COMPUTER VISION BASED AUTOMATED PLAYER TRACKING IN RUGBY

Galkissage Manik Tharaka Fernando

(168221X)

Master of Science in Computer Science

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

February 2020

COMPUTER VISION BASED AUTOMATED PLAYER TRACKING IN RUGBY

by

Galkissage Manik Tharaka Fernando

(168221X)

Dissertation submitted in partial fulfillment of the requirements for the Master of Science in Computer Science

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

February 2020

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Master of Science thesis under my supervision.

Name of Supervisor: Dr.Amal Shehan Perera

Signature: Date:

COMPUTER VISION BASED AUTOMATED PLAYER TRACKING IN RUGBY

by

Galkissage Manik Tharaka Fernando

Abstract

Sports related analytics have become a main component of the present professional sporting domain. Teams continuously rely on the knowledge provided by analytics systems to gain a competitive edge over the opposing team. One of the main aspects of sports analytics is automated player tracking which can be achieved by computer vision based techniques by analyzing video footage of sporting events. Multiple object tracking in itself is a non trivial problem due to the large number of variables involved. This is further amplified by the high number of occlusions, trajectory changes that occur in a highly physical sport such as Rugby. We set out to solve the problem of automated player tracking using a tracking by detection approach. We make use of an object localisation model named YOLO and retrain it to suit the specific scenarios in Rugby. In order to solve the data association problem we compute an appearance based metric using an identity embedding encoder network. A Kalman filter is used along with the appearance based metric to establish the associations between tracks and detections. We conduct several experiments to evaluate the implemented solution and report the results. We discuss the limitations,further improvements and areas that present further research opportunities.

Thesis Supervisor: Dr.Amal Shehan Perera
Title: Senior Lecturer

Acknowledgments

I express my heartfelt gratitude to my supervisor, Dr. Amal Shehan Perera for the guidance, encouragement and mentorship provided. I would like to acknowledge and convey my appreciation towards members of my family for the continuous support and encouragement given to me throughout the course of this thesis. I would like to thank my wife Madhavi Andradi and my brother Deegha Galkissa for the support provided in annotating the data. Last but not least I thank my friends for the support they have extended to me in numerous ways.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Automated player tracking

Automatic player tracking is an area which has many uses in the modern sports science and professional and semi professional domains. From a sports scientific perspective tracking helps identify aspects such as demands on energy expenditure [1]. In the professional sporting domain which has become a multi-million dollar industry, analytics play a major role in the outcome of a sporting event [2, 3, 4]. Consequently the performance of the sporting team or franchise may be significantly affected by the leverage provided by the analytics system employed from the phase of player drafting, training to on-field real time analytics. Most of the decisions made traditionally through experience of management and coaching staff are being supplemented by analytics. Data such as player positioning, meters gained, speed and trajectories can be valuable with regard to analytics applications. For instance previous season statistics can be used when drafting and recruiting new players and at the training phase the data can be used to build a profile of each player, and at game day the real time data can be used to guide the strategy and the tactics used in the game.There are many technologies that may be employed to track players in sporting events or in training and collect the data. These technologies may range from global positioning systems (GPS), radio frequency tags, magnetic sensor tracking, ultrasound, video analysis based tracking. The most widely used technologies from the above are GPS and computer vision based automated video tracking.

GPS has the advantage of being accurate and the ability to provide data in near real time. However these approaches require the players to wear a specialised tracking device. In the case of GPS the device receives signals from a constellation of earth orbiting satellites, the location can be determined by using data from 3 of the satellites. Computer vision based tracking methods have made significant progress in recent times. They can provide significant accuracy but are still not superior to the performance of GPS tracking[1].

1.2 Problem and opportunity for research

With regard to automated player tracking previous studies have been conducted employing various technologies. However a large number of studies are conducted on non contact or semi contact sports such as Basketball or Soccer , and we find only a handful of research on full contact sports such similar to Rugby.

Rugby is a highly physical full contact game which generally requires strength, stamina, speed and agility. It is a dynamic game which may consist of bouts of high intensity activity followed by low intensity aerobic activity and rest[5].Individual rugby players are usually specialized in a certain skill set that is demanded of the position they play, for example Props need to be strong and sturdy and usually of heavy physical stature which is in contrast to a player playing wing who needs to be fast and agile. This characteristic of rugby demands a specialized training regimen that can accommodate the required fitness capacity of players at an individual level. From a physiological viewpoint knowledge gained from analyzing player patterns both in training and game time can be used in designing an enhanced training program.

Furthermore in a game of rugby there are unique playing aspects such as scrums, rucks and mauls where multiple players may physically engage with each other. From a technical point of view positional and trajectory data along with information about the various playing scenarios can be used to gain knowledge and to plan out the strategy and tactics of the game. Identifying these playing situations is important for the purpose of analytics and poses a unique challenge for computer vision applications due to the multiple occlusions that occur at these situations.

The study conducted by Edgecomb and Norton[1] Illustrates the difficulties of applying some of the aforementioned technologies to Australian Football. In highly physical contact sports such as rugby, and Australian Football, using tracking technologies which requires the players to wear specialized devices is not practical in most situations due to risk of injury. Most of the players who take part in these sports wear little to no protective gear, therefore the places to safely embed these devices are minimal. Using a tracking technology which requires players to wear specialized equipment has the additional disadvantage of not being able to track the players of the opposition.

Considering the above mentioned factors it is evident that although GPS and other device based technologies are generally considered superior for the purpose of automated player tracking, there still are some limitations with regard to these aproaches and they may not be suitable for every situation. Computer vision based tracking provide a non invasive approach to tracking eliminating the need for specialised devices to be worn by players. However due to the complexity associated with CV based player tracking there are still areas that can be improved from further research. Thus CV based approaches provide ample opportunity for further research while at the same time being advanced enough to permit practical applications. Therefore the focus of this research will be to develop a computer vision based automated player tracking system which can be employed in analytics for rugby matches.

1.3 Motivation to automate player tracking for Rugby

The scarcity of research conducted on applying automated player tracking for rugby and the lack of a specialized system that can identify various playing aspects of rugby is the main motivating factor for this study. Furthermore the usage of computer vision based techniques for player tracking still has challenges that needs solutions. Solving the interesting problems faced when applying computer vision based player tracking, such as tracking and identifying moving entities can be applied to other domains such as surveillance or autonomous vehicle navigation. In recent times professional sports has developed into a multi million dollar industry. With these developments managing bodies of professional teams are constantly looking out for ways to minimise mistakes and to gain a competitive advantage over their opponents. One area that has caught the attention of professional sporting teams is data driven decision making. To this end data science and analytics can provide value additions in a large range of applications ranging from scouting/drafting of players, team selection to on field decision making. This high demand for data analytics and tools that enable data analytics in professional sports also acts as further motivation for this study.

1.4 Objectives

The purpose of this research is to develop an automated system that can collect data related to players positioning and trajectories. This research focuses on computer vision based automated player tracking, which uses computer vision to extract player positioning data during the timeline of a game from video footage. The study will be conducted on video footage of Rugby Union matches. The main objectives can be listed as follows.

1. Detect players on the field

2. Track player movement trajectories and positioning across time

Chapter 2

LITERATURE REVIEW

Automated player tracking is a domain where a considerable amount of research has been conducted in recent years. The focus of this chapter is to have a discussion on literature and past studies that are concerned with computer vision based tracking.

Object tracking can be defined as estimating the trajectory of an object of interest as it moves within a scene. This can be achieved by assigning consistent labels to the object within each frame of a video [6]. Depending on the approach that is employed we can classify the literature concerning object tracking into many different classes. One of the most frequently occuring classifications is the tracking-by-detection paradigm versus tracking-by-estimation paradigm. In the tracking by detection approach,the problem of object tracking is broken down to two subproblems of object detection and object tracking [7, 8, 9, 10]. The object detection component is concerned with continuously detecting the objects of interest in the input and the tracking component involves estimating and associating the detected trajectories across the frames.

The techniques and the approach used for the purpose of computer vision based tracking is reliant on the specific domain that the problem is applied to [6]. Therefore most of the literature discussed will be related to the area of multi-target tracking applied to various sports along with studies conducted in generic domains. First the general literature related to the problem of automated player tracking will be discussed followed by discussion on player detection and player tracking.

2.1 Automated multi target tracking

The comparison done on automated surveillance systems and automated player tracking systems by Barris and button [11] helps us to put the challenges faced with automated player tracking into perspective. According to the authors the technologies used in surveillance have seen some changes during the recent past. Attempts at capturing human motion has led to articulate 3D models from the previous 2D models. Tracking of objects has seen a shift towards sampling based techniques from deterministic linear tracking frameworks. It is emphasized that The improvements in motion analysis is contributed by the usage of machine learning approaches and that this has led to systems that can cope with outdoor scenes and situations where there are multiple occluding people. Authors also state that these improvements are attributed towards advancements in segmentation technologies while some studies credit the improvements in model based pose estimation.

The main differences that are seen between surveillance systems and sports tracking systems is that the targets being tracked in a sporting event may be fast moving and might exhibit erratic unpredictable movements wheres in a pedestrian surveillance setting the targets may be slow moving with smoother trajectories. The review on sports tracking in both indoor and outdoor settings illustrate the technological trends that are evident since recent times, some of these studies will be discussed in depth.The authors have observed that scene event analysis using image processing has become increasingly common. The review also discusses limitations that are commonly faced by numerous studies. Most of the studies have conducted the research with stationary cameras due to the added complexity with the use of a moving camera. Player occlusion seems to be another major hurdle when applying computer vision based algorithms. When a player blocks another player visually in the field of view of the camera it is known as a player occlusion. Situations where two players occlude or where multiple players cluster around a certain point have yielded low tracking accuracy.

2.2 Player detection

Some of the research conducted with regard to automated multi target tracking involves first detecting the object and associating a tracker with the detected object [7, 8] and others have incorporated an object detection component as part of the overall tracking mechanism [9, 10]. In this section the various approaches made with regard to player detection will be discussed.

Within the research literature a varying number of features are used for the task of object detection. Significant success has been achieved when they are incorporated with a machine learning approach. The research conducted buy Dalal and Triggs [12] proposes the usage of Histogram of Oriented Gradients as the features from an image as input features. The HOG features are used in a linear SVM in order to detect the pedestrians. The HOG features share similarities with orientation histograms, SIFT descriptors and shape contexts. The HOG features are computed on a dense grid of uniformly spaced cells where's the others are sparse descriptors. The local object appearance and shape can be characterized with the distribution of local intensity gradients and edge directions. The authors make a note to normalize the calculated features before using them in detection. The detection window is tiled with a dense overlapping grid of HOG feature descriptors. The HOG representation is good at capturing edges in the local vicinity as it takes into consideration the gradients in the local cells and translations and rotations makes little difference if they are smaller than the local spacial orientation bin size. The experiments were conducted on the MIT pedestrian database[13] as well as a data set that was compiled by the authors themselves with 1805 images. the methodology followed involved training a set of preliminary detectors and searching a set of 1218 negative image samples for false positives. The training set is supplemented with any false positives found in the data set and the SVM is retrained. The retraining process helped to increase the performance by 5%.

The framework proposed by Zhu et al. [7] uses support vector classification and segmentation of the play field to detect players on the field. In order to identify the pixels of the playing field Gaussian Mixture Models (GMM) were used. Once the general playing field pixels are identified, region growing algorithm is used to connect the pixels into areas and to refine the edges. In order to identify the players from other objects in the frame a Support Vector Machine model in the form of a classifier is used. Training examples were manually extracted from footage, images of players were selected as positive examples and parts of the playing field was used as negative examples.The features used in the classification include a color model that is built from the Hue-Saturation-Value color space using a histogram.

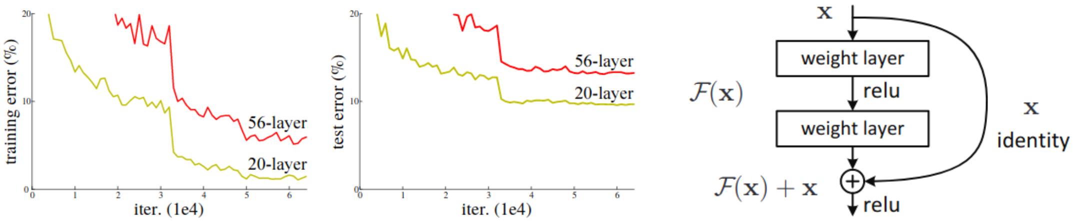
In a similar study conducted by Liu et al. [8] on broadcast soccer videos apart from player detection player labeling was also achieved. Initially the dominant color is identified and the playing field is extracted with the use of dominant color segmentation, morphological filtering and connect component analysis. The size of the play field and non-playing field may vary due to changing camera angles and shots. Therefore a decision tree was used to classify each view into global, medium, close-up and out of view. In order to identify players, a boosted cascade of Haar features based on the work by Viola and Jones[14] were used. A set of player images that capture the different motion variations as much as possible was used as positive training samples along with the images of various markers and signs on the playing field as negative examples. The background playing field was filtered out of the training examples which has made the process faster. The detector scans over the image at varying sizes, which sometimes resulted in multiple detection instances around a player. These duplicate detection instances were merged together to obtain a single detection.The experiments were done on the 2006 soccer world cup videos. Frames were randomly selected from the footage with about 50 to 100 frames difference between selected frames and the ground truth was manually labeled. The videos that were used were from the 2006 soccer world cup from matches between France and Spain and between Brazil and Japan.The authors were able to achieve significant figures precision (88.65%,92.38%) and recall (92.19%,88.82%) values as well as high F-score values

(90.39%,90.57%).

In the widely cited study conducted by Okuma et al.[9] a cascaded Adaboost process is used for detection of players [14]. A training set of 6000 figures of hockey players scaled to a standard size of 10 X 24 pixels is used for the purpose of training the cascaded classifier. There is a considerable difference between the intensities of the pixels that represent a player and the hockey rink. The researchers have used this property to device a simple extraction method to extract player images for training, where small low intensity areas(players) which are surrounded by high intensity areas(hockey rink) are detected and extracted. However the authors point out that this approach is not the ideal one for accuracy, as the trained Adaboost estimator had a number of false positives around the edge of the rink. The false positives of the spectators can be corrected from the classification if plausible motions of the hockey players are taken into consideration. In order to achieve this the Adaboost classifier was incorporated to the proposal of the mixture of particle filter. In the final mixture model the Adaboost algorithm performed well in detecting players. However there are some weaknesses of Adaboost such as susceptibility to variations in intensity and occlusion.

2.2.1 Deep learning based object recognition

In recent times a surge in the success rate of deep neural network based object classification algorithms can be observed. Researchers have been able to achieve very high accuracy levels in a very short time span. The model named Alexnet [15] gave the breakthrough results in the 2012 imagenet competition [16]. The model consisted of five convolutional layers and three fully connected layers. An interesting contribution from the Alexnet model was that by using ReLu (Rectified Linear unit) activation functions deep CNNs can be trained much faster with gradient descent, as opposed to the tanh and sigmoid activation functions that were usually used. The researchers also used dropout in the first two fully connected layers to reduce overfitting. They were able to achieve a top-5 error rate of 15.3% at the ILSVRC-2012 competition. The VGG network [17] was



(a) Number of layers vs error rate

(b) Resnet building block

Figure 2-1: (a) Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training and test errors.(b) Residual unit with skip connection provide identity mapping from the output of the previous layer.Source [18]

among the best performing models in the imagenet competition. The model used convolutional layers with 3X3 filters as opposed to large receptive fields used in previous models. The network had two models of 16 layers and 19 layers. Due to difficulty in training the networks with large depths, the authors trained smaller models with fewer layers and used these smaller models as an initialisation step to train the larger models.

The ResNet architecture introduced by He et al. [18] is another revolutionary approach. The authors observe that the recent success in image recognition models can be attributed to the large depth of the networks, which results in a larger parameter space and more complex functions. However naively increasing the number of a neural network does not result in better performance. He et al. illustrate this by plotting the training and test error of a 20 layer neural network and a 50 layer neural network, where the 50 layer network performs worse than the shallower 20 layer network. They identify that this hindrance does not occur due to overfitting either and observe that as the network depth increase the the accuracy becomes saturated and rapidly degrades. The authors introduce residual learning comprised of residual blocks to overcome this *degradation* problem. These blocks use skip connections to obtain identity mapping from the output of the previous layer which is added to the output of the next layer.

2.2.2 Object localisation and detection

These achievements in object classification paved for further advances in object detection and localisation tasks where the objective is to identify the location and bounds of the object of interest. In classical computer vision approaches object localisation was mainly achieved by sliding window techniques, where a window encompassing a subset of the image pixels slides across full image. The pixel information is input to a detector model to identify if the window bounding box contains an object of interest. This approach is quite cumbersome, which provided an opportunity for further research. One of the prominent approaches that set out to solve the problem of object localisation is the Regions with CNN (R-CNN) by Girshick et al. [19]. The authors use selective search algorithm to propose approximately 2000 regions of interest. These regions are extracted from the image.The extracted image patches are warped to the dimensions 227x227 and input to a CNN which outputs a 4096 dimensional feature vector. A set of class specific SVM models are used to classify whether there is an object of interest in the selected region and and a regression model is used to refine the bounding box dimensions. In this approach the CNN acts a s rich feature extractor and the actual classification was done by the SVM. They were able to achieve a mean average precision (mAP) of 53.7% on the PASCAL VOC 2010 dataset [20]. Despite its good performance in terms of accuracy R-CNN had several drawbacks. Object detection at test time was slow, the training process was a multi stage pipeline with several componentes and the training process was computationally expensive. In a later study the authors present Fast R-CNN [21] as an improved approach over R-CNN. In the R-CNN model each image proposal patch is fed forward through a deep convolutional neural network, this method results in a performance bottleneck. If there were N proposals, the CNN feedforwad computation has to occur N times. The new model solves this problem by first feeding the full image through the CNN and creating a convolutional feature map. Then for each proposal of interest a layer named RoI pooling extracts the features and resizes them to a fixed aspect ratio. The Fast

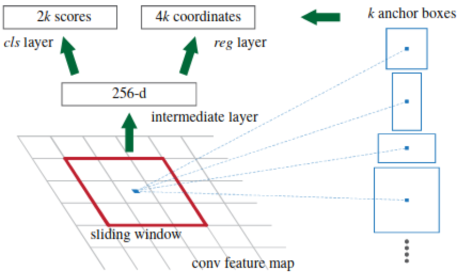


Figure 2-2: Region proposal done by the Region Proposal Network.Source [22]

R-CNN makes another change in terns of the output ,where R-CNN relied on class specific SVM estimators and regression models, the Fast R-CNN has two sibling output layers. One layer outputs a vector of probabilities for the classes the network is trained on, and the other layer outputs the bounding box regression offsets. With these improvements the authors were able to obtain speed increase of 10x to 100x at test time. There performance on the VOC 2007 dataset is reported at mAP 70% , on VOC 2010 at mAP 68.8% and on VOC 2012 mAP 68.4%.

Another research that continued to make more improvements over the Fast R-CNN is the aptly named Faster R-CNN [22] model by Ren et al. Despite having a similar name to the previous discussed methods Faster R-CNN makes some important changes. Instead of using selective search for region proposals, the model uses a region proposal network which slides a window over the convolutional feature map. The sliding windows are mapped to a lower dimensional feature which in turn becomes the input to the final output layers. At each sliding window center multiple region proposals named anchors are predicted. Each region proposal consists of four bounding box coordinate values and two scores that estimate the probability if the region contains an object 2-2. In this study 3 aspect ratio variations and 3 scale variations amounting to 9 anchor boxes are predicted by each sliding window. The Faster R-CNN model exhibits improved performance with 73.2% mAP for the PASCAL VOC 2007 dataset and 70.4% mAP on the VOC 2012 test set.

The work done by Redmon et al. resulted in an object detection and localisation model dubbed You Only Look Once (YOLO) [23]. This model is capable in handling the end to end object localisation process in a single network, taking in the full image data as input only once, which is in contrast to the region proposal and sliding window approaches discussed above. In other words the model looks at the image only once, which was the inspiration for the YOLO model name. The model divided the image into a grid of NxN cells. Each grid cell predicts B bounding boxes, each of which consists the coordinates for the bounding box and a object confidence score. Each grid cell predicts a set of C class conditional probabilities as well.Only one set of class probabilities are predicted per grid cell, irrespective of how many bounding boxes the grid is predicting. The YOLO model was capable of obtaining a high FPS performance of 45 frames per second with an NVIDIA TITAN X GPU and a faster model was capable of performing at 150 FPS.

2.3 Player tracking

As stated by Hue et al. [24] the main challenge of the problem comes down to the assignment of the measurement observations to the target. Tracking in the sense of object tracking involves the state estimation of an unknown object/objects. The problem involves the identification and recursive localisation of an object from sequential data that is available to the observer [25] . The problem can be considered as having two aspects, namely the data association and estimation which needs to be considered together.

2.3.1 Recursive bayesian estimation based player tracking

In the research conducted by Zhu et al.[7] a particle filter which uses Support Vector Regression is used for the tracking of players which are detected from the object detector discussed above. Particle filters need a large sample set in order to be accurate. Since the computational costs increase with the size of the sample set, the authors have employed the sample re-weighting approach that was proposed by Vapnik [26] and have developed the Support Vector Regression particle filter. This approach strives to minimize the noise of the posterior distribution by combining support vector regression into Sequential Monte Carlo algorithm.The experimental data includes broadcast videos from sports such as soccer,hockey,basketball and American Football. From 4 videos 13 clips were extracted which summed up to a set of 3599 frames. The accuracy of the test results were consistently high in the ranges above 80% accuracy for all tests. Some of the limitations stated by the authors is that when the player pixel area is very small which could happen due to the camera perspective, players will be dismissed because of wrongly identifying them as noise. As future work the authors plan on solving the problems posed by occlusion.

In a similar study conducted by Liu et al. [8] on broadcast soccer videos the problem of tracking players is formulated as a data association problem. The players are detected and labeled as mentioned previously, and these results are used as input to the tracking problem. The graph structure is defined over the observational data where each node represents single observation and the edges represent a relationship between neighboring observations. In order to find the optimal solution Markov Chain Monte Carlo (MCMC) based strategy is used. As mentioned earlier experiments were conducted on video from the 2006 soccer world cup matches. From the match between Spain and France 100 consecutive frames were extracted and manually labeled with the ground truth. The system was able to accurately detect and track players even under occlusion. But authors have not mentioned how long the occlusion had occurred. They were able to obtain a precision of 99.32% precision and 94.43% recall against the ground truth.

In the study conducted by Wu et. al [27] they employ Relative Discriminative Histogram of Oriented Gradients (RDHOG) based Particle Filter tracker to vehicles and to resolve situations when partial occlusion occur. The RDHOG descriptor is an extension to the HOG descriptor. The difference between the two can be seen when considering the blocks of cells. The RDHOG takes into consideration the relationship between the central block and the surrounding blocks. Background subtraction is performed on the image to identify foreground pixels and RDHOG is used to identify the target objects, in this instance vehicles. The detected objects are input to the particle filter which initiates the tracking procedure.

In contrast to the studies discussed above where a separate object detection mechanism was used in the research done by Okuma et. al [9] uses a cascading Adaboost algorithm to enhance the particle filter.At a higher level their study employs a combination of mixture particle filters and Adaboost method which the authors refer to as Boosted Particle Filter (BPF). Hue-Saturation-Value color histograms are used to develop observational models. The HSV space is is used due to the lack of susceptibility to variations in illumination in the HSV space. The Bhattacharya similarity coefficient is calculated in order to measure the distance between the reference color model and the candidate color model. The distance is then used to model the likelihood distribution.The authors recall that particle filters have a tendency to perform poorly when the posteriors are multi modal (tracking multiple targets).In order to mitigate this issue and handle multiple targets the mixture model approach is used which is derived from the Mixture Particle Filter (MPF) developed by Vermak et al [25]. In this approach each component is modeled by and individual particle filter and the resampling step is delegated to the individual filters so as to mitigate the effects of sample depletion.

The experimental results show that the model is robust at handling new objects moving into the frame. When a player enters the scene the Adaboost mechanism detects the player, and the player is tracked by the BPF after assigning particles to it. The Adaboost does not perform very well under situations where players appear in close quarters and the susceptibility of Adaboost towards large variations in intensity is also evident in a situation where a camera was flashed during the game, at which point Adaboost lost detection of several players.However the BPF handled and tracked the players in these situations making up for the lack of failures of the Adaboost detector. The authors propose several improvements, such as incorporating non-uniform backgrounds when training the Adaboost detector and dynamic value selection for the weighting parameter.

Although the BPF system by [9] demonstrated a robust hybrid tracking system it did not fare well under mutual target occlusions. The study done by Cai et al. [10] picks up from where [9] left and proposes several improvements to the system, mainly focused towards improving the limitations in situations of mutual occlusion.One of the main flaws pointed out by the authors in the previous research is that the MPF used by Okuma et al. [9] has a fixed number of particles and when a new object enters into the frame the some particles should be shared between multiple entities leading to lower accuracy. During occlusions the merge and split of particles in the MPF structure causes the loss of identities of the players. Therefore the researchers have continued using the boosting particle filter as the main basic filtering component while incorporating independent particle sets instead of using MPF. There is difficulty involved with modeling target motion dynamics due to the motions involved with a non stationary camera. In order to circumvent this they have mapped the locations of the target players in the image coordinates to the coordinate domain of the hockey rink which behaves like a stationary reference frame.Once this is achieved the motion of the hockey players may be predicted with a constant velocity autoregressive model as the players follow the physical laws of momentum. However hockey (ice hockey) is a special case where the playing surface has a very low friction. Therefore application of this technique may yield different results in other sporting domains. Autoregressive models consider historical data to predict the current state of the value.

$$x_t = Ax_{t-1} + Bx_{t-2} + CN(0, \Sigma) \tag{2.1}$$

The authors formulate the autoregressive model in eq. (2.1) for the motion dynamics of the target players. here A, B, C are autoregression coefficients and $N(0, \Sigma)$ is a Gaussian noise term with mean 0 and standard deviation 1. A color model similar to [9] is built and where the target is split vertically into two parts and histograms for which are built separately. The authors have resorted to embedding the mean-shift algorithm in order to stabilize the tracking results with the expectation of improving tracking accuracy and reducing the impact from background clutter and occlusion.

The RDHOG based particle filter by [27] utilizes the output provided by RDHOG detector to initialize trackers. If the frame rate of the video is sufficiently high a moving vehicle should be detected in subsequent frames and the detected objects should spatially overlap. The researchers use this premise to assign trackers to targets. If a detected object appears in consecutive frames with the targets overlapping in each frame a particle is assigned to that target.

The Simple Online Realtime Tracking, a research conducted by [28] highlight the importance of the object detection module in a tracking by detection approach. They present an online multiple object tracking framework that uses the afore discussed Faster RCNN[22] as the object detection component with a Kalman Filter and Hungarian algorithm for the object tracking component. The authors focus on the problem of pedestrian tracking and do not consider detections of other classes for the tracking problem. They approximate displacement of an object between frames with a linear constant velocity model independent of other object motion and camera motion. The state of each target is modeled by a vector x, as in eq. (2.2).

$$x = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T \tag{2.2}$$

In this state representation u and v represent the center coordinates of the bounding box in image coordinate space. The variable s represents the area of the bounding box and r the aspect ratio. In this model aspect ratio is considered to be constant. The Kalman filter predicts what the next state of the target, associations are then made with the new set of detections. When a successful association is made, the detection bounding box coordinates are used to update the target state. The assignment cost matrix is calculated by using the intersection over union (IOU) as a distance metric. Once the cost matrix is obtained the hungarian algorithm is used to solve the assignment problem and make associations between target tracks and detections. The experiments are evaluated on the MOT15 [29] benchmark. The authors were able to obtain a MOTA value of 33.4 % and MOTP of 72.1% . In comparison there results exhibited a better performance than other online tracking approaches that were used for the comparison. The researchers state that with the use of a state of the art object detector they were able to achieve good results even with a rudimentary approach in the tracking component.

2.3.2 Data association

Data association in the context of target tracking is the problem of making correspondence between new detections and track hypothesis. An approach that has had success in finding associations between targets and observations is the joint probabilistic data association (JPDA) [30] which calculates a joint probability score for the association of targets to the observations. In the basic form, JPDA considers all possible associations between measurements and targets when calculating a joint probability score[31]. This approach can be quite costly and computationally hard. Various techniques have been employed in numerous studies to mitigate the complexity involved in calculating the probability score.

Hamid et al. [31] tries to tackle the problem by an approximate solution. The JPDA consists calculating a marginalized probability score over the possible associations of targets and measurements, which may end up with a large number of terms. The approach taken to mitigate this was to select the m best probability hypotheses and to approximate the marginal probability as the sum over these m hypotheses. First the problem was formulated as an integer linear programming (ILP) problem. Once this is achieved, solution to the ILP gives the best likelihood for the data association. However the above approximate solution does not fare well under a time-varying number of targets.An extension named integrated JPDA (IJPDA) has been used in literature, which introduces extra computational complexity to the model.Instead a set of heuristics were used. If an observation is not claimed by a target, it is initiated as a new target and if a the number of consecutive missed detections for a target reaches a certain threshold the corresponding track will be terminated. In order to evaluate the results optimal sub-pattern assignment for tracks (OSPA-T) as well as CLEAR MOT [32] was used . A number of experiments were carried out on both simulated data and real world data. The a single frame and a 3-frame JPDA (3F-JPDA) proposed by the authors were compared against a Multiple Hypothesis Tracking model (MHT) and Interacting Multiple Model Joint Probabilistic Data Association (IMM-JDPA) model in previous studies. Overall the JPDA methods exhibited good performance than the MHT method due to the handling of long occlusions. The proposed JPDA model showed accuracy similar to the IMM-JPDA model with significantly lower time.The 3F-JPDA had better accuracy, but it took the longest time for the computation.

2.3.3 Neural networks based target tracking

The latest study conducted by Milan et al. [33] takes a very different approach towards the problem of multi target tracking. Although deep learning has achieved quite a lot of traction in the recent past in various domains including computer vision [34, 15], applications towards the problem of multi target tracking is scarce in literature. There study focus on employing Recurrent Neural Networks (RNN) for the purpose of tracking multiple targets. From a high level perspective it can be considered that RNNs work in a sequential manner.It is stated that RNNs are good at predicting the state but faces troubles when handling the data association task. In order to handle the data association component authors have used Long Short Term Memory (LSTM) unit. The input vector at a given time state is defined as $x_t \in \mathbb{R}^{N.D}$ where N is the number of targets to be tracked. Each target is represented by the corresponding bounding box coordinates (x, y, w, h).

A temporal RNN is used for the state prediction and update component. The RNN learns the dynamic model of targets along with an indicator for particle death and birth rates. The prediction for each frame is dependent of the state of the previous frame as well as the state of the hidden layers. A loss function was developed that is based on the multi object tracking accuracy (MOTA) [32] metric. An LSTM network is proposed by the authors for the data association component. The step by step functionality of the LSTM is used to predict and assign the target.A softmax layer was applied to the outputs along with normalisation.The expected output is a vector consisting of probabilities of each target for each observation. In order to calculate the loss due to false assignments, a negative log likelihood function was used. The approach was tested on data from the MOTChallenge 2015 benchmark.As baseline methods two variations of Kalman filters and an implementation of JPDA [30] method was used. One variation of the network based on the hungarian data association and a network that uses both RNN and LSTM fully were used in the experiments. The Neural models developed by the authors exhibited a higher recall than other methods at 37.8% and 37.1%. However the precision values of 75.2% and 73.5% was not enough to outperform the precision by other methods.

Recent advances in the applications of convolutional neural network has led researchers to incorporate features from CNN feature maps as object appearance models. One such research was conducted by Hong et al [35] where they try to use a discriminative saliency map as an object appearance model . The model uses CNN that is pre trained for large scale image classification. Another layer consisting of an online trained SVM is added to the network, this SVM is used to learn the target appearance at each time step. The features relevant to the target are identified by the parameters of the SVM and propagated backwards to obtain a saliency map that highlights the object of interest. The saliency maps of the positive examples are aggregated to build the target specific saliency map . In order to perform the tracking the authors employee sequential Bayesian Filtering by considering the saliency map as an observation. Target appearances in previous saliency maps are used to learn a generative model and a dense likelihood map is calculated by convolution between the appearance model and the target-specific saliency map. Once this is achieved the SVM model and the generative model are updated. The authors were able to achieve a good performance under occlusion with this approach. However the tests are focused on a single tracking scenarios, but it provides a good approach that needs to be extended towards a multi target scenarios.

Using a similar approach Wang et al. [36] work on a study to exploit the convolutional features of a fully convolutional network for online target tracking. The researches take a CNN pretrained on a large dataset and analyse the feature maps to find out which features respond best at discriminating a given target. They select the features maps from conv4-3 and conv5-3 layers of the VGG-16 network for this purpose. The obtained features are reshaped into a d-dimensional vector. The features are also used to create a mask of the foreground object by a sparse encoding method. They conduct a series of experiments to identify that layers higher up and closer to the output, in this case conv5-3 are better at object localisation as opposed to lower layers such as conv4-3 which are more sensitive to intra class appearance differences. With these findings the authors create two networks SNet and GNet which are fed the convolutional features conv4-3 and conv4-5 respectively. The GNet is tasked with capturing the category information of the target while the SNet works to discriminate target from similar looking targets in the background. When a new frame comes into the tracking pipeline a region of area centered around the previous target location with surrounding background context is input to the network. The image first goes through the VGG CNN and output the convolutional features, which then goes throug the GNet and SNet. At the output layer the GNet and SNet networks outputs a heatmap for the region activated by the target.Finally the target is determined by a distracter detection scheme that decides which heat map to be used from the GNet and SNet outputs.

Looking at these studies it is evident the tendency in the research community to use already computed CNN features for target object appearance modelling. On of the biggest appealing aspects of these features is the reduced computational cost associated with training a full deep neural network. The research work done by Wojke et al. [37] is also of interest to our research as it strives to develop tracking by detection model with frame by frame data association. They focus on the simple framework used in the SORT by Bewley et al. [28] approach and try to use those principals to build a robust tracking model. The authors state that one of the main drawbacks with sort is the high number of identity switches. The reason for this drawback is because the association metric accurate only when the state estimation uncertainty is low. The solution the authors provide is to employee a more robust association metric that takes into consideration both motion and appearance information. With this approach the authors are able to build a system that is robust against misses, and occlusions. The model uses a Kalman filter with an 8 dimensional state space vector $(u, v, h, \gamma, \dot{x}, \dot{y}, \dot{\gamma})$ where (x, y) are the center coordinates of the bounding box, γ the aspect ratio and h the height of the bounding box. For each track hypotheses an age counter is maintained which indicates the number of iterations since it was last associated with a detection. This age value is incremented at each Kalman prediction and reset to 0 when a detection is associated with the track. If the age exceeds the maximum age threshold, that track will be removed from the set of hypotheses. The authors try to solve the data association between the Kalman estimates and the new detections by incorporating motion and appearance information. The motion aspect is taken into consideration by calculating the Mahalanobis distance between the Kalman predictions and the detections. In order to take the target appearance information into consideration the authors compute an appearance embedding descriptor for each detection, furthermore a dictionary consisting the appearance descriptors of previous detections associated with a given track is also maintained. The cosine distance between the new detection and the track is calculated as an appearance metric for data association. A CNN is used to generate the appearance embedding descriptor, which is a 128 dimensional vector. After each detection the image patch bounded by the detection dimensions is extracted and fed to the CNN to obtain the appearance descriptor. The authors combine the two distance metrics to obtain a single cost value for each detection track pair and use a cascading matching algorithm to prioritise tracks with smaller ages over tracks that has gone a long time without associations.

2.4 Evaluation

One of the important studies conducted recently towards building a set of common metrics for measuring the performance of multiple object trackers is the research conducted by Bernadin and Stiefelhagen [32]. The authors initially state the qualities that are expected from a multiple object tracker. At all times the tracker should be able to find the correct number of actual objects, estimate the location of each target and it should be able to uniquely keep track of the objects. The criteria for the evaluation metrics are derived based on these requirements as follows.

1. The metrics should allow to determine a trackers precision with regard to estimating a target location.
2. They should exhibit the trackers ability to consistently track object through time, correctly trace one trajectory for each single object.

given a set of estimated object hypothesis $h_1, ..., h_n$ and ground truth objects $o_1, ..o_m$ the procedure is to first determine the best possible correspondence between h_i and o_j . Then compute the error for each object position estimation and accumulate all correspondence errors. If a hypothesis was not output for an object it will be considered as a missed calculation. Hypothesis to which no real objects exists will be calculated as false positives. If a detected hypothesis changed for an object compared to the previous frame, it will be calculated as a mismatch error. Once this is achieved authors state that the tracking performance can be expressed in two values, namely tracking precision and tracking accuracy.

In this chapter we started our discussion by looking at the main approaches used in literature for multi target tracking. We focussed our attention to discuss the main problems of sports player tracking in comparison to pedestrian tracking and tracking objects in a controlled setting. Emphasis was given to the importance of breaking down the problem into it's main components of object detection and tracking. We looked at classical approaches used for object detection and how the recent advances in object recognition and localisation frameworks utilising deep learning approaches have shaped the research landscape. With regard to object tracking, we looked at several scenarios where studies were conducted on sports specific events and we observe the success of recursive bayesian estimation based techniques in these studies . We briefly discuss the data association methods used in some of the studies and the recent trend in using deep learning approaches such as RNNs in tracking scenarios as well. Delving further into the domain of deep learning we observe the recent approaches of repurposing deep convolutional features for target appearance modelling. Finally we have a look at evaluation metrics that are commonly used in literature such as MOTA and MOTP. In the next chapter the discussion will be focussed on the methodology we propose to solve the multi target tracking problem. Our aim is to treat the problem as a tracking by detection problem due to the success this approach has had in literature. We have an in depth discussion about the details of the object detection module followed by the tracking module.

Chapter 3

METHODOLOGY

The problem of multiple player tracking in rugby has some similarities with other sporting domains that were discussed in the previous chapter such as the unpredictable and erratic movements of players. At the same time it poses some unique problems of its own to be solved. A game of rugby may involve two types of playing situations referred to as set play and open play. Set play is a predefined playing scenario usually originating from a scrum or a lineout with a static field position. Open play is when a sequence of playing scenarios occur without the stoppage of time, or interruption by the referee.

Playing scenarios that can occur during open play are Maul, Ruck and Tackle which are collectively referred to as breakdowns as they indicate a breakdown of open play. The general movements of players are similar to the sports discussed. Main distinguishing factor between an open play and a set play is that set play is started with the referees whistle whereas open play occurs as a continuous sequence of uninterrupted events. Due to the dynamic nature involved with open play scenarios it is vital to keep track of how players position themselves, arrive at the break downs and how they react to certain events. In contrast set play does not need in depth analysis as factors such as player positioning and running angles are predefined and are originating from a static position. Therefore the study will be mainly focused towards events related to open play events. In the discussion that follows the term "breakdown" will be used to refer to all open play events such as rucks,mauls and tackles.

Considering the main approaches discussed in the literature and the success it has exhibited in the past it is evident that a player detection and tracking approach is more suitable for this study in order to discriminate between players movements and identify breakdown events. Therefore our tracking system will be consisting of two main subsystems that will be acting in two steps .

In the first step the player detector will detect players and breakdowns in video frames when they occur. In the second step the object tracker will be making associations between the newly detected observations and the established track hypotheses. If a tracked player is not detected for a certain amount of time the track hypothesis associated to the player will be removed. Similarly when a previously unknown player enters the frame a new track hypothesis will be added by the tracker subsystem. In order to obtain the best tracking performance the data association step makes use of several approaches such as bounding box intersection over union, Kalman filter motion prediction and object appearance embedding based distance for track re-identification.

The next sections will be describing in detail the different components of the full tracking system. We will start with object detection and move on to the tracking system where the data association and tracking components will be discussed in detail.

3.1 Player and break down detection

A large amount of research has been conducted in the domain of object recognition and localisation in the recent past. As discussed in the literature review the recent success of methods such as R-CNN,Fast R-CNN,Faster R-CNN and YOLO [19, 21, 22, 23] has provided a platform on which to tackle the problems of object recognition in specialised domains. One such method, YOLO[23] provides a set of capabilities that are ideal for our research. This approach is primarily different from other approaches mentioned above, the object detection, classification and localisation occur in one single network. Since YOLO employees a grid based approach to localise objects of interest the need for cumbersome iterative approaches such as sliding window and computationally expensive approaches such as region proposals can be eliminated. The specific YOLO model that we will be using is called YOLO V3 [38] which has several incremental improvements over the original version.

Since the final objective of the research is to develop a system that is able to track rugby players across time we focus our attention mainly on using a trained model based approach such as YOLO for object detection and focus our efforts on the holistic system . To this end we do not focus on classical computer vision based approaches to solve the object detection problem.

3.1.1 Object detection and localisation with YOLO

At a higher level the YOLO model makes use of a fully convolutional neural network that acts as a feature extractor for an input image. The output layer predicts the bounding box coordinates for detected objects, the class probabilities an object confidence score.In the next few subsections we will have an in depth look at each component of YOLO model.

Feature extraction at multiple scales

The YOLO model makes use of a network architecture dubbed Darknet-53. The Darknet-53 network is a fully convolutional neural network (FCN) which consists of 53 convolutional layers (table 3.1) along with residual and upsampling layers. Since the network does not make use of pooling layers miniscule details of the input image are preserved relative to the downsampling that occured with the use of pooling. This network architecture is able to make predictions on features generated at three different scales. At some predefined layers, the layer output is routed to the output layer directly with skip connections. These routed output feature maps have dimensions (13,13), (26,26) and (52,52) corresponding to the grid sizes they represent. We are able to make detections at three different scales by this increase in grid cell resolution.

Table 3.1: Darknet 53

	Type	Filter	Size	Output
	Convolutional	32	3 x 3	256 x 256
	Convolutional	63	3 x 3 / 2	128 x 128
1 x	Convolutional	32	1 x 1	
	Convolutional	64	3 x 3	
	Residual			128 X 128
	Convolutional	128	3 x 3 / 2	64 x 64
2 x	Convolutional	64	1 x 1	
	Convolutional	128	3 x 3	
	Residual			64 x 64
	Convolutional	256	3 x 3 / 2	32 x 32
8 x	Convolutional	128	1 x 1	
	Convolutional	256	3 x 3	
	Residual			32 x 32
	Convolutional	512	3 x 3 / 2	16 x 16
8 x	Convolutional	256	1 x 1	
	Convolutional	512	3 x 3	
	Residual			16 x 16
	Convolutional	1024	3 x 3 / 2	8 x 8
4 x	Convolutional	512	1 x 1	
	Convolutional	1024	3 x 3	
	Residual			8 x 8
	Convolutional	512	3 x 3 / 2	16 x 16
	Avgpool		Global	
	Connected		1000	
	Softmax			

Object localisation with anchor boxes and grid

As mentioned in the introduction one of the main advantages of YOLO is that it goes through the image only once as opposed to other approaches that employee region proposals or sliding windows that are used to localise the objects of interest. In order to achieve this the image is divided into a grid of $N \times N$ cells. Each cell is responsible for predicting the bounding box coordinates along with the confidence score of how confident the cell is about the predicted bounding box containing and object and class probabilities. This is illustrated in figure (3-1).

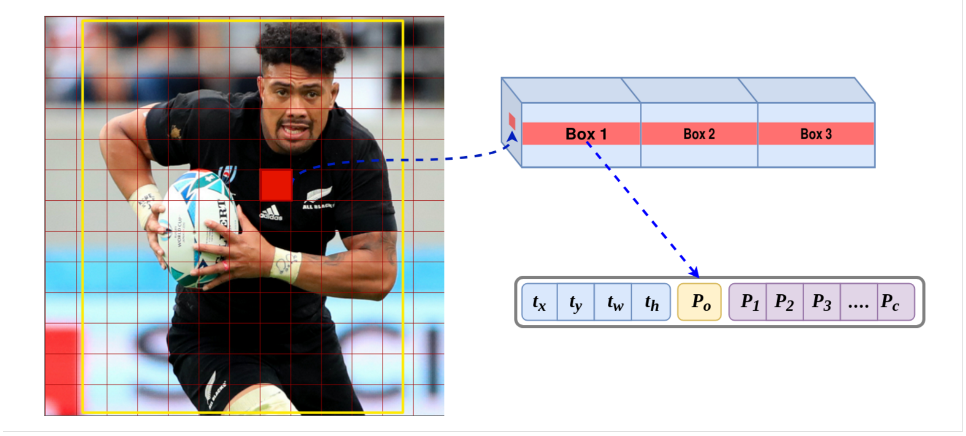


Figure 3-1: Each grid cell is responsible for predicting 3 bounding boxes. Each predicted bounding box estimation consists of the bounding box coordinates t_x, t_y, t_w, t_h, objectness score P_o and class probabilities $P_1, P_2, P_3, ...P_c$

.

Rather than predicting the coordinates of the bounding boxes directly relative to the global coordinates of the image YOLO uses an interesting technique with the use of *Anchor Boxes* . Anchor box dimensions are calculated at the preprocessing time by taking the average dimensions of the ground truth boxes. At prediction time each cell predicts a vector of bounding box coordinates (t_x, t_y, t_w, t_h) corresponding to the number of anchor boxes (3 in this case).These predicted coordinate values needs to be transformed to the image coordinate space in order to obtain the detection bounding boxes. The x,y coordinates of the detection bounding box is based on the offset of the corresponding cell from the global image coordinates and the width and height dimensions are derived from the anchor box dimensions. If the detected bounding box coordinates are given by b_x, b_y, b_w, b_h where (b_x, b_y) is the x, y coordinates of the box and b_w, b_h are the width and height of the bounding box, c_x, c_y are the offset of the cell making the prediction from the image origin coordinates and p_w, p_h are the width and height of the anchor box associated with the prediction. The coordinate transformation can be formulated as follows.

$$b_x = \sigma(t_x) + c_x \tag{3.1}$$

$$b_y = \sigma(t_y) + c_y \tag{3.2}$$

$$b_w = p_w e^{t_w} \tag{3.3}$$

$$b_h = p_h e^{t_h} \tag{3.4}$$

The figure 3-2 helps to illustrate this transformation relationship further. In this scenario the σ is the sigmoid function.

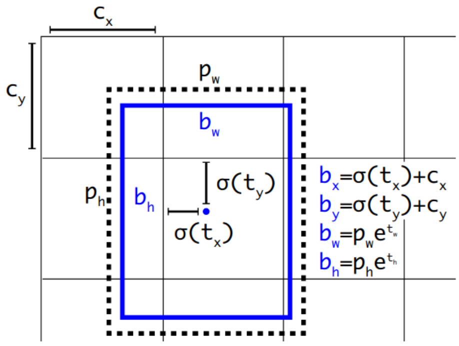


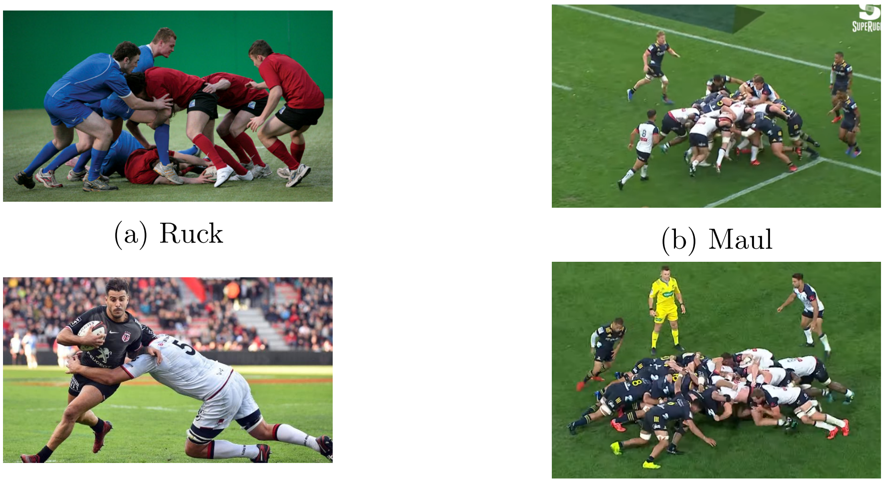
Figure 3-2: Bounding box prediction with dimension priors and location prediction. The center coordinates of the bounding box are predicted relative to the cell location with the use of sigmoid function. The width and height are derived from the anchor box (bounding box prior) dimensions. source: [38]

Therefore if we consider a single cell at the output it would output a tensor of shape $N \times N \times [A(4+1+C)]$, in which N is the number of grid cells , A number of anchor boxes, C number of classes. We add 4 bounding box dimensions and 1 for objectness score. In our training cycles the values for the above parameters are as follows. We are training for 5 classes, therefore $C = 5$, with $A = 3$ anchor boxes. The grid size N takes on the values 13,26 and 52 at 3 different scales as explained in the next section.

Training the network

One of the problems that we try to tackle in this research is the problem of tracking under high occlusion. The visual occlusion occurs at different play situations of the game. For the sake of brevity we shall refer to these situations as breakdown events. In each of the breakdowns due to the high physical engagement between the players a large amount of visual occlusion occurs. In these situations rather than trying to identify individual players we try to detect and identify what kind of breakdown event occured. We keep tracking individual players who are in the open field and not part of the breakdown area, and consider the players engaged in the breakdown as a single breakdown entity. In order to differentiate between individual players and breakdown events we aim to train the YOLO model to classify the five classes one for individual players and other four for breakdown events ruck,scrum,maul and tackle.

The breakdown events that we are concerned with are as shown in figure fig. 3-3. A *Ruck* (fig. 3-3a) is defined when the ball is on the ground and two or more players are physically engaged over it. A *Maul* (fig. 3-3b) occurs when a player carrying the ball is held by one or more opponents and one or more of the ball carrying players team mates bind on the ball carrier, all players should be on their feet. The defending team would try to obtain the possession of the ball by tackling the ball carrier of the attacking team. A *Tackle* (fig. 3-3c) occurs when a ball carrier is held and brought to ground by a player from the opposing team. A scrum (fig. 3-3d) is used to restart play after a minor infringement. In a scrum eight players from each team physically engage (bind) forming a tunnel in the middle. The non-offending team will put the ball to into the tunnel to restart play.



(a) Ruck

(b) Maul

(c) Tackle

(d) Scrum

Figure 3-3: Different breakdown events occur during a game of Rugby. sources [39, 40, 41]

For the purpose of training, broadcast video footage from the public domain is used. The videos are broken down into individual frames and saved as image files. These image files were later used to create the annotations. Given the depth and the large number of parameters in the Darknet network it would be counter productive to train the network from a randomly initialised state. Therefore we employ a technique named transfer learning, where a network pretrained on a large dataset is used to get a head start in the training process. In this research a Darknet model trained on pascal VOC [42] dataset was used as the base model. We remove the output layer of the Darknet model and retrofit the final layers with a customised layer that can accommodate detecting and classifying of the five classes of interest. We train only the head portion of the network by freezing the weights of the body portion.

3.2 Player tracking

Once we obtain the detection bounding boxes and classes for each individual frame it is the task of the tracking module to identify and make associations with existing tracking hypothesis and the incoming set of detections. There are two main approaches available to achieve this. First approach is a post processing approach, in which we feed all the available frames to a detector and obtain the detections which are then passed into the tracking module which will make the data associations. This approach can have a higher accuracy but is not suitable for a real time application. The other approach is to get the detections at each frame and feed them into the tracker, where all the detection and associations are done online. This method will be fast but getting a higher tracking accuracy with traditionally available methods is not an easy feat.

For the purpose of data association between detections and track hypotheses we use a combination of several approaches that were proven to be effective in literature, let's discuss these approaches step by step starting from the least complex approach.

3.2.1 Bounding box intersection and euclidean distance based data association

At the simplest level consider two subsequent frames F_i and F_j with their corresponding detection sets $D_i = \{d_i1, d_i2, ..., d_in\}$ and $D_j = \{d_j1, d_j2, ..., d_jn\}$. We calculate a pair of distance metrics for each pair of detections from D_i and D_j. One distance metric is the intersection over union of the two bounding boxes. The intersection of the two bounding boxes is divided by the union of the two bounding boxes as defined in eq. (3.5).

$$IOU(A, B) = \frac{A \cap B}{A \cup B} \tag{3.5}$$

Along with the intersection over union we use the euclidean distance between center coordinates of the two bounding boxes as the second distance metric. Lets consider that the center of bounding box A is given by $p_a = (x_a, y_a)$ and the center of the bounding box B is given by $p_b = (x_b, y_b)$, the euclidean distance can be calculated by eq. (3.6).

$$ED(A, B) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \tag{3.6}$$

We incorporate this distance measurement for the final data association assignment.

3.2.2 Track prediction with Kalman Filters

Kalman filter is a special form of Bayesian filtering that assumes the data is normally distributed. It has been used in numerous research with varying success. We use a kalman filter assigned to each track for state estimation as used in [37]. With this approach we are able to take into consideration the motion of the object of interest at the time of data association. The Kalman filter consists of two steps, namely prediction step and the update step. In order to represent the state of objects we use an eight dimensional vector $x = (x, y, h, \gamma, \dot{x}, \dot{y}, \dot{h}, \dot{\gamma})$ where (x, y) are the center coordinates of the bounding box, γ the aspect ratio and h

the height of the bounding box.

For each track hypotheses a counter is maintained that would keep track of the number of iterations since the last detection association. At each prediction step of the Kalman filter if the last detection association exceeds the threshold A_{max} that track hypotheses is considered to have died and will be removed from the set of tracks. In order to make the association between the predicted track and existing hypotheses the squared Mahalanobis distance is used. In the eq. (3.7) we denote the projection of the ith track distribution into measurement space by (y_i, S_i) and the j-th bounding box detection by d_j. The Mahalanobis distance checks how many standard deviations away from the mean track location the detected bounding box is.

$$d^{(1)}(i,j) = (d_j - y_i)^T S_i^{-1}(d_j - y_i) \tag{3.7}$$

3.2.3 Object re-identification

One of the main features of the data association step with regard to the tracking subsystem is the entity re-identification mechanism. In the research conducted in [37] the Wojke et al. extract the image patch from the bounding box and obtain an appearance descriptor r_j, where $\|r_j\| = 1$. The appearance descriptor is obtained by passing the image patch through a CNN, which outputs the 128 dimensional re-identification vector. The architecture of the re-identification network which is based on the work done by Wojke et al. [43] is described in table 3.2. Figure 3-4 illustrates the end to end process from the input of the image to the output of the detected bounding boxes and the identity embeddings.

Each track keeps a gallery $R_k = \{r_k^{(i)}\}_{k=1}^{L_k}$ of the last $L_k = 100$ association descriptors associated with that track. After each detection iteration we look for the track and detection pair with the smallest cosine distance between the corresponding identity descriptors in the appearance space eq. (3.8).

Table 3.2: Deep Cosine network

Type	Patch Size/Stride	Output
Convolutional	3 x 3/1	32x128x64
Convolutional	3 x 3/1	32x128x64
Max Pool 3	3 x 3/2	32x64x32
Residual 4	3 x 3/1	32x64x32
Residual 5	3 x 3/1	32x64x32
Residual 6	3 x 3/2	64x32x16
Residual 7	3 x 3/1	64x32x16
Residual 8	3 x 3/2	128x16x8
Residual 9	3 x 3/1	128x16x8
Dense 10		128
Batch and l_2 normalisation		

$$d^{(2)}(i,j) = min\{1 - r_j^T r_k^{(i)} | r_k^{(i)} \in R_k\} \tag{3.8}$$

Once we the Mahalanobis distance for object locations and the cosine distance for object appearance is obtained we calculate a weighted sum of the two distance metrics to determine the final combined distance as in eq. (3.9). The use of Mahalanobis distance derived from motion information and cosine distance derived from appearance information provides solutions to two aspects of the association problem. The motion information can be helpful in predicting possible object locations and appearance information can be helpful to recover identities of tracks after periods of occlusion [37].

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j) \tag{3.9}$$

The problem of associating detections to tracks is solved as a series of cascading subproblems instead of considering it as a global problem. The intuition for this approach as described in [37] is as follows: If an object is occluded for a long period of time the uncertainty of the Kalman filter increases, the peak of the observation likelihood decreases. In a situation like this the association metric should account for the increased uncertainty and increase the distance between the track and the detection. However given a scenario where two tracks

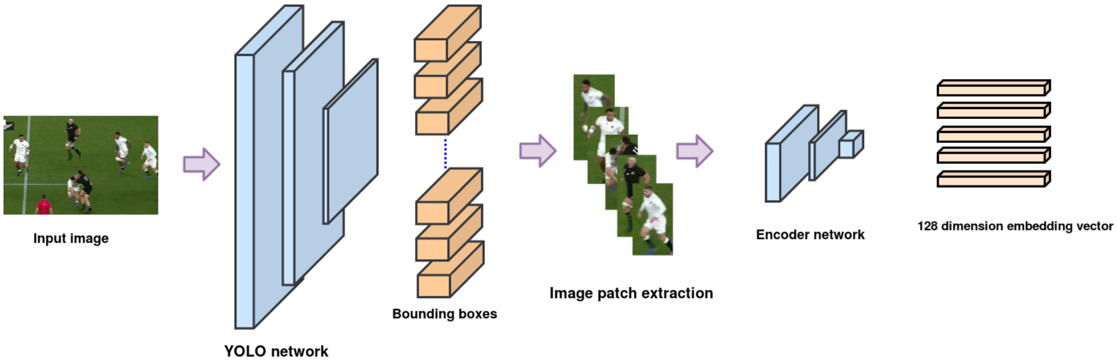


Figure 3-4: Visualisation of the end to end process of object detection and object identity embedding encoding

are competing for a detection the Mahalanobis distance tends to favour the larger uncertainty, since it effectively reduces the distance in standard deviations of any detection towards the track mean. This behaviour may lead to unstable tracks and track fragmentation. Therefore a matching cascade is introduced which prioritises the more frequently observed tracks. The age of the track a_i is the number of iterations it has gone without being associated with a detection. We iterate through the tracks starting from the tracks that has the lowest age and try to solve a linear assignment problem between the track and the detections and at the end update the sets of matched and unmatched detections. If any tracks have reached the A_{max} age it without being associated with a new track, it will be considered to have died out. Finally the IOU metric and euclidean distance is used to make associations in the set of unmatched tracks with $a_i = 1$ age.This increases robustness against erroneous initialisation, and helps to account for sudden appearance changes.

3.3 Evaluation of tracking performance

The data for training and evaluation will be extracted from rugby match videos. Some of the generic tracking systems have done the evaluation on simulated data. But due to the specific goal that is being pursued in this research evaluating on actual footage is of importance. But the primary concern is the lack of data in the domain of rugby. Therefore footage will be manually tagged and used for evaluation.

With the increase in research being conducted with regard to multiple object tracking several approaches to evaluate the tracking performance has been developed. Considering these numerous performance measures it is evident that the measures that may be useful to a certain end application may not be suitable for a different application. We use one of the most popular measures used in multi object tracking literature, the CLEAR MOT [32] metrics. Apart from that we use a another measure named identification F1 score [44] which provides us a different perspective about the tracking performance.

3.3.1 CLEAR MOT tracking measure

The CLEAR MOT metric was specifically designed to fill the lack of a widely accepted standard metric for multiple target tracking. The designers of CLEAR multiple object tracking metric was specifically aiming for two main objectives. Specifically the ability of the metric to evaluate the tracking algorithm's precision in determining object locations, and the ability to show the tracker's performance to correctly track object trajectories, creating exactly one track hypothesis for an object of interest. Based on the above the the authors identify the main measurements as tracking precision, which measures how well the exact positions of targets are represented and the tracking accuracy which reflects the number of mistakes the tracker makes in terms of false positives, false negatives, mismatches and failure to recover tracks.

The multiple object tracking precision (MOTP) is the concrete manifestation of the tracking precision measure the authors designed. As a first step towards evaluating the performance a correspondence mapping is created between the track hypotheses and ground truth. Since we are using bounding boxes to represent objects of interest, the intersection over union between the track and ground truth bounding boxes will be used to establish correspondence. The MOTP metric is calculated as shown in eq. (3.10).

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{3.10}$$

It is the accumulated error for the position estimation for pairs of matched hypothesis and target pairs divided by the sum of matches made.It isolates the ability of the tracker to estimate target positions precisely, independent of other aspects such as keeping consistent trajectories. The second metric Multiple Object Tracking Accuracy ($MOTA$) expresses how many errors the tracker has made in terms of misses, false positives, mismatches, failures to recover tracks etc. Where number of misses, false positives and mismatches are defined as m_t, fp_t, and mme_t respectively, we can consider the error ratio for misses in the sequence as eq. (3.11),

$$\bar{m} = \frac{\sum_t m_t}{\sum_t g_t} \tag{3.11}$$

and the ratio of false positives as,

$$\bar{fp} = \frac{\sum_t fp_t}{\sum_t g_t} \tag{3.12}$$

and the ratio of mismatches as,

$$\bar{mme} = \frac{\sum_t mme_t}{\sum_t g_t} \tag{3.13}$$

and by incorporating the above three error measures eqs. (3.11) to (3.13) the $MOTA$ can be obtained by eq. (3.14).

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{3.14}$$

In this regard the $MOTA$ takes into consideration all object configuration related errors that the tracker is making.

3.3.2 Identification F1 score

In event based measures such as the previously discussed MOTA measure the errors are calculated irrespective to the identity assigned to the track , but are informative in pinpointing the events where some errors occur. A ground truth track that switches between two computed tracks over n frames can incur at least 1 penalty and at most $n-1$ penalties without identifying the true track associated with the object. There can be situations where this may not be perfectly representative of the performance of the tracker. The ID measures [44] focuses on evaluating how the computed tracks conform to the ground truth tracks. Given a scenario where two tracks track the same ground truth object at two different segments of time, it decides which track to associate with the ground truth for the purpose of the evaluation process. Once this association is made every frame that the track deviates from the ground truth object is penalised.

In order to obtain the optimal matches between ground truth and computed tracks a bipartite graph is created by minimising the number of mismatched detections between one ground truth and one computed track. The bipartite graph $G = (V_T, V_C, E)$, where the vertex set V_T consists a node τ for each ground truth value and one false positive node f_γ^+ for each computed value γ. Similarly the vertex set V_C consists of a node γ for each computed track value and one false negative node $f_\gamma^-$for each ground truth track τ [44].

At a higher level for a given pair of ground truth nodes τ and γ at time t if the intersection over union between the two bounding boxes does not exceed a predefined Δit is considered a miss, and is represented by eq. (3.15),

$$m(\tau, \gamma, t, \Delta) = 1 \tag{3.15}$$

If there is a match between τ and γ it is defined as $m(\tau, \gamma, t, \Delta) = 0$. Furthermore if either f_γ^+ or f_γ^- is considered, any other detection in the other trajectory is considered a miss. The overall cost on an edge between τ in the set T_τ and γ in the set T_γ is represented by eq. (3.16)

$$c(\tau, \gamma, \Delta) = \sum_{t \in T_\tau} m(\tau, \gamma, t, \Delta) + \sum_{t \in T_\gamma} m(\tau, \gamma, t, \Delta) \tag{3.16}$$

The first term represents the false negatives and the second term represents the false positives. A minimum cost solution to this bipartite graph will provide the best match between the ground truth and computed tracks. Once the best fit solution is found, false negative IDFN , false positive IDFP and true positive IDTP can be calculated as eqs. (3.17) to (3.19).

$$IDFN = \sum_{\tau \in AT} \sum_{t \in T_\tau} m(\tau, \gamma_m(\tau), t, \Delta) \tag{3.17}$$

$$IDFP = \sum_{\tau \in AC} \sum_{t \in T_\tau} m(\tau_m(\gamma), \gamma, t, \Delta) \tag{3.18}$$

$$IDTP = \sum_{\tau \in AT} len(\tau) - IDFN = \sum_{\tau \in AC} len(\gamma) - IDFP \tag{3.19}$$

Where AT is the set of all ground truth identities and AC is the set of all computer track ids. The precision IDP, recall IDR and the F1 score IDF1 can be derived with eqs. (3.20) to (3.22).

$$IDP = \frac{IDTP}{IDTP + IDFP} \tag{3.20}$$

$$zIDR = \frac{IDTP}{IDTP + IDFN} \tag{3.21}$$

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{3.22}$$

In order to get a wider perspective of the performance of our tracking model we report the IDF1 score along with the MOTA and MOTP measure that were discussed above.

This chapter focussed on the methodology that we are using for this study. We had an in depth look at the YOLO model that we are using for the object detection module and the benefits it offers over the other object detection and localisation modules. The discussion was then focussed on the individual components of the tracking module. The discussion touched on incorporating appearance information and motion information for the data association step. In order to incorporate motion information we use a Kalman Filter and a object entity embedding encoder is used to calculate an object appearance embedding for each detection. We looked at the technical details of the two different evaluation techniques we are using for evaluation of the experiments. In the next chapter we will discuss the experimental setup used in the study. Then we will observe and analyse the results we obtained.

Chapter 4

EXPERIMENTAL SETUP

In this chapter we will continue from where we left off in the methodology section to cover the implementation aspects of the experimental setup. The discussion will delve deep into the implementation aspects of the object detector and the tracking approach we used for this study. With regards to the object detector we were able obtain a high success rate from the transfer learning approach we employed. In order to solve the data association problem we conduct several experiments and obtain good results with the re-identification embedding descriptor approach.

4.1 System architecture

The system consists of two main components as mentioned in previous chapters. The object detector searches and detects the objects of interest such as individual players and the object tracker assigns tracklets to detected objects and make associations between target objects and tracks across the frames. The process will be executed in a single pass online manner. We use video footage obtained from the public domain to train the YOLO object detector model. We were able to get good results with a pre-trained model for the entity descriptor model, therefore we did not pursue retraining or fine tuning of the person re-identification model in the tracking module.

4.2 Training data

The dataset comprises of broadcast footage of rugby matches that were taken from the public domain. Since these are videos recorded for the purpose of broadcasting we tend to get a lot of sudden camera movement. Another aspect of the dataset is that the camera angle keeps interchanging between a wide angle shot and a close up shot, therefore it was difficult for us to extract a clip with a consistent camera angle for an extended length of time. The wide angle shots give a better view of the overall field with visual information about the bounds of the playing field, while the close up shot gave us more detailed pixel data at the loss of overall location information. For the purpose of training we resize all the frames to the dimensions of 640 x 360 pixels. When we resize the wide angle shot video clips to this dimension we observed a loss of detector accuracy relative to the close up shot. Therefore the study was focused on clips that were recorded with a close up angle.

We used two well known image annotation tools, CVAT[45] and VATIC[46], both of which has the capability to interpolate the bounding boxes over a small number of frames once the starting and ending frames are defined. The bounding boxes were annotated with label types Player,Ruck,Maul,Scrum and Tackle. Note that we consider the referee as a player as well. We used 11 clips for training consisting of a total of 2157 frames.

The detailed breakdown of the different class instances used for the training of the object detector can be seen in the table 4.1. It is evident that there is a class imbalance between player class and the other breakdown classes. This imbalance is a side effect of breakdown events occurring less frequently relative to the individual player instances. However we did not observe a significant lack of performance in discriminating between players and breakdown events.

Table 4.1: Training Dataset

Clip	frames	mauls	players	rucks	scrums	tackles
clip1	52	0	416	0	0	0
clip2	175	0	350	0	0	0
clip3	79	0	711	0	79	0
clip4	42	0	420	0	0	0
clip5	165	0	2145	0	165	165
clip6	303	0	1818	0	303	0
clip7	162	0	2592	162	0	0
clip8	73	0	657	0	0	0
clip9	189	168	1425	0	0	0
clip10	522	265	3939	152	0	0
clip11	395	0	4887	106	0	189
total	2157	433	19360	420	547	354

4.3 Experiments

To the best of our knowledge this is the first study that has focused on applying computer vision aided automated tracking to the domain of Rugby Union. Most of the tracking literature found in recent times focus mainly on pedestrian tracking and in the domain of self driving cars. Therefore we do not have a suitable benchmark for a fair comparison. We conduct several experiments, following the path of followed making improvements to the models incrementally. We believe this would help to illustrate the performance gains we obtained against a simple baseline.

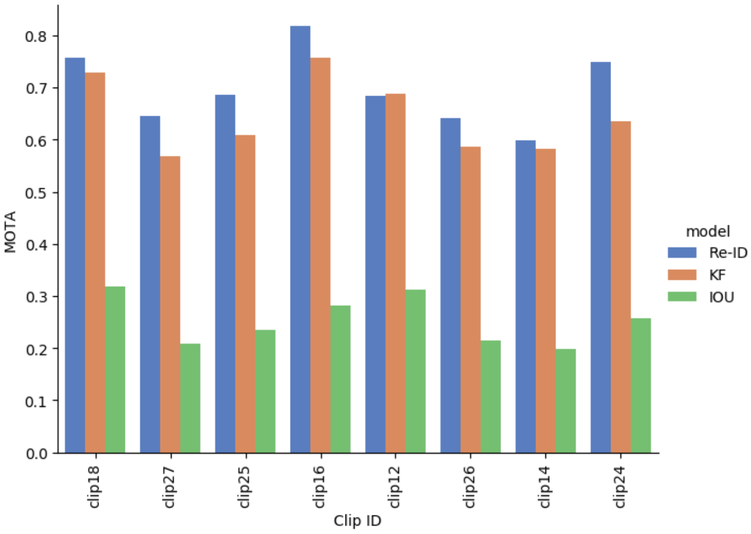
As was mentioned in the methodology section the first experiment makes use of a tracking module that makes associations between track hypotheses and detections with the intersection over union metric. We refer to this experimental model as IOU. The second approach named KF uses a Kalman Filter associated with each track to predict the next state of the track, it also employees the cascaded association algorithm of [37] to prioritise the latest detections when making associations. The final approach named Re-ID uses image embedding feature descriptors along with the methods used in the previous models to make data associations.

We primarily use the well known MOTA and MOTP metrics to evaluate the performance of our models, apart from theses metrics we calculate other measures such as number of track switches, detection precision and recall, ID F1 score and the mean frames per second.

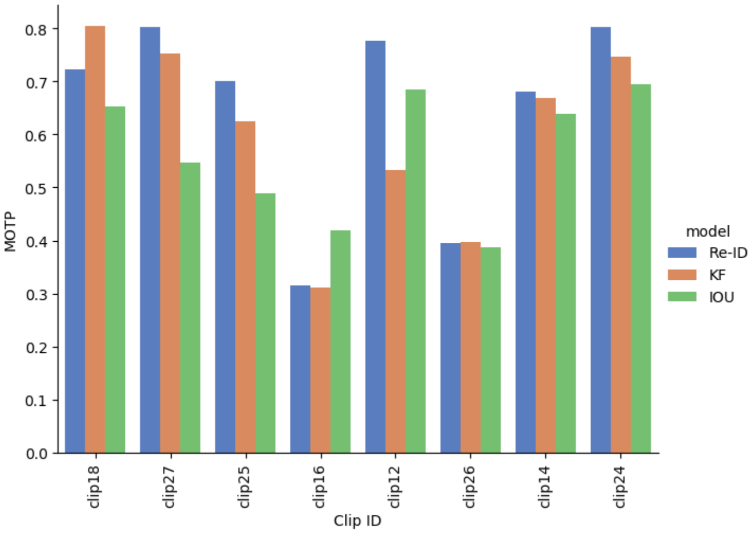
4.4 Results

The performance results for the different experiments conducted are as shown in the table 4.2.The overall results in terms of the multiple object tracking accuracy measure is above 0.5 on average for the Re-ID model. We can see that only in clip12 the MOTA value of KF approach has surpassed Re-ID by a very small margin. However a similar success of the Re-ID model is not exhibited when considering the multiple object precision measures. Looking at the data we can observe that in the instances where the MOTP value of the Re-ID model is reported below the other two models, the margins are very low. In order to interpret this relative lack of performance of Re-ID, it is worth revisiting the definition of the MOTP metric eq. (3.10). In short MOTP averages the accumulated error in position estimation between matched pairs of predicted track and ground truth by the number of total matches made. If we take the instances where Re-ID has under performed in MOTP, clip16, clip18 and clip26 we can observe that Re-ID was able to perform better in terms of the number of matches table 4.3. The reason for this maybe that during the track to detection association step Re-ID was able to successfully associate the tracks with difficult detections which the other two models were not able to associate at all, in doing so Re-ID has sacrificed in overall position estimation performance and in turn the tracking precision MOTP. This can be further established by observing that Re-ID has a better MOTA value in clips16 ,clip18 and clip26. We can get a more intuitive understanding of the MOTA results with the figure 4-1a and the results of MOTP with the figure 4-1b.

The table 4.3 shows other key measures calculated from the experiments. One other important measure that would help us gauge the performance is the number



(a) MOTA



(b) MOTP

Figure 4-1: Results for MOTA and MOTP measures

Table 4.2: MOTA and MOTP for each experiment

Clip	Frames	MOTA			MOTP		
		IOU	KF	Re-ID	IOU	KF	Re-ID
clip12	132	0.3113	**0.6871**	0.6840	0.6843	0.5328	**0.7756**
clip14	558	0.1975	0.5828	**0.5996**	0.6397	0.6685	**0.6803**
clip16	42	0.2811	0.7580	**0.8185**	**0.4196**	0.3102	0.3155
clip18	186	0.3183	0.7294	**0.7575**	0.6536	**0.8050**	0.7230
clip24	185	0.2576	0.6347	**0.7483**	0.6949	0.7466	**0.8019**
clip25	60	0.2350	0.6082	**0.6865**	0.4881	0.6251	**0.6999**
clip26	133	0.2141	0.5866	**0.6418**	0.3869	**0.3964**	0.3952
clip27	239	0.2091	0.5677	**0.6455**	0.5469	0.7528	**0.8029**

of track matches and track switches. The track matches measures the number of successful matches made between detections and tracks, a higher track matches values exhibit a good performance. The number of switches measure quantifies the number of tracks that switched the detections, a lower switches value is an indicator of good track performance. We can get a better visualisation of the results with the figures 4-2a and 4-2b for switches and matches respectively. It can be seen that both KF and Re-ID has a better performance than the simpler IOU approach and Re-ID has a slightly better performance than KF.

The IDF1 score provides us a better picture with how well the the tracker preserves the identity over the course of the track lifetime. The figure 4-3a exhibits very well the performance of the Re-ID model over the other two models. This is a reflection on the gain in tracking performance obtained by employing an object entity embedding network for object re-identification. If we take into consideration this figure along with plots for MOTA and MOTP 4-1, we can make an informed decision on instances where MOTA and MOTP alone does not help us to determine the best performing model.

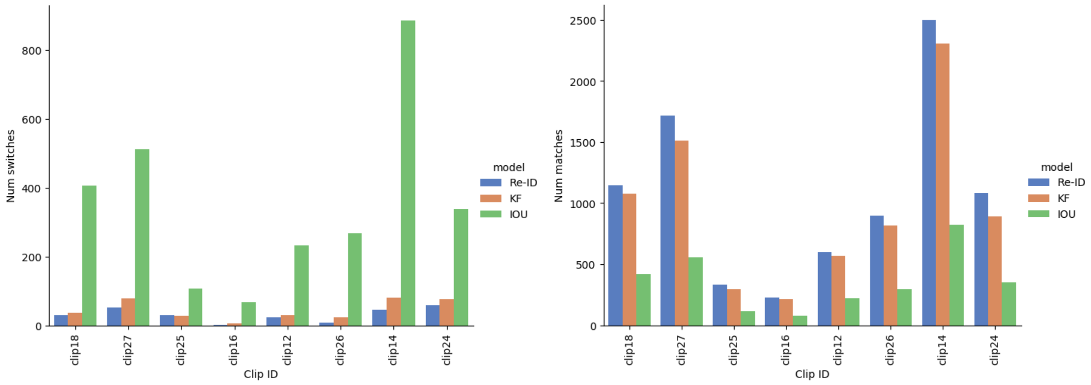
Since the aim of the application is to create a model that can perform well in real time to near real time scenarios looking at the execution speed is important. We obtain the mean frames per second value for each experiment to gauge the expected performance in terms of processing speed. The figure 4-3b shows that

Table 4.3: Experimental results

model	Clip	Matches	Switches	Precision	Recall	IDF1	mean FPS
IOU	clip12	224	233	0.9560	0.7009	0.0320	11.03
	clip14	825	886	0.9052	0.5232	0.0163	10.85
	clip16	79	68	1.0000	0.5231	0.0658	11.35
	clip18	420	406	0.9987	0.6276	0.0168	10.88
	clip24	350	337	0.9956	0.5100	0.0276	10.92
	clip25	114	108	1.0000	0.4577	0.2357	11.37
	clip26	299	267	1.0000	0.4054	0.0490	10.99
	clip27	557	512	1.0000	0.4014	0.0246	10.93
KF	clip12	570	30	0.8310	0.9202	0.6370	10.64
	clip14	2306	81	0.8564	0.7299	0.5268	12.26
	clip16	213	5	1.0000	0.7758	0.7323	11.26
	clip18	1076	37	0.9056	0.8457	0.6331	10.40
	clip24	891	76	0.9641	0.7178	0.4258	10.73
	clip25	295	29	1.0000	0.6680	0.7980	10.75
	clip26	819	23	1.0000	0.6031	0.6234	10.76
	clip27	1512	78	1.0000	0.5970	0.4502	10.50
Re-ID	clip12	600	24	0.8020	0.9570	0.7521	7.43
	clip14	2499	46	0.8254	0.7782	0.6965	3.74
	clip16	230	1	1.0000	0.8220	0.8980	10.25
	clip18	1143	30	0.8893	0.8913	0.8051	4.44
	clip24	1083	59	0.9383	0.8478	0.7267	5.78
	clip25	333	31	1.0000	0.7505	1.0224	9.74
	clip26	896	9	1.0000	0.6482	0.7812	5.98
	clip27	1719	52	1.0000	0.6650	0.6958	4.86

the overall performance of the Re-ID model with regard to FPS is lower than the other approaches. However considering that the FPS values used in broadcast videos range from 24 fps to 30 fps, this performance can be deemed suitable for near realtime applications where the speed of delivery is not mission critical.

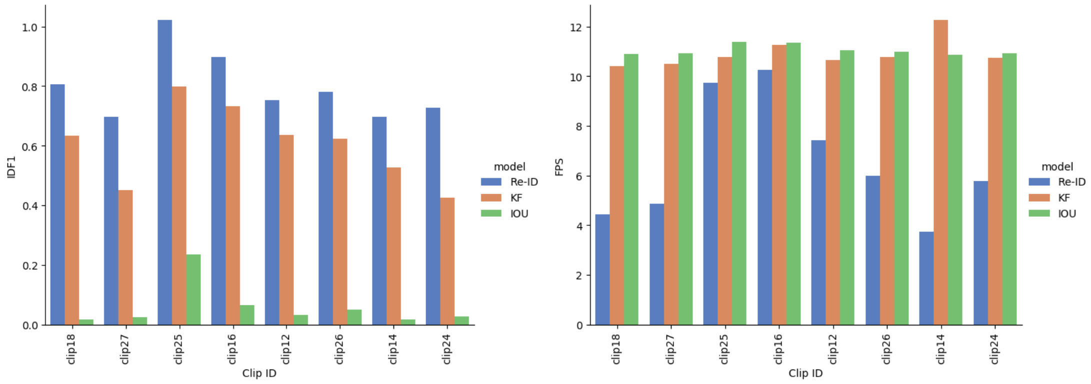
When considering the overall results the superior performance of the Re-ID model is evident.However the low FPS rate in comparison to the other two approaches may be of concern. The main reason for this lack of speed is because of the object entity re-identification network. For our application of automated player tracking we conclude that the performance gain in terms of tracking accuracy far outweighs the hindrances caused by the low processing speed. Possible approaches to improve the performance in terms of speed will discussed in the



(a) Number of switches

(b) Number of matches

Figure 4-2: Results for number of switches and matches



(a) ID F1 score

(b) frames per second

Figure 4-3: Results for ID measures and FPS values

next chapter.

Since this study is unique in it's application of multi target tracking to the domain of Rugby, there are no other studies that we can make a fair comparison of the results. However we are able to report the results from MOT16 [47], which is a benchmark for multi object tracking. The MOT16 uses 14 sequences of videos containing footage of pedestrians in a street setting. The results from the MOT16 benchmark in comparison to our Re-ID model cab be seen in table 4.4.

Table 4.4: Re-ID model performance in comparison to the MOT16 benchmark

Tracker	MOTA	MOTP	IDF1
BD_MOT16	73.2967	78.9515	69.4906
PA_MOT16P	71.4550	77.6016	65.6695
HT_SJTUZTE	71.2722	79.2697	67.6229
LMP_p	70.9968	80.2248	70.0714
FM16	68.7050	80.2782	70.4039
KDNT	68.1781	79.4059	59.9653
PT16	68.1603	80.1060	67.3698
POI	66.1142	79.5237	65.1117
CNNMTT	65.2146	78.4287	62.1947
TAP	64.7960	78.6537	73.5462
Re-ID	69.7712	64.9287	79.7225

As discussed earlier there is a vast difference when comparing a pedestrian setting with that of footage from Rugby. We believe that tracking Rugby players from broadcast footage is an immensely difficult task due to the high number of occlusions, high speed movements of players and sudden changes of directions. With the use of a state of the art object detector retrained for the specific case of Rugby and a data association approach which takes into account both appearance and motion information we were able to achieve results comparable to that of the MOT16 benchmark. In the next chapter we discuss the contributions made by our research and possible avenues for further research.

Chapter 5

CONCLUSIONS AND RECOMMENDATIONS

The increasing influence of various professional sports and the multi million dollar economies surrounding these respective sports leagues has created a demand for informed decision making at strategic and tactical levels. With the availability of numerous channels of data collection, we are seeing increased involvement of data analytics in various professional sports teams.

Tracking players during practice sessions and during matches can be considered an important tool in the analytics toolbox and provide insights when making high level strategic decisions or low level tactical decisions. Although a significant number of studies focussing on player tracking have been conducted with regard to professional sports such as American Football, Basketball and Soccer, to the best of our knowledge there are no studies that has applied automated computer vison based player tracking to Rugby. This can be explained by the fact that Rugby Union is a relatively new sports to enter into the domain of professional sports. The lack of studies conducted on Rugby player tracking provided us a unique opportunity, to apply research in computer vision and to develop a system to track players. The fact that Rugby is still a growing professional sport and the increasing demand for data analytics and tools served as a motivating factor for this research.

Our contributions from this research is threefold. We have developed a framework that can be utilised for tasks in multi target tracking. The framework contains modules for training object detector models, tracking players on new footage and for evaluating the end to end tracking performance. The system architecture is developed in a modular manner, with minimal coupling enabling the integration of different object detection and tracking modules if necessary. The most impactful contribution from this study is a YOLO model trained on Rugby footage with the use of transfer learning. We were able to leverage the performance of the trained object detector to achieve considerable results in tracking rugby players. Our final contribution is an annotated dataset that can be used for detection and tracking tasks applied to Rugby. One of the biggest impediments that we had to tackle early on in the research was the lack of tracking data on Rugby. We have painstakingly annotated and compiled a dataset from broadcast footage available in the public domain. We believe that these contributions will be valuable for further research in this domain. The authors plan to release the code, models and the dataset to the public domain in the future .

5.1 Further research

One of the areas that was deemed out of scope of this study was the feature to visualise player trajectories on a birds eye view projection of the play field. One of the main barriers we faced with this regard was that the camera angles did not clearly capture play field markings all the time, thus the problem of estimating player location relative to the play field became a non trivial problem which would require a separate research on it's own.

Another area for improvement would be to reduce the performance bottleneck that occurs at the entity embedding network. Currently each image patch for a detected player is sent through the network to obtain the 128 dimension embedding vector. Further research can be carried out test the ability to reuse the CNN features corresponding to the detection bounding boxes and route them via the last layer of the embedding network. If this is achievable it will save computational cost as the need to feed the image patches through many layers of the embedding network is bypassed.

Although the current iteration of this system is limited to tracking players the authors envision a future system that is capable of tracking playing field movements in realtime and provide coaches with outcome predictions of the next playing scenario as a game progresses. To this end the work done so far provides a good foundation.

5.2 Conclusion

The main objective of this study was to develop a system capable of automatically tracking players from broadcast footage, to this end we have been successful in developing a system with a significantly high accuracy. Considering the high number of occlusions that occur in a game of rugby we believe the success rate reported by the experiments are impressive. To the best of our knowledge this is the first study of this nature that tries to solve the problem of automated player tracking in Rugby Union. We believe that we have succeded in making contributions for further research applied to Rugby and to the domains of computer vision and multiple object tracking in general.

REFERENCES

[1] S. Edgecomb and K. Norton, "Comparison of global positioning and computer-based tracking systems for measuring player movement distance during australian football," *Journal of science and Medicine in Sport*, vol. 9, no. 1, pp. 25–32, 2006.

[2] T. H. Davenport, "Competing on analytics," *harvard business review*, vol. 84, no. 1, p. 98, 2006.

[3] B. McKenna. (2014) SAP helps germany lift the world cup. Web Page. [Online]. Available: http://www.computerweekly.com/news/2240224421/SAP-helps-Germany-lift-the-World-Cup

[4] S. R. Choudhury. (2016) Euro 2016: SAP and german football team worked to develop big data analytics. Web Page. [Online]. Available: http://www.cnbc.com/2016/07/07/euro-2016-sap-and-german-football-team-worked-to-develop-big-data-analytics.html

[5] C. W. Nicholas, "Anthropometric and physiological characteristics of rugby union football players," *Sports Medicine*, vol. 23, no. 6, pp. 375–396, 1997.

[6] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *European Conference on Computer Vision.* Springer, 2002, pp. 661–675.

[7] Zhu, Guangyu, C. Xu, Q. Huang, and W. Gao, "Automatic multi-player detection and tracking in broadcast sports video using support vector machine and particle filter," in *2006 IEEE International Conference on Multimedia and Expo.* IEEE, 2006, pp. 1629–1632.

[8] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 103–113, 2009.

[9] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *European Conference on Computer Vision.* Springer, 2004, pp. 28–39.

[10] Y. Cai, N. de Freitas, and J. J. Little, "Robust visual tracking for multiple targets," in *European conference on computer vision.* Springer, 2006, pp. 107–118.

[11] S. Barris and C. Button, "A review of vision-based motion analysis in sport," *Sports Medicine*, vol. 38, no. 12, pp. 1025–1043, 2008.

[12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2010, pp. 886–893. [Online]. Available: http://ieeexplore.ieee.org/document/1467360/

[13] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, Jun 2000. [Online]. Available: https://doi.org/10.1023/A:1008162616689

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.

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

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Ima-

geNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

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

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

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[20] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[21] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[24] C. Hue, J.-P. Le Cadre, and P. Pérez, "Tracking multiple objects with particle filtering," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 791–812, 2002.

[25] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multimodality through mixture tracking," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on.* IEEE, 2003, pp. 1110–1116.

[26] V. N. Vapnik and V. Vapnik, *Statistical learning theory.* Wiley New York, 1998, vol. 1.

[27] B.-F. Wu, C.-C. Kao, C.-L. Jen, Y.-F. Li, Y.-H. Chen, and J.-H. Juang, "A relative-discriminative-histogram-of-oriented-gradients-based particle filter approach to vehicle occlusion handling and tracking," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 8, pp. 4228–4237, 2014.

[28] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP).* IEEE, 2016, pp. 3464–3468.

[29] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942

[30] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE journal of Oceanic Engineering*, vol. 8, no. 3, pp. 173–184, 1983.

[31] S. Hamid Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid, "Joint probabilistic data association revisited," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3047–3055.

[32] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[33] A. Milan, S. H. Rezatofighi, A. Dick, K. Schindler, and I. Reid, "Online multi-target tracking using recurrent neural networks," *arXiv preprint arXiv:1604.03635*, 2016.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[35] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *International conference on machine learning*, 2015, pp. 597–606.

[36] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3119–3127.

[37] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[38] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[39] W. Rugby. Rucking. World Rugby. [Online]. Available: https://coaching.worldrugby.org/images/keyfactors/rucking.jpg

[40] S. RUGBY. Super rugby 2020 | highlanders v rebels - rd 5 highlights. SANZAR. [Online]. Available: https://www.youtube.com/watch?v=0s3CaSrtxZU

[41] F. Sports. Toulouse's sofiane guitoune is tackled around the waist at ernest wallon stadium. Fox Sports. [Online]. Available: https://cdn.newsapi.com.au/image/v1/987a9673f3aa09843438f805362f6454

[42] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[43] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE, 2018, pp. 748–756.

[44] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision.* Springer, 2016, pp. 17–35.

[45] A. Z. Nikita Manovich, Boris Sekachev, "Powerful and efficient computer vision annotation tool (cvat)," https://github.com/opencv/cvat, 2018.

[46] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowd-sourced video annotation," *International journal of computer vision*, vol. 101, no. 1, pp. 184–204, 2013.

[47] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831