Classification

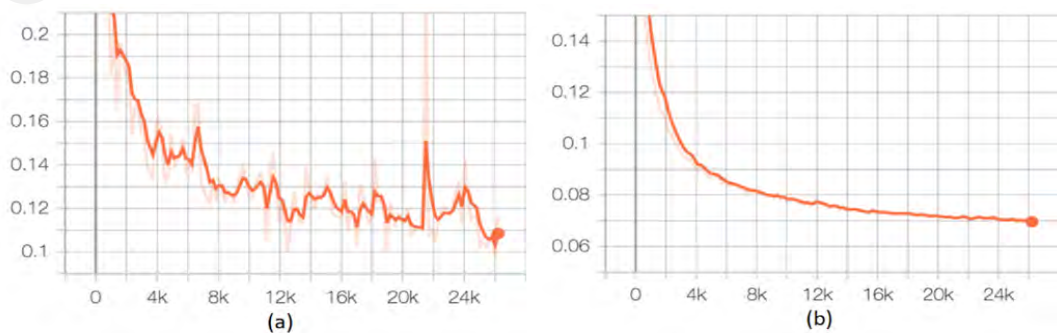
The proposed model in Section 4.2 is a combination of two subnetworks, the Two-Stream-CNN and Dilated-RNN. To better analyze the impact of each of these network components in the final accuracy, we conducted a set of experiments with each

subnetwork to provide empirical results on how much each network contributes to the final accuracy.

For pure CNN based models (e.g., Two-stream CNN), we use a SoftMax layer with a four-class output to classify four events (card event, goal event, substitution event, and the background event - i.e., no event). For RNN based models (e.g., Dilated RNN + Two-stream CNN), we do not use a sliding window. Instead, at each time step, “t”, the input to the network is the frame at time “t”. The output of the network is a four-dimensional classification layer with a SoftMax activation function. To identify the associated class for the entire video chunk, we can either average the results from the SoftMax layers along the time axis, or just take the maximum. In our experiments, we find out that max operation results in better classification accuracy. For this reason, the results provided below use max operation to fuse the outcomes of all-time steps.

For a fair comparison, we used similar one-minute video chunks used in (Giancola et al., 2018). The test subset includes 326, 453, 579 videos for goals, cards, and substitutions respectively. We also added 460 random background video clips as well.

Figure 5.1 shows the average training and validation loss of our full model (dilated RNN + two-stream CNN) per-epoch. Note that since the validation loss is computed over the entire validation set, it is smoother compared to the training loss.



**Figure 5.1: Average lost per training epoch. (a) training loss, (b) validation loss.**

Considering each class separately, one can compute the average precision (AP) by casting the results as a binary classification problem. For example, when considering the “card” event, we consider all other events as the “non-card” event. For this, we can compute the AP for the “card event”. As a result, we have four AP values for each event. Taking another average over the classes results in the mean average precision or mAP.

Table 5.2 presents the result for event classification in the SoccerNet dataset for one-minute videos chunks.

**Table 5.2: Accuracy result of the proposed models for event classification presented as mean average precision (mAP).**

Approach	mAP	Improvement
ResNet-Max Pool	52.4	-
(Giancola et al., 2018)	67.8	-
2SNet-Max Pool	57.8	5.4
D-RNN-Res	60.8	8.4
D-RNN-2SNet	62.7	10.3
D-LSTM-Res	65.3	12.9
<b>D-LSTM-2SNet (Proposed model)</b>	<b>69.9</b>	<b>17.5</b>

We compare different variants of our models with the result reported in (Giancola et al., 2018). The first variant is a two-stream CNN with max-pooling used as the fusion layer. This is referred to as “2SNet-Max Pool”. The other two models are two different dilated RNN models with different CNN backbones, ResNet and Two-stream CNN referred to as D-LSTM-Res and D-LSTM-2SNet. The “NetRVLAD” is the model proposed in (Giancola et al., 2018).

### 5.2.1.1 Summary and Discussion of Event Classification

As it is shown in (Table 5.2) the Dilated-LSTM with two-stream CNN backbone outperforms baselines and different variants of our approach. This is specifically important comparing D-LSTM-2SNet and D-LSTM-Res which supports our hypothesis

that modeling short-range correlation using a specific temporal CNN model is important and it results in 4.6% accuracy improvement. Another supporting evidence for this is the 5.4% improvement from “2SNet-Max pool” compared to “ResNet-Max pool” which is gained by applying a temporal CNN based model.

On the other hand, comparing pure CNN-based models and Dilated-RNN models we observe a large accuracy improvement. D-LSTM-Res improves the mAP accuracy 4.5% -12.5% compared to the other variance of the ResNet-based models, and D-LSTM-2SNet improves the accuracy 12.1% compared to two-stream CNN. This clearly shows that considering long-range dependencies is very important and clearly contributes to the overall accuracy.

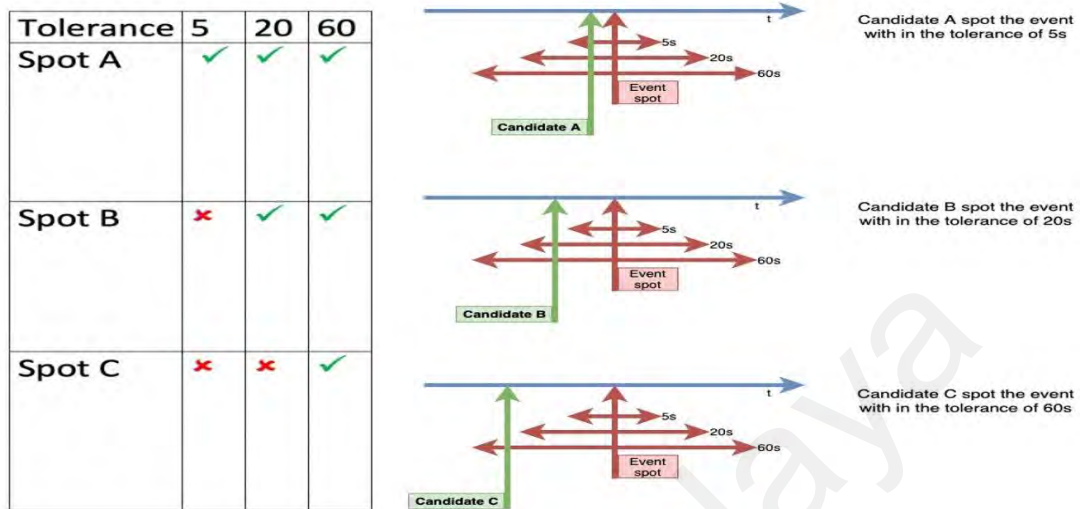
Finally, comparing results from D-LSTM and D-RNN shows 4.5-7.2% accuracy improvement. This is an important finding as it shows that LSTM units are important to capture the mid-range dependencies.

To summarize, our classification model which uses the dilated LSTM network, and the two-stream CNN component together was able to improve the overall accuracy by 2.1% compared to the state of the art and up to 17.5% compared to our simple baseline. The ablation study shows that each component has contributed to the overall accuracy. On average, the two-stream CNN improved the accuracy by 4%. The LSTM has improved the accuracy by 5.85% on average among different models. And finally, the dilated RNN improves the accuracy by at least 2.1% and up to 17.5%.

### **5.2.2 Event Spotting**

Our proposed event spotting approach is built on top of the event classification model. As explained in Section 3.5, an event is spotted correctly, if the estimated event frame

falls within a threshold of the ground truth frame. Figure 5.2 shows different variations of event spotting and various thresholds defined as correctness tolerance.



**Figure 5.2: Illustration of an anchor point in event spotting. For a given anchor point, we consider multiple error tolerance thresholds. A candidate event spot is correct if it falls in the error tolerance window and it is incorrect otherwise.**

Similar to event classification, since the model consists of multiple subnetworks, we evaluate each subnetwork separately. For event spotting, we have three different approaches for temporal localization.

1. One approach is based on an unsupervised temporal segmentation, “WaterShed” algorithm (Chien, Huang, & Chen, 2003), to create temporal segments from a long video.
2. Second approach is based on a “uniform one-minute” segmentation from the original video.

Given a temporal segmentation from one of the above algorithms, we use an event classification models to classify frames within the segment.

The combination of different event classification models and different temporal segmentation results in various event spotting approaches. We conduct a comparison with

several baselines and variations in order to evaluate the effectiveness of our model. In the following, we first explain different baselines and variations of our approach in detail. Later we will provide multiple comparison results between various approaches.

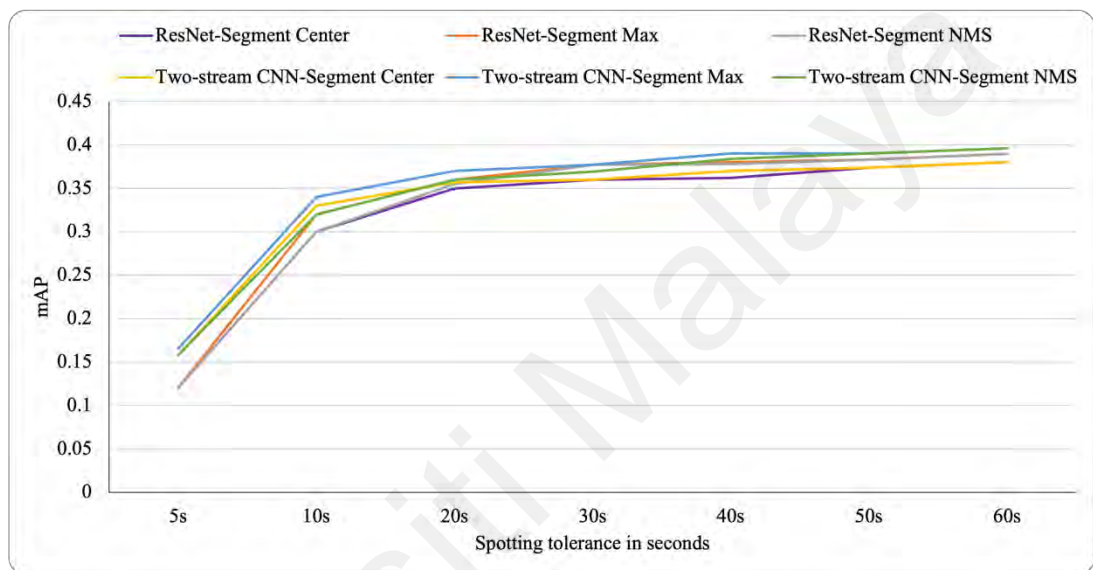
### 5.2.2.1 Baselines

The baseline approaches are all frame-based models. This means that they do not use any recurrent component. These models use a CNN backbone (two-stream CNN or ResNet) to classify frames in the given temporal segment. Table 5.3, introduces various baselines, and explains the CNN model and the segmentation approach used in each.

**Table 5.3 :Definition of different event spotting baselines.**

Model Name	CNN backbone	Temporal Segmentation	Description
‘2SNet-Segment Center’	Two-stream CNN	Watershed	Classifies the segment based on the event scores of the center frame in the segment.
‘2SNet-Segment Max	Two-stream CNN	Watershed	Classifies the segment based on the event of the frame with maximum event scores among the frames in the segment.
‘2SNet-Segment NMS’	Two-stream CNN	Uniform one-minute	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment.
‘ResNet-Segment Center’	ResNet	Watershed	Classifies the segment based on the event scores of the center frame in the segment.
‘ResNet-Segment Max	ResNet	Watershed	Classifies the segment based on the event of the frame with maximum event scores among the frames in the segment.
‘ResNet-Segment NMS’	ResNet	Uniform one-minute	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment.

Figure 5.3, demonstrates the mAP results for different baselines explained in Table 5.3. The plots show the mAP results for different error tolerance represented in seconds. Note that the best result in (Giancola et al., 2018) is achieved using an additional NetVLAD model after the classification layer. Similar to the baselines, we also apply PCA dimensionality reduction on CNN features with 512 dimensions. Similarly, we used the threshold of 50% for the watershed algorithm.



**Figure 5.3: Spotting results for variants of our single-frame models (mAP vs error tolerance threshold). All models are trained on 60-second videos.**

One important observation is that for lower tolerance thresholds, all two-stream neural network-based baselines outperform the equivalent baselines which operate on ResNet features. This suggests the importance of capturing both spatial and temporal information.

### 5.2.2.2 Variations of Our Approach

Evaluation result from baseline approaches allows us to demonstrate the effectiveness of two-stream CNN in capturing the local spatiotemporal features. To better understand the efficacy of the DialtedLSTM in modeling mid-range and long-range correlations between frames, we define multiple variations of our model, each using distinct recurrent network components. Unlike the baselines, we only use the NMS spotting algorithm for

recurrent methods for the following two reasons. First, the boundaries of the watershed segments are not well-aligned with the ground truth event spots. Second, this limits the information flow from previous frames which are not in the watershed segment. More importantly, the key idea behind using recurrent models is to allow the network itself to learn how to regulate the information flow. In our view, providing the pre-processed segments is not well-explained with the theory behind the recurrent models.

We study two variants of LSTM networks: 1) Standard LSTM network, and 2) DilatedLSTM which is a DilatedRNN with LSTM cells. We ground these recurrent networks on the features computed using two-stream CNN or ResNet50. In total, combining the recurrent component and the CNN backbone defines six different variants of our proposed spotting algorithm.

Table 5.4, provides details of the ML model and the temporal algorithm in multiple variations of our approach.

**Table 5.4 : Definition of different variation of our approaches.**

<b>Model Name</b>	<b>CNN backbone</b>	<b>Description</b>
LSTM-Res	ResNet	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment. A plain recurrent model with LSTM units which operates on the features extracted from a ResNet model is used to classify each frame.
LSTM-2SNet	Two-stream CNN	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment. A plain recurrent model with LSTM units which operates on the features extracted from a Two-stream CNN model is used to classify each frame.



**Table 5.4 Continued**

<b>Model Name</b>	<b>CNN backbone</b>	<b>Description</b>
D-RNN-Res	ResNet	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment. A dilated recurrent model with vanilla RNN cells which operates on the features extracted from a ResNet model is used to classify each frame.
D-RNN-2SNet	Two-stream CNN	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment. A dilated recurrent model with vanilla RNN cells which operates on the features extracted from a Two-stream CNN model is used to classify each frame.
D-LSTM-Res	ResNet	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment. A dilated recurrent model with LSTM cells which operates on the features extracted from a ResNet model is used to classify each frame.
D-LSTM-2SNet	Two-stream CNN	Classifies each one-minute segment by applying a non-maximum suppression (NMS) over the estimated event classes for frames in one-minute segment. A dilated recurrent model with LSTM cells which operates on the features extracted from a Two-stream CNN model is used to classify each frame.

Table 5.5, demonstrates mAP result for spotting football events trained on 5, 20, and 60 seconds, averaged over three different error thresholds of 5 seconds, 20 seconds, and 60 seconds for the event spotting algorithms proposed in Chapter 4. For this experiment, we use the uniform “one-minute” segmentation followed by a non-maximum suppression of “one-minute” window. As it is shown in the table, among the non-recurrent models, the 2SNet model outperforms the ResNet model. This shows that considering a temporal stream and identifying local spatiotemporal features helps the model to classify the events with more accurate confidence which results in better event spotting. Comparing the recurrent variants of our model (i.e., LSTM-Res and LSTM-2SNet) with the frame-based

baselines show that including mid-range temporal information increases the overall spotting accuracy up to 2.3%. Finally, dilated recurrent based models (i.e., D-LSTM-Res and D-LSTM-2SNet) gained up to 9.8% accuracy improvement compared to vanilla recurrent models. One key observation is that for smaller error thresholds, the accuracy gain of using dilated recurrent model is smaller than the accuracy gain compared to larger error thresholds. We believe this is due to the fact that for smaller error thresholds fine grain details of local spatiotemporal features are more important in accuracy gain. For larger thresholds, the gain of using dilated recurrent networks is more evident and more significant as long-range correlation is more important to consider. The best results are obtained with a model that combines all three ranges of correlation, D-LSTM-2SNet. Note that since the authors in (Giancola et al., 2018), did not provide a tabular result for their spotting model, the numbers in Table 5.5, are based on the plots reported in the original paper and their text. Compare to (Giancola et al., 2018) our best model achieves 5.4%-6.9% improvement.

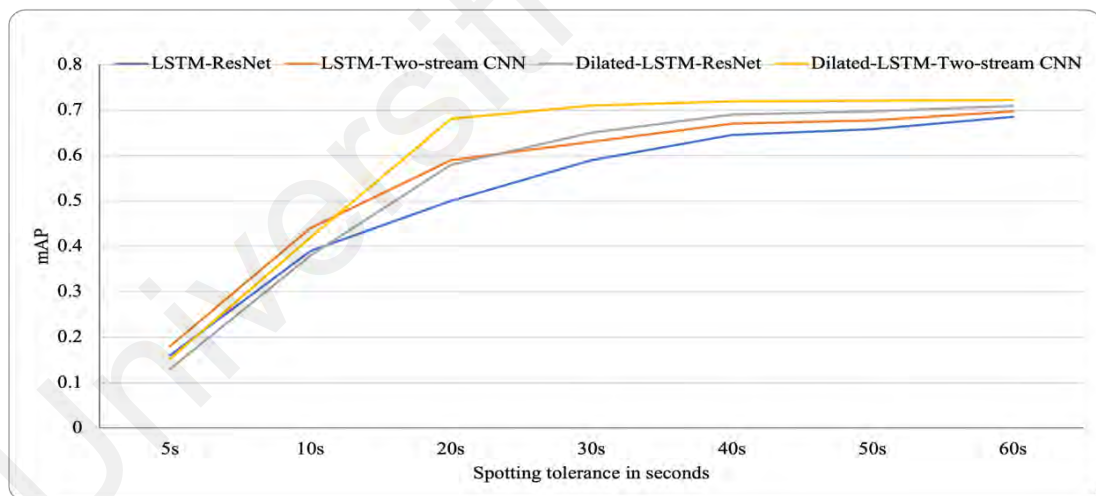
**Table 5.5: mAP result for spotting football events trained on 5, 20, and 60 seconds, averaged over three different error thresholds of 5 seconds, 20 seconds, and 60 seconds.**

Model	Error Tolerance = 5s	Error Tolerance = 20s	Error Tolerance = 60s
<b>(Giancola et al., 2018)</b>	3	35	59
ResNet	2	29.7	57.2
2SNet	7.7	31.1	60.6
LSTM-Res	4.3	30.5	60.2
LSTM-2SNet	8.7	33.4	61.7
D-LSTM-Res	4.5	40.3	63.9
<b>D-LSTM-2SNet</b>	<b>9.2</b>	<b>41.9</b>	<b>64.4</b>

Figure 5.4, illustrates the mAP results of the model trained on 60-second videos as a function of error tolerance for our recurrent models, which allows us to understand the positive and negative contributions of each network.

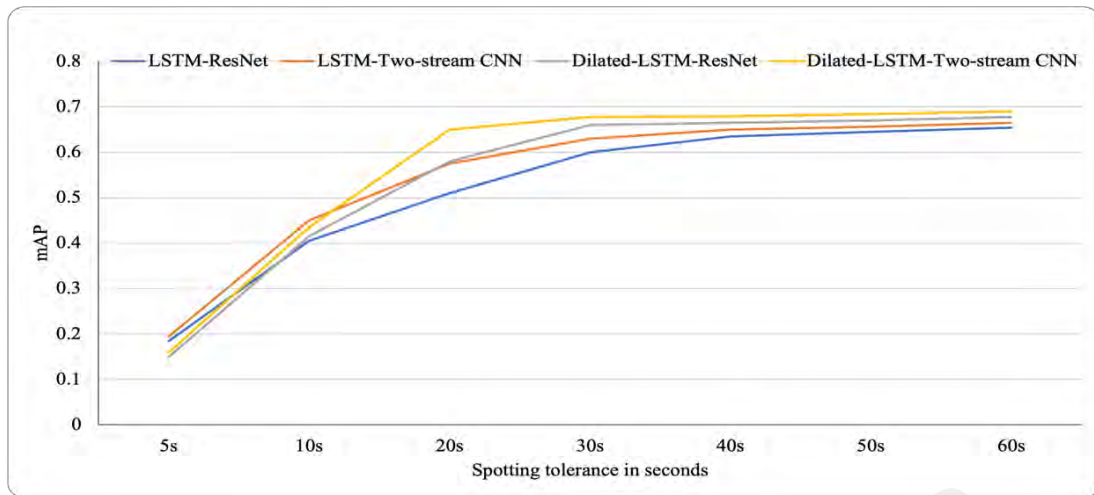
As it is shown, the standard LSTM variants outperform the dilated variants for error thresholds smaller than 10 seconds. We believe that fine-grain details captured by local spatiotemporal features provide more accuracy gain for smaller error thresholds. Our understanding is that although the accuracy difference is insignificant compared to other variants, considering the information from faraway frames potentially adds more noise which decreases the accuracy.

For greater threshold values, the gain of using dilated recurrent network is more evident and more significant, which indicates that long-range correlations are more critical to consider. At first glance, this seems counterintuitive. After a deeper analysis of the result, we identified that the DilatedRNN variants make fewer mistakes if the error threshold is between 10-20 seconds. On the other hand, if the error threshold is less than 10 seconds, local spatiotemporal models are more accurate.

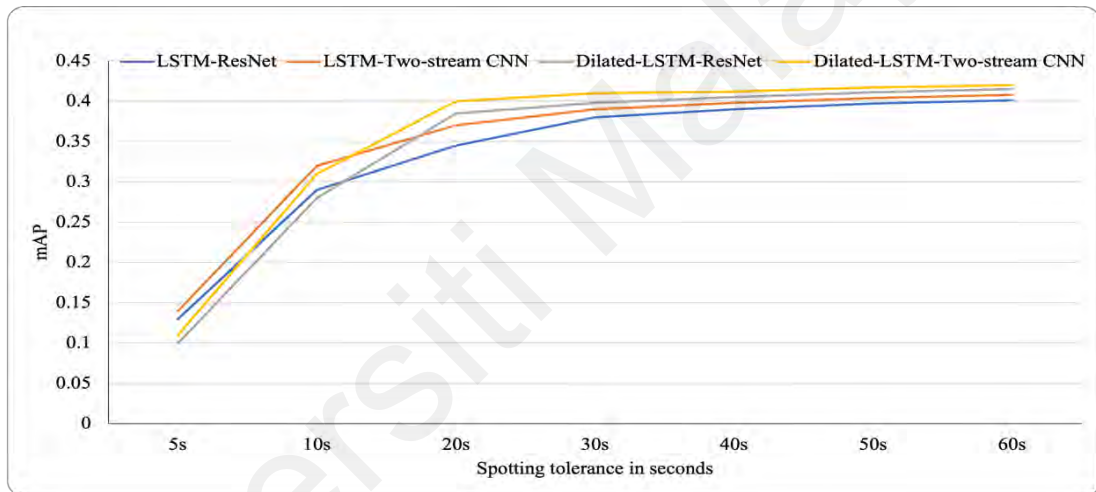


**Figure 5.4: Spotting results for variants of our recurrent-based models (mAP vs error tolerance threshold). All models are trained on 60-second videos.**

As discussed in Chapter 4 for training dilated RNN based models, we use “one-minute” video clips. To better understand the effect of the clip’s length, we also trained models on shorter clips of five and twenty seconds. The mAP results for various error tolerance for these models are shown in Figure 5.5 and Figure 5.6.



**Figure 5.5: Spotting results for variants of our recurrent-based models (mAP vs error tolerance threshold). All models are trained on 20-second videos.**



**Figure 5.6: Spotting results for variants of our recurrent-based models (mAP vs error tolerance threshold). All models are trained on 5-second videos.**

We observe a similar pattern between mAP and error tolerance for models trained with five-seconds and twenty-seconds videos compared to the models trained with one-minute clips. Also, comparing the results from models trained on videos of 20 and 60 seconds, it seems that capturing long-term dependencies is more successful in longer videos and the results are generally better for error thresholds above 25 seconds.

In addition, Table 5.6 demonstrates the average mAP results across different error tolerance for variations of our event spotting algorithm. In other words, this table

represent area under curve for mAP plot (AUC). We train each model with video clips of 5, 20, and 60 seconds. Five-second video clips represent settings where the model is limited to short-range dependencies during training. Training with 20- seconds and 60-seconds video clips allows the model to explore mid-range and long-range dependencies as the corresponding models have access to 4 times and 12 times more previous frames each.

**Table 5.6: Comparison of our proposed event spotting approach compared to state of the. The models are trained on 5, 20, and 60 second videos. The results show the area under curve (AUC) of the mAP plots shown in Figure 5.4.**

Model	5s videos	20s videos	60s videos
<b>(Giancola et al., 2018)</b>	35.1%	49.7%	40.6%
ResNet	29.4%	32.1%	33.2%
2SNet	29.7%	32.4%	34.0%
LSTM-Res	34.3%	54.1%	55.1%
LSTM-2SNet	36.2%	57.3%	58.1%
D-LSTM-Res	36.1%	58.3%	59.2%
<b>D-LSTM-2SNet</b>	<b>37.3%</b>	<b>60.1%</b>	<b>63.3%</b>

Similar to the studies of the baseline methods, the two-stream CNN model outperforms the ResNet model. In particular, the margin is higher for shorter clips of sizes 5 and 20 seconds. These results verify that if we only consider short-term dependencies, the models with explicit motion features provide a richer representation of the underlying raw frames.

Based on the results presented in Table 5.6, recurrent variants of our approach perform better than the single-frame variant across different training settings. One interesting observation is that for models trained on five-second videos, the best result is 7%, but the average mAP increases to more than 13% for models trained on 20-second videos. We believe this shows the importance of capturing mid-range correlations between video frames. Finally, dilated LSTM variants gain up to 9.8% accuracy improvement compared to the standard LSTM models. One key observation is that for short video clips, the

accuracy gain of using dilated recurrent model is insignificant compared to the accuracy improvement for longer video clips. This observation verifies our original hypothesis that sophisticated modeling of long-range dependencies is essential for more accurate event spotting.

Table 5.7 compares our best performing model, which grounds DilatedLSTM on two-stream CNN features, to the state of the art. We provide the best reported results from three prior work (Giancola et al., 2018), (Cioppa et al., 2020) and (Vats et al., 2020). While the authors in (Giancola et al., 2018) specifically mentioned that their best result is obtained training on 20-second videos, for (Cioppa et al., 2020) and (Vats et al., 2020) it is not clear from the text which video lengths the model is trained with. Our full model improves the result by 13.6% compared to (Giancola et al., 2018). We also outperform the models proposed in (Cioppa et al., 2020) and (Vats et al., 2020) by 0.8% and 3.2% respectively.

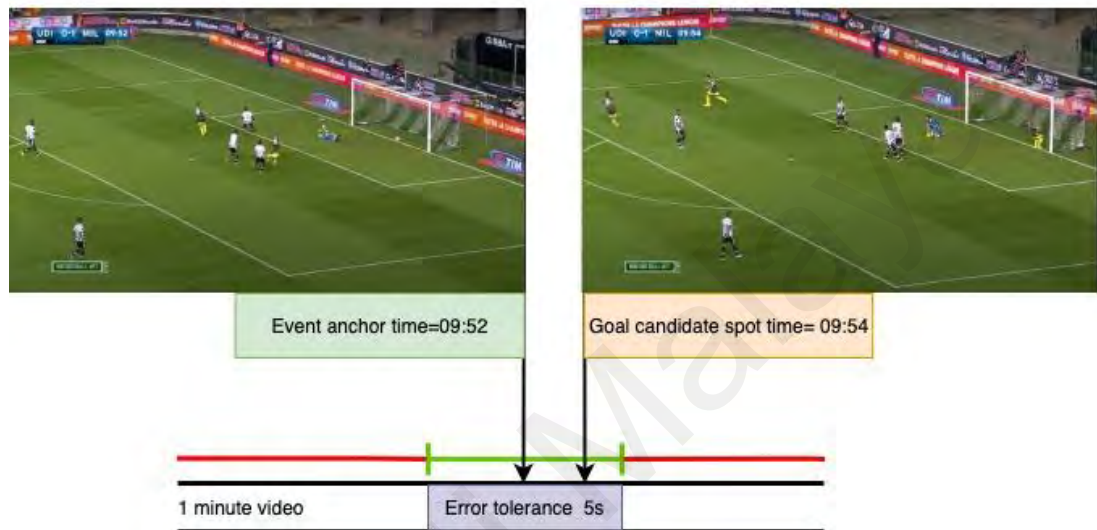
**Table 5.7: Comparison of our proposed event spotting approach compared to state of the art. The reported mAP results from the state of the art are from published results. Note that the best result reported in (Giancola et al., 2018) is trained on 20-second videos. For the results in (Cioppa et al., 2020) and (Vats et al., 2020) it is not clear what video length the models are trained on. Our best result is obtained from models trained on 60-second videos. This shows that explicit modeling of long-range dependencies improves the accuracy for longer videos.**

Model	mAP
(Giancola et al., 2018) [trained on 20-second videos]	49.7%
(Vats et al., 2020)	60.1%
(Cioppa et al., 2020)	62.5%
<b>Ours [trained on 60-second videos]</b>	<b>63.3%</b>

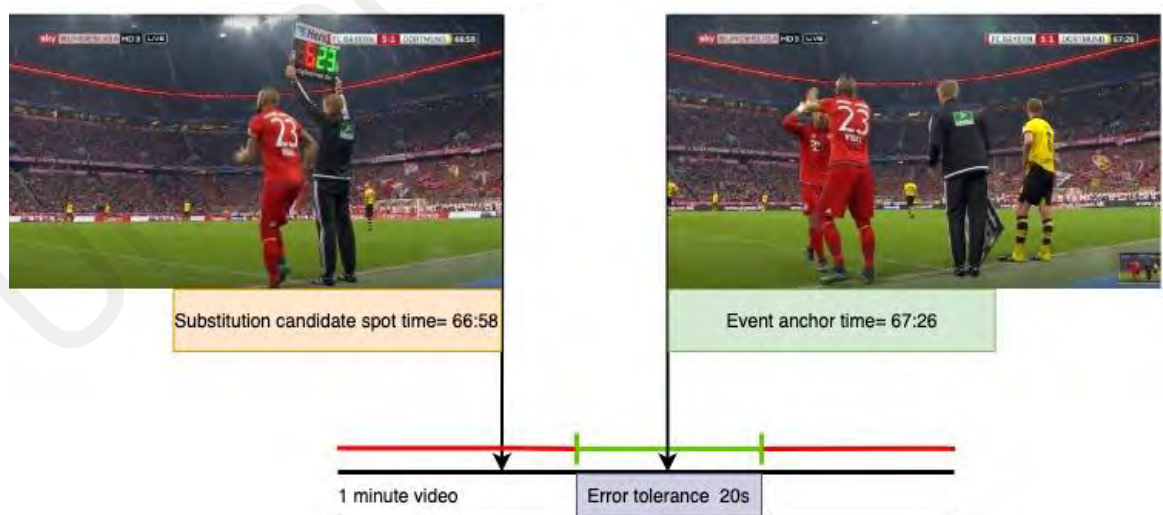
### 5.2.2.3 Qualitative results

To better understand the spotting results and to explain the outcome of our proposed model, we randomly sampled some positive and negative examples and visualized the results. In the following, we will provide a few qualitative spotting examples of our model (two-stream CNN + Dilated RNN) and the center segment spotting baseline for 1-minute video clips. Figure 5.7 shows a successful event spotting example. The sample candidate

point (goal candidate spot time = 9:54) in Figure 5.7 landed within an error tolerance of **5 seconds** around the anchor of the event (goal event anchor time = 9:52). Figure 5.8 shows a failed event spotting sample. The candidate spot (substitution candidate spot time = 66:58) in ) Figure 5.8 is out of the error tolerance of **20 seconds** around the anchor of the event (substitution event anchor time = 67:26).



**Figure 5.7: Successful goal spotting example for 5s tolerance (Italy Série A, 2015-2016/2015-09-22, 21:45, Udinese 2 - 3 AC Milan.**



**Figure 5.8: Failed substitution spotting example for 20s tolerance (Germany Bundesliga, 2015-2016/2015-10-04, 18:30, Bayern Munich 5 - 1 Dortmund).**

The presented examples in Figure 5.9, demonstrate successful spotting within 60 seconds of the event anchor. Given these examples, we can see that even when the error tolerance is 60 seconds, the spotted event is well-correlated with the event boundaries.



**Figure 5.9 : Successful examples of goal, substitution, and card event spotting ((a) presents goals sample, (b) presents substitution sample and (c) presents cards sample).**

Finally, it is informative to better understand the important spatial regions in video frames which impact the neural network output. This has been mostly investigated in the sub-field of explainable AI (XAI). Three different categories of explainable AI have been studied in the literature, namely "pre-modeling explainability", "explainability modeling", and "post-modeling explainability". The majority of deep learning models are trained with estimation performance as the main criteria, resulting in black-box ML models. As a result, the "post-modeling explainability" is studied in more detail in the literature. Since we use two-stream CNN and RNN based architectures, we are also categorized as the former. More details regarding investigating XAI in our work is provided in APPENDIX C.



#### 5.2.2.4 Summary and Discussion of Event Spotting

To summarize our observations, from the results provided in Figure 5.3-Figure 5.6 and Table 5.5-Table 5.8, we can clearly see that the two-stream CNN network component, contributes to the improvement of accuracy which results in up to 5.7% accuracy improvement compared to the ResNet variants. This contribution is more obvious for smaller threshold error tolerances. On the other hand, the dilated recurrent models outperform all variants of single-frame approaches. Comparing to the state of the art, while the accuracy improvement is more significant and consistent using all components (i.e., 5.4%-6.9% accuracy improvement with Dilated RNN+ LSTM+ Two-stream CNN), we still observe a considerable amount of improvement (1.5%-5.7%), by only using two out of three components. This also clearly shows that considering mid-range and long-range temporal correlation between frames improves the accuracy in general. Note that as pointed out, among the dilated recurrent models, the one with LSTM cells which is grounded on Two-stream CNN outperforms all others due to the fact that it considers short-range, mid-range, and long-range correlations together.

In addition to the above, we have also identified multiple approaches for spotting the events including an unsupervised “watershed” segmentation followed by the center of the segment spotting, max confidence spotting or “non-maximum suppression” (i.e., NMS) spotting which relies on scanning a “one-minute” window. Based on the results the NMS-based models outperform the unsupervised segmentation models. As we can see the NMS with Dilated-RNN grounded on two-stream CNN achieves an average accuracy of 38.5% (best 64.4%) compared to the result reported in (Giancola et al., 2018) which is 32.3% (best 59%). Table 5.8, summarizes improvements achieved using different networks and error tolerance.

**Table 5.8 :A summary of improvements based on the results presented in Table 5.5 compare to different baselines and state of the art.**

<b>Short term, Mid Term and Long-Term Comparison</b>	<b>Average improvement for 5 second</b>	<b>Average improvement for 20 second</b>	<b>Average improvement for 60 second</b>	<b>Average improvement for all error tolerance</b>
ResNet CNN vs 2-stream CNN networks <sup>11</sup> :	4.93%	1.96%	1.8%	2.9 %
RNN with LSTM vs CNN models <sup>12</sup> :	1.65%	1.55%	2.05%	1.75%
Dilated RNN vs non-Dilated networks <sup>13</sup> :	0.35	9.15%	3.55%	4.35%
Proposed model vs baseline (ResNet)	7.2%	11.4	7.2%	8.6%
Proposed model vs state of the art (Giancola et al., 2018)	6.2%	6.9%	5.4%	6.2 %

### 5.3 Summary

In this chapter, we presented the evaluation result of the proposed models for event classification and spotting in long football videos. This includes the evaluation of event classification using single-frame approaches (e.g., Two-stream CNN) and temporal models (e.g., LSTM-Two-stream). Also, for event spotting, we evaluated multiple different neural network models as well as different event spotting approaches.

To better analyze the impact of each component in our proposed models, we identified multiple baselines and variations of our approach. Through extensive evaluation, we showed that all Two-stream CNN models outperform the ResNet models for event

---

<sup>11</sup> 2SNet vs ResNet, LSTM-2SNet vs LSTM-Res and D-LSTM-2SNet vs D-LSTM-Res

<sup>12</sup> LSTM-Res vs ResNet and LSTM-2SNet vs 2SNet

<sup>13</sup> D-LSTM-Res vs LSTM-Res and D-LSTM-2SNet vs LSTM-2SNet

classification by 17.5%. This confirmed our original hypothesis that local spatiotemporal features are important to be considered specifically. We observe a similar accuracy improvement in event spotting experiments.

Another set of experiments also showed that specifically modeling the temporal correlation for mid-range and long-range dependencies can improve the accuracy of the classification by 8.7% and the accuracy of the spotting 3.8%-12.2% compared to the baselines.

Universiti Malaya

## CHAPTER 6: CONCLUSION AND FUTURE DIRECTION

In this thesis, we addressed the problem of event localization in long football (soccer) videos. In Chapter 1, we explained that event localization in football videos is an active research problem with a strong industry impact. We formally explained the research questions and the objectives that motivated us to work on this research. In the rest of this chapter, we will review our research questions and objectives and provide details on how we achieved our goals. We will then highlight our contribution and provide some insight into our research findings. At the end, we will conclude with future works that can be built on top of this work.

### 6.1 Achievement of Research Objectives

As discussed in Section 1.4, the goal of this research is to propose an approach based on modern machine learning techniques to improve the quality of the automatic event classification and localization in football videos. The proposed model should increase the accuracy of localizing and classifying of the three important events (goal, card, and substitution) in football videos. As shown in Figure 4.1, to achieve this goal, a unified neural network model was carefully analyzed and designed. The research objectives outlined in Section 1.4 are discussed as follows:

**“To investigate the state-of-the-art feature extraction models in videos to improve the event classification and spotting in football videos”-** As it is stated in Chapter 2, we reviewed a considerable number of prior works on football/sport analysis as well as multiple relevant papers from non-sport event detection in videos. We reviewed both prior approaches which were built on top of hand-crafted features (e.g., HOD, HOF) as well as modern approaches which mainly use deep neural networks (e.g., CNN, RNN). We also reviewed prior work on football video analysis in more details to better understand the latest research in this area and to identify the limitations of the current approaches. Based

on our reviews, we concluded that specifically modeling short-range, mid-range, and long-range dependencies is the key to improve localization accuracy in long football videos. While prior approaches have addressed various frame dependencies individually, to the best of our knowledge and based on our literature review of the prior work, none of these models have addressed all these frame correlation categories together.

In Addition, our study shows that events in football videos are highly **correlated** in time. These temporal **correlations** are either **short-range, mid-range, or long-range**. Short-range correlations consider the dependencies between frames which are in a narrow neighborhood window (e.g., 1-5 frames). Mid-range correlations consider the dependencies that are beyond 5 frames but still happen in a short interval (e.g., 5-10 seconds). Long-range correlations are the dependencies between frames that are beyond the mid-range dependencies.

**“To design and implement a neural network model to improve event localization in long football videos”**- Based on our findings that no prior work has considered all types of frame correlation in a single model, we proposed a novel unified neural network architecture, which we implemented using TensorFlow framework, and trained it to classify and spot events in football videos (Chapter 4). The proposed model uses one of the most successful CNN architectures, Two-stream CNN, to compute local spatiotemporal features which describe the short-range dependencies. It also uses dilated RNN with LSTM cells to model the mid-range and long-range dependencies. While LSTM’s memory cell allows the model to handle mid-range dependencies, the dilated RNN architecture with skip connections allows the information flow from distant frames. This enables the model to better understand the long-range correlations among frames.

**“To evaluate accuracy of the proposed algorithm for event localization and classification in long football videos”**- To demonstrate the effectiveness of our proposed

model, we evaluated our approach on the largest publicly available football dataset, SoccerNet in Chapter 5. The SoccerNet dataset contains a total number of 764 hours of football videos from major European leagues. To better understand the impact of each component in our proposed neural network, we perform an extensive ablation study. We identified multiple baselines and variations of our model with different components and evaluated our model for both event classification and spotting. For event spotting we identified three different spotting methods based on three different temporal segmentation algorithms. As reported in Chapter 4 and Chapter 5, using the correct feature extraction and classification model and modeling various correlation among video frames in football videos, enabled us to improve the accuracy of both event classification and spotting in long football videos. We were able to improve the event classification mAP by at least 12.1% and up to 17.5% compared to baselines and 2.1% compared to the state of the art. The event spotting accuracy (between different error tolerances) was improved by at least 3.8% up to 12.2%. compared to baselines and by at least 5.4% up to 6.9%. compared to state of the art.

## 6.2 Main Contribution

We believe that this research extended the prior work in event localization in football videos and made several key contributions. The following summarizes the main contributions of this work compared to prior work.:

- We identified that modeling various correlation among frames is the key to improve the event localization accuracy: By reviewing the previous work in event localization and studying the limitations of each model, we realized that in order to improve localization accuracy it is important to model various correlations among frames, ranging from short to long-range.

- We used two-stream networks to model short-term dependencies between frames. Two-stream convolutional neural networks are powerful local spatiotemporal feature extraction models. We successfully exploit the descriptive power of this feature extraction method to improve the model's accuracy specifically for short-range dependencies.
- We model the mid-range correlation between frames using LSTM units in recurrent neural networks. The main strength of an LSTM unit is the explicit memory unit which allows the follow of information from previous time steps. We benefitted from this and used LSTMs to model the mid-range dependencies between frames.
- We proposed to use dilated-RNN, a hierarchical recurrent neural network with skip connections, to model the long-range dependencies between frames. While LSTM is a powerful model to capture mid-range dependencies, it is limited in memorizing capacity. This results in poor representation of long-range dependencies between time steps. By augmenting our model with the dilated RNN architecture, we were able to address this limitation and allow long-range information from previous time steps to contribute to the classification of the most recent frame.
- We performed a thorough evaluation of our proposed model using the largest publicly available football dataset for the research community. We implemented and trained our models using Tensorflow framework. We ran an extensive ablation study. Different components of our network are studied in isolation and together and we compared the results with various baselines and state of the art. This allowed us to analyze the contribution of each component separately.

### **6.3 Research findings and Outcome of the Research**

We believe that given the accuracy improvement achieved using the proposed model in this research, and the fact that the evaluation is done on a dataset of long videos from major European leagues, the outcome of this research has strong academic and industry impacts. As for the academic impact, although our experiments have been performed on football videos, the importance of addressing long-range, mid-range and short-range dependencies is not limited to football events. For any long video with multiple events, it is possible to use a similar architecture for event classification and localization and model the temporal correlations between frames and sub-events. One can simply apply our proposed model on different sports or on other problem domains such as video surveillance.

As for the industry impact, given the quantitative and qualitative results, it is evident that our approach provides a more accurate event localization estimates for long football videos, and as a result could effectively be used in any industry application for semi-automatic event annotation or highlight generation. This means that less human annotation effort is required to provide meta-information for industry applications.

### **6.4 Limitation and Future Discussion**

While our proposed model improves the accuracy of event classification and spotting in football videos, it is still possible to improve the results for small error tolerance thresholds. One key area of improvement is to combine the temporal segmentation process and the event classification into a single end-to-end architecture. This can improve the accuracy of spotting by eliminating the errors occur in the boundaries of the unsupervised temporal segmentation. Another area of focus could be the computational efficiency of the proposed approach. While our neural network inference is fast for each frame, for long videos with high a frame frequency rate, it is still computationally



expensive to run the pre-processing Opticalflow algorithm. One potential exploration would be to compute the Opticalflow using CNN models as well. Multiple recent work compute Opticalflow using a variation of CNN models which can be combined to provide the input for the temporal stream of the two-stream CNN.

Universiti Malaya

## REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Assfalg, J., Bertini, M., Colombo, C., Del Bimbo, A., & Nunziati, W. (2003). Semantic annotation of soccer videos: automatic highlights identification. *Computer vision and image understanding*, 92(2-3), 285-305.
- Assfalg, J., Bertini, M., Del Bimbo, A., Nunziati, W., & Pala, P. (2002). *Soccer highlights detection and recognition using HMMs*. Proceedings. IEEE International Conference on multimedia and expo.
- Atmosukarto, I., Ghanem, B., Saadalla, M., & Ahuja, N. (2014). Recognizing team formation in American football. In *Computer Vision in Sports* (pp. 271-291): Springer.
- Awad, G., Fiscus, J., Joy, D., Michel, M., Smeaton, A., Kraaij, W., . . . Ritter, M. (2016). *Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking*.
- Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2010). *Action classification in soccer videos with long short-term memory recurrent neural networks*. International Conference on Artificial Neural Networks.
- Bengio, Y., Boulanger-Lewandowski, N., & Pascanu, R. (2013). *Advances in optimizing recurrent networks*. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.
- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2), 157-166.
- Brendel, W., Fern, A., & Todorovic, S. (2011). *Probabilistic event logic for interval-based event recognition*. CVPR 2011.
- Buch, S., Escorcia, V., Shen, C., Ghanem, B., & Carlos Niebles, J. (2017). *Sst: Single-stream temporal action proposals*. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.
- Caba Heilbron, F., Carlos Niebles, J., & Ghanem, B. (2016). *Fast temporal activity proposals for efficient detection of human actions in untrimmed videos*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). *Activitynet: A large-scale video benchmark for human activity understanding*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Campos, V., Jou, B., Giró-i-Nieto, X., Torres, J., & Chang, S. F. (2018). Skip RNN: Learning to skip state updates in recurrent neural networks. *In ICLR*.

- Carreira, J., & Zisserman, A. (2017). *Quo vadis, action recognition? a new model and the kinetics dataset*. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Chang, S., Zhang, Y., Han, W., Yu, M., Guo, X., Tan, W., . . . Huang, T. S. (2017). *Dilated recurrent neural networks*. Advances in Neural Information Processing Systems.
- Chaudhry, R., Ravichandran, A., Hager, G., & Vidal, R. (2009). *Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions*. 2009 IEEE Conference on Computer Vision and Pattern Recognition.
- Chen, L., Zhai, M., & Mori, G. (2017). *Attending to Distinctive Moments: Weakly-Supervised Attention Models for Action Localization in Video*. Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on.
- Chen, S., Feng, Z., Lu, Q., Mahasseni, B., Fiez, T., Fern, A., & Todorovic, S. (2014). *Play type recognition in real-world football video*. IEEE Winter Conference on Applications of Computer Vision.
- Chen, S., Fern, A., Mahasseni, B., & Todorovic, S. (2013). *Detecting the Moment of Snap in Real-World Football Videos*. IAAI.
- Chen, S., Fern, A., & Todorovic, S. (2014). *Multi-object tracking via constrained sequential labeling*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Chen, W., Xiong, C., Xu, R., & Corso, J. J. (2014). *Actionness ranking with lattice conditional ordinal random fields*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Chien, S.-Y., Huang, Y.-W., & Chen, L.-G. (2003). Predictive watershed: a fast watershed algorithm for video segmentation. *IEEE Transactions on circuits and systems for video technology*, 13(5), 453-461.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Chung, J., Ahn, S., & Bengio, Y. (2016). Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*.
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Cioppa, A., Deliege, A., Giancola, S., Ghanem, B., Droogenbroeck, M. V., Gade, R., & Moeslund, T. B. (2020). *A context-aware loss function for action spotting in soccer videos*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

- D’Orazio, T., & Leo, M. (2010). A review of vision-based systems for soccer video analysis. *Pattern recognition*, 43(8), 2911-2926.
- Dai, Q., Zhao, R.-W., Wu, Z., Wang, X., Gu, Z., Wu, W., & Jiang, Y.-G. (2015). *Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning*. MediaEval.
- Dai, W., Chen, Y., Huang, C., Gao, M.-k., & Zhang, X. (2019). *Two-Stream Convolution Neural Network with Video-stream for Action Recognition*. 2019 International Joint Conference on Neural Networks (IJCNN).
- Dai, X., Singh, B., Zhang, G., Davis, L. S., & Qiu Chen, Y. (2017). *Temporal context network for activity localization in videos*. Proceedings of the IEEE International Conference on Computer Vision.
- Dalal, N., & Triggs, B. (2005). *Histograms of oriented gradients for human detection*. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05).
- Dalal, N., Triggs, B., & Schmid, C. (2006). *Human detection using oriented histograms of flow and appearance*. European conference on computer vision.
- Deng, J. a. D., W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *CVPR09*.
- Deng, Z., Vahdat, A., Hu, H., & Mori, G. (2016). *Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). *Long-term recurrent convolutional networks for visual recognition and description*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). *Recognizing action at a distance*. null.
- Ekin, A., Tekalp, A. M., & Mehrotra, R. (2003). Automatic soccer video analysis and summarization. *IEEE Transactions on Image processing*, 12(7), 796-807.
- Fan, Y., Lu, X., Li, D., & Liu, Y. (2016). *Video-based emotion recognition using CNN-RNN and C3D hybrid networks*. Proceedings of the 18th ACM International Conference on Multimodal Interaction.
- Feichtenhofer, C., Pinz, A., & Wildes, R. P. (2017). *Spatiotemporal multiplier networks for video action recognition*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). *Convolutional two-stream network fusion for video action recognition*. Proceedings of the IEEE conference on computer vision and pattern recognition.

- Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381-395.
- Gal, Y., & Ghahramani, Z. (2016). *A theoretically grounded application of dropout in recurrent neural networks*. Advances in neural information processing systems.
- Gao, J., Yang, Z., Chen, K., Sun, C., & Nevatia, R. (2017). *Turn tap: Temporal unit regression network for temporal action proposals*. Proceedings of the IEEE International Conference on Computer Vision.
- Gavrilyuk, K., Ghodrati, A., Li, Z., & Snoek, C. G. (2018). *Actor and action video segmentation from a sentence*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Giancola, S., Amine, M., Dghaily, T., & Ghanem, B. (2018). *Soccernet: A scalable dataset for action spotting in soccer videos*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.
- Glorot, X., & Bengio, Y. (2010). *Understanding the difficulty of training deep feedforward neural networks*. Proceedings of the thirteenth international conference on artificial intelligence and statistics.
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action.
- Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). *Hybrid speech recognition with deep bidirectional LSTM*. 2013 IEEE workshop on automatic speech recognition and understanding.
- Graves, A., Mohamed, A.-r., & Hinton, G. (2013). *Speech recognition with deep recurrent neural networks*. 2013 IEEE international conference on acoustics, speech and signal processing.
- Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., . . . Sukthankar, R. (2018). *Ava: A video dataset of spatio-temporally localized atomic visual actions*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- Hong, Y., Ling, C., & Ye, Z. (2018). *End-to-end soccer video scene and event classification with deep transfer learning*. 2018 International Conference on Intelligent Systems and Computer Vision (ISCV).
- Huang, C.-L., Shih, H.-C., & Chao, C.-Y. (2006). Semantic analysis of soccer video using dynamic Bayesian network. *IEEE Transactions on Multimedia*, 8(4), 749-760.

- Huang, C.-P., Hsieh, C.-H., Lai, K.-T., & Huang, W.-Y. (2011). *Human action recognition using histogram of oriented gradient of motion history image*. 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control.
- Huang, Y., Llach, J., & Bhagavathy, S. (2007). *Players and ball detection in soccer videos based on color segmentation and shape analysis*. International Workshop on Multimedia Content Analysis and Mining.
- Huang, Y., Yang, H., & Huang, P. (2012). *Action recognition using hog feature in different resolution video sequences*. 2012 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring.
- Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., & Mori, G. (2016). *A hierarchical deep temporal model for group activity recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 221-231.
- Jiang, H., Lu, Y., & Xue, J. (2016). *Automatic soccer video event detection based on a deep neural network combined cnn and rnn*. 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI).
- Kalchbrenner, N., & Blunsom, P. (2013). *Recurrent continuous translation models*. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). *Large-scale video classification with convolutional neural networks*. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.
- Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H., Eskofier, B. M. J. D. M., & Discovery, K. (2017). Activity recognition in beach volleyball using a Deep Convolutional Neural Network. *31(6)*, 1678-1705.
- Klaser, A., Marszałek, M., & Schmid, C. (2008). *A spatio-temporal descriptor based on 3d-gradients*.
- Kolekar, M. H., & Sengupta, S. (2015). Bayesian network-based customized highlight generation for broadcast soccer videos. *IEEE Transactions on Broadcasting*, 61(2), 195-209.
- Koutnik, J., Greff, K., Gomez, F., & Schmidhuber, J. (2014). *A clockwork rnn*. International Conference on Machine Learning.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. Advances in neural information processing systems.

- Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., . . . Wiskott, L. (2012). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1847-1871.
- Kumar, S. S., & John, M. (2016). *Human activity recognition using optical flow based feature set*. 2016 IEEE international Carnahan conference on security technology (ICCST).
- Lan, T., Sigal, L., & Mori, G. (2012a). *Social roles in hierarchical models for human activity recognition*. 2012 IEEE Conference on Computer Vision and Pattern Recognition.
- Lan, T., Sigal, L., & Mori, G. (2012b). *Social roles in hierarchical models for human activity recognition*. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- Lertniphonphan, K., Aramvith, S., & Chalidabhongse, T. H. (2011). *Human action recognition using direction histograms of optical flow*. 2011 11th International Symposium on Communications & Information Technologies (ISCIT).
- Li, C., Wu, X., Zhao, N., Cao, X., & Tang, J. (2018). Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing*, 281, 78-85.
- Li, X. (2007). HMM based action recognition using oriented histograms of optical flow field. *Electronics Letters*, 43(10), 560-561.
- Liu, P., Wang, J., She, M., & Liu, H. (2011). *Human action recognition based on 3D SIFT and LDA model*. 2011 IEEE Workshop on Robotic Intelligence In Informationally Structured Space.
- Liu, T., Lu, Y., Lei, X., Zhang, L., Wang, H., Huang, W., & Wang, Z. (2017). *Soccer video event detection using 3d convolutional networks and shot boundary detection via deep feature distance*. International Conference on Neural Information Processing.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91-110.
- Lu, C., Shi, J., & Jia, J. (2013). *Abnormal event detection at 150 fps in matlab*. Proceedings of the IEEE international conference on computer vision.
- Ma, Z., Yang, Y., Xu, Z., Yan, S., Sebe, N., & Hauptmann, A. G. (2013). *Complex event detection via multi-source video attributes*. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.
- Maheswari, S. U., & Ramakrishnan, R. (2015). Sports video classification using multi scale framework and nearest neighbor classifier. *indian Journal of Science and Technology*, 8(6), 529.

- Martens, J., & Sutskever, I. (2011). *Learning recurrent neural networks with hessian-free optimization*. Proceedings of the 28th international conference on machine learning (ICML-11).
- Mleya, M. V., Li, W., Liang, J., Liu, K., Sun, Y., Jin, G., & Wang, J. (2019). *Online Aggregated-Event Representation for Multiple Event Detection in Videos*. International Conference on Advanced Data Mining and Applications.
- Neil, D., Pfeiffer, M., & Liu, S.-C. (2016). Phased lstm: Accelerating recurrent network training for long or event-based sequences. *arXiv preprint arXiv:1610.09513*.
- Nguyen, T.-N., & Meunier, J. (2019). *Anomaly detection in video sequence with appearance-motion correspondence*. Proceedings of the IEEE/CVF International Conference on Computer Vision.
- Oreifej, O., & Liu, Z. (2013). *Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Pallavi, V., Mukherjee, J., Majumdar, A. K., & Sural, S. (2008a). Ball detection from broadcast soccer videos using static and dynamic features. *Journal of Visual Communication and Image Representation*, 19(7), 426-436.
- Pallavi, V., Mukherjee, J., Majumdar, A. K., & Sural, S. (2008b). Graph-based multiplayer detection and tracking in broadcast soccer videos. *IEEE Transactions on Multimedia*, 10(5), 794-805.
- Pascanu, R., Mikolov, T., & Bengio, Y. (2013). *On the difficulty of training recurrent neural networks*. International Conference on Machine Learning.
- Qian, X., Hou, X., Tang, Y., Wang, H., & Li, Z. (2012). Hidden conditional random field-based soccer video events detection. *IET Image Processing*, 6(9), 1338-1347.
- Qian, X., Liu, G., Wang, Z., Li, Z., & Wang, H. (2010). *Highlight events detection in soccer video using HCRF*. Proceedings of the Second International Conference on Internet Multimedia Computing and Service.
- Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research*, 44(11), e107-e107.
- Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., & Fei-Fei, L. (2016). *Detecting events and key actors in multi-person videos*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Rapantzikos, K., Avrithis, Y., & Kollias, S. (2009). *Dense saliency-based spatiotemporal feature points for action recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition.
- Raptis, M., & Sigal, L. (2013). *Poselet key-framing: A model for human activity recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.



- Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.
- Scovanner, P., Ali, S., & Shah, M. (2007). *A 3-dimensional sift descriptor and its application to action recognition*. Proceedings of the 15th ACM international conference on Multimedia.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization*. Proceedings of the IEEE international conference on computer vision.
- Serpush, F., & Rezaei, M. (2020). Complex Human Action Recognition in Live Videos Using Hybrid FR-DL Method. *arXiv preprint arXiv:2007.02811*.
- Shou, Z., Wang, D., & Chang, S.-F. (2016). *Temporal action localization in untrimmed videos via multi-stage cnns*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Shri, S. J., & Jothilakshmi, S. (2018). *Video Analysis for Crowd and Traffic Management*. 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA).
- Sigurdsson, G. A., Russakovsky, O., & Gupta, A. (2017). *What actions are needed for understanding human actions in videos?* Proceedings of the IEEE International Conference on Computer Vision.
- Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). *Hollywood in homes: Crowdsourcing data collection for activity understanding*. European Conference on Computer Vision.
- Simonyan, K., & Zisserman, A. (2014). *Two-stream convolutional networks for action recognition in videos*. Advances in neural information processing systems.
- Simonyan, K., & Zisserman, A. J. a. p. a. (2014). Very deep convolutional networks for large-scale image recognition.
- Soomro, K., Idrees, H., & Shah, M. (2018). Online localization and prediction of actions and interactions. *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 459-472.
- Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.
- Sultani, W., Chen, C., & Shah, M. (2018). *Real-world anomaly detection in surveillance videos*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Sun, C., & Nevatia, R. (2013). *Active: Activity concept transitions in video event classification*. Proceedings of the IEEE International Conference on Computer Vision.

- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). *Going deeper with convolutions*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Tang, D., Qin, B., & Liu, T. (2015). *Document modeling with gated recurrent neural network for sentiment classification*. Proceedings of the 2015 conference on empirical methods in natural language processing.
- Tang, K., Fei-Fei, L., & Koller, D. (2012). *Learning latent temporal structure for complex event detection*. 2012 IEEE Conference on Computer Vision and Pattern Recognition.
- Tavassolipour, M., Karimian, M., & Kasaei, S. (2014). Event detection and summarization in soccer videos using bayesian network and copula. *IEEE Transactions on circuits and systems for video technology*, 24(2), 291-304.
- Thi, T. H., Zhang, J., Cheng, L., Wang, L., & Satoh, S. (2010). *Human action recognition and localization in video using structured learning of local space-time features*. 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance.
- Todorovic, S., & Mahasseni, B. (2013). *Latent multitask learning for view-invariant action recognition*. Proceedings of the IEEE International Conference on Computer Vision.
- Todorovic, S., & Mahasseni, B. (2016). *Regularizing long short term memory with 3D human-skeleton sequences for action recognition*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Tovinkere, V., & Qian, R. J. (2001). *Detecting semantic events in soccer games: Towards a complete solution*. null.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). *Learning spatiotemporal features with 3d convolutional networks*. Proceedings of the IEEE international conference on computer vision.
- Tran, D., Wang, H., Torresani, L., & Feiszli, M. (2019). *Video classification with channel-separated convolutional networks*. Proceedings of the IEEE/CVF International Conference on Computer Vision.
- Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). *A closer look at spatiotemporal convolutions for action recognition*. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.
- Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access*, 6, 1155-1166.
- Vats, K., Fani, M., Walters, P., Clausi, D. A., & Zelek, J. (2020). *Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

- Wang, C.-h., Wang, Y., & Guan, L. (2011). *Event detection and recognition using histogram of oriented gradients and hidden markov models*. International Conference Image Analysis and Recognition.
- Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). *Action recognition by dense trajectories*. CVPR 2011.
- Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1), 60-79.
- Wang, J., Liu, P., She, M., & Liu, H. (2011). *Human action categorization using conditional random field*. 2011 IEEE Workshop on Robotic Intelligence In Informationally Structured Space.
- Wang, J., Xu, C., Chng, E., Wah, K., & Tian, Q. (2004). *Automatic replay generation for soccer video broadcasting*. Proceedings of the 12th annual ACM international conference on Multimedia.
- Wang, L., Xiong, Y., Wang, Z., & Qiao, Y. (2015). Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*.
- Wang, L., Zhou, F., Li, Z., Zuo, W., & Tan, H. (2018). *Abnormal event detection in videos using hybrid spatio-temporal autoencoder*. 2018 25th IEEE International Conference on Image Processing (ICIP).
- Wang, T., Li, J., Diao, Q., Hu, W., Zhang, Y., & Dulong, C. (2006). *Semantic event detection using conditional random fields*. 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06).
- Wang, W., Bao, F., & Gao, G. (2016). *Mongolian named entity recognition with bidirectional recurrent neural networks*. 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI).
- Wang, Z., Yu, J., & He, Y. (2016). Soccer video event annotation by synchronization of attack–defense clips and match reports with coarse-grained time information. *IEEE Transactions on circuits and systems for video technology*, 27(5), 1104-1117.
- Wickramaratna, K., Chen, M., Chen, S.-C., & Shyu, M.-L. (2005). *Neural network based framework for goal event detection in soccer videos*. Seventh IEEE International Symposium on Multimedia (ISM'05).
- Wu, J., Hu, C., Wang, Y., Hu, X., & Zhu, J. (2019). A hierarchical recurrent neural network for symbolic melody generation. *IEEE Transactions on Cybernetics*, 50(6), 2749-2757.
- Wu, P., Liu, J., & Shen, F. (2019). A deep one-class neural network for anomalous event detection in complex scenes. *IEEE transactions on neural networks and learning systems*, 31(7), 2609-2622.

- Xie, L., Chang, S.-F., Divakaran, A., & Sun, H. (2002). *Structure analysis of soccer video with hidden Markov models*. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing.
- Xu, H., Das, A., & Saenko, K. (2017). *R-c3d: Region convolutional 3d network for temporal activity detection*. Proceedings of the IEEE international conference on computer vision.
- Xu, H., Das, A., & Saenko, K. (2019). Two-stream region convolutional 3d network for temporal activity detection. *IEEE transactions on pattern analysis and machine intelligence*, 41(10), 2319-2332.
- Xu, P., Xie, L., Chang, S.-F., Divakaran, A., Vetro, A., & Sun, H. (2001). *Algorithms And System For Segmentation And Structure Analysis In Soccer Video*. ICME.
- Y.-G. Jiang, J. L., A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah,, & Sukthankar, a. R. (2014). THUMOS challenge: Action recognition with a large number of classes. <http://crcv.ucf.edu/THUMOS14/>.
- Yan, S., Smith, J. S., Lu, W., & Zhang, B. (2018). Abnormal event detection from videos using a two-stream recurrent variational autoencoder. *IEEE Transactions on Cognitive and Developmental Systems*, 12(1), 30-42.
- Yang, X., & Tian, Y. L. (2012). *Eigenjoints-based action recognition using naive-bayes-nearest-neighbor*. 2012 IEEE computer society conference on computer vision and pattern recognition workshops.
- Yang, X., Zhang, C., & Tian, Y. (2012). *Recognizing actions using depth motion maps-based histograms of oriented gradients*. Proceedings of the 20th ACM international conference on Multimedia.
- Yang, Y., Lin, S., Zhang, Y., & Tang, S. (2007). *Highlights extraction in soccer videos based on goal-mouth detection*. Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on.
- Yeung, S., Russakovsky, O., Mori, G., & Fei-Fei, L. (2016). *End-to-end learning of action detection from frame glimpses in videos*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Yu, W., Yang, K., Bai, Y., Xiao, T., Yao, H., & Rui, Y. (2016). *Visualizing and comparing AlexNet and VGG using deconvolutional layers*. Proceedings of the 33rd International Conference on Machine Learning.
- Zawbaa, H. M., El-Bendary, N., Hassanien, A. E., & Abraham, A. (2011). *SVM-based soccer video summarization system*. 2011 Third World Congress on Nature and Biologically Inspired Computing.
- Zhan, Y., Liu, J., Gou, J., & Wang, M. (2016). A video semantic detection method based on locality-sensitive discriminant sparse representation and weighted KNN. *Journal of Visual Communication and Image Representation*, 41, 65-73.

- Zhang, J.-T., Tsoi, A.-C., & Lo, S.-L. (2014). *Scale invariant feature transform flow trajectory approach with applications to human action recognition*. 2014 International Joint Conference on Neural Networks (IJCNN).
- Zhang, Y., Chen, G., Yu, D., Yaco, K., Khudanpur, S., & Glass, J. (2016). *Highway long short-term memory rnn for distant speech recognition*. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- Zhao, R., Ali, H., & Van der Smagt, P. (2017). *Two-stream RNN/CNN for action recognition in 3D videos*. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).
- Zhou, X., Zhuang, X., Yan, S., Chang, S.-F., Hasegawa-Johnson, M., & Huang, T. S. (2008). *Sift-bag kernel for video event analysis*. Proceedings of the 16th ACM international conference on Multimedia.
- Zhu, M. (2004). Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, 2, 30*.