AFL Player Detection and Tracking

Hayden Faulkner
School of Computer Science
University of Adelaide
Email: hayden.faulkner@adelaide.edu.au

Anthony Dick
School of Computer Science
University of Adelaide
Email: anthony.dick@adelaide.edu.au

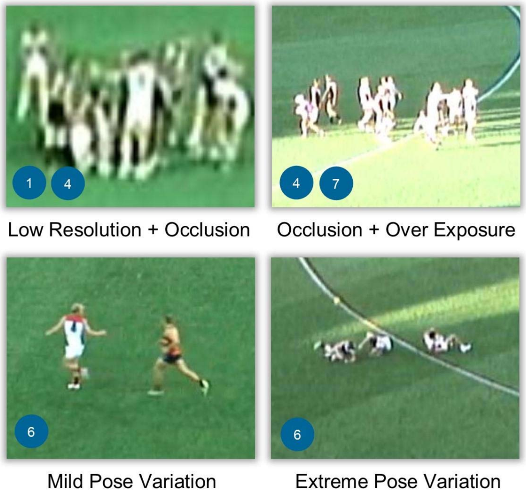
***Abstract*—This paper is an empirical study of the application of visual detection and tracking methods to the problem of locating and tracking all AFL players during a game. While most person detection and tracking algorithms are designed for pedestrians, we show that with appropriate modifications, state of the art methods can be adapted to a more challenging domain where motion is significantly more varied and occurs in a much wider area.**

I. Introduction

Driven by the priority of the surveillance and vehicle navigation applications, most tracking research focuses specifically on pedestrian tracking and detection. To a much lesser extent, the tracking of players in sports video has also received some attention, with most work revolving around popular international sports such as soccer and basketball. In Australia however, the most popular sport is Australian Rules Football run by the Australian Football League (AFL), with a supporter base of over 800,000 [1]. The motivation behind the use of visual tracking systems in sports, including for the AFL application, is to provide a foundation for a system that is able to automate game statistics for match and player analysis.

AFL football is a unique game that is currently only played professionally in Australia. This uniqueness presents some distinct properties and challenges that aren't found in other sports and pedestrian tracking problems. Specific challenges related to the AFL situation are (Figure 1):

1) the large size of the field makes covering the entire field at sufficient resolution difficult;
2) the number of persons constantly needing to be tracked is close to 50;
3) the fast movement of players, with sudden direction changes based on play, is generally more erratic than in other sports and for pedestrians whom often follow relatively straight paths;
4) the regular bunching of players into dense packs causing many difficult, often long lasting occlusions;
5) the lack of identifiable appearance differences between players on the same team, and sometimes players on opposing teams;
6) players are more deformable and take a more varied set of shapes, for example when making large strides whilst running or when lying on ground after contact with another player; and
7) the variability of the outdoor environment, for example the bright and dark areas of the field with sunny and shadowy conditions.



Low Resolution + Occlusion

Occlusion + Over Exposure

Mild Pose Variation

Extreme Pose Variation

Fig. 1. Some of the AFL's most challenging and unique situations. Numbers refer to the list numbers above.

In this paper we describe how some of these challenges can be overcome to create a functioning player tracking system for the AFL, who partly funded the research.

A. Related Work

We use a tracking-by-detection approach in our system. Firstly, an object detector is applied to individual video frames separately to obtain target positions, and potentially, target appearance information. Secondly, a tracker uses the position information to correlate detections referring to the same target over a period of time.

Using a detector has a number of advantages including being able to better handle clutter, occlusions and varying backgrounds, while not being subject to model drift [2] caused by trackers altering their detection model online. Recent advances in detection and classification methods [3] [4] [5], have made this an increasingly viable approach and in fact such methods now produce the best results for many tracking benchmarks [6] [7] [8] [9].

Multi-target tracking approaches can be broadly separated into two categories: local methods that use information from past frames to estimate the current state recursively; and global methods that estimate the state based on an optimal association for all tracks within a temporal sliding window. In this work we investigate both approaches, implementing a basic local Kalman Filter method, and comparing it to a global energy minimisation technique proposed by Milan et al. [10] [11].

978-1-4673-6795-0/15/$31.00 ©2015 IEEE

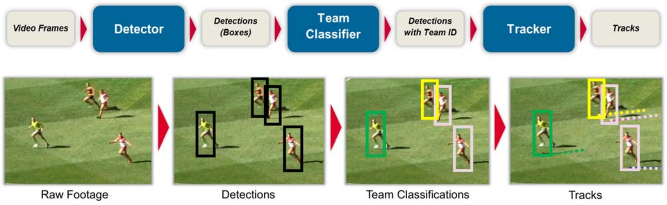


Fig. 2. The overall pipeline. Top: main steps in the method. Bottom: example output of each step.

1) Detection & Tracking for Sports: One of the earliest publications involving sports tracking was by Nillius et al. [12], in which they proposed the use of a track graph. The nodes in the graph denoted the tracks and the edges represented how each of the tracks spilt and merged together. Then identifying the problem as one of inference, they aimed to find the most likely set of paths for the targets, given the appearance vector of such targets. A number of the authors of this paper went on to develop TRACAB[1], a commercially available application of sports tracking that has been deployed on sports including soccer, tennis, basketball and cricket. It was trialled for AFL but not used as it wasn't able to cope with the complexity of the AFL problem.

Okuma et al. [13] proposed a self-learning framework which improves player localisation, allowing an unconstrained number of target objects to be tracked with non-static cameras. They tested their framework on broadcast footage of ice hockey and basketball, and were able to classify players using a parts model based on team. AFL players are much less likely to be captured at the resolution necessary for a parts model due to the large size of AFL field. Also the lighting conditions in an outdoor AFL match environment are problematic for this method, especially in the team classification process.

Hamid et al. [14], [15] proposed an approach for robust localisation of soccer players using a set of cameras viewing the field from different angles. They set up a complete K-partite graph, with each partition corresponding to one of the K cameras. This is likely to be beneficial for many broader multi-camera applications, including AFL, however for this particular project it was unfeasible to setup cameras at multiple spots around the ground.

Liu et al. [16] introduced a set of Game Context Features (GCF) to describe the current state of play, based on the expected player movements. Using the current track information in combination with the GCF they were able to select a simplified affinity model for each player at any time instant using a random decision forest. Their approach was tested on field hockey and basketball match footage, showing improvements in tracking accuracy by 10% when using the GCF. This type of higher level information abstraction and utilisation may in future help the AFL case, however it is out of the scope of this work.

II. System Overview

Our system is summarised in Figure 2. The main modules are as follows:

- The **detector** finds individual players in each frame and provides a bounding box for each player;
- The **team classifier** examines the contents of each bounding box and determines the team that the player belongs to;
- The **tracker** uses the bounding box positions and team classifications to build paths for each player across the sequence.

We describe the implementation of each of these steps in Sections III, IV and V.

A. *Capturing Footage*

The project was conducted in collaboration with the AFL, who provided access to the Adelaide Oval. Over five matches, approximately 50 hours of footage was acquired, covering a broad range of AFL scenarios. Broadcast footage was not suitable for this work due to the lack of control over the footage such as the large number of cuts, unknown camera positions, fast camera movement and other factors. Instead play was captured with a number of static cameras on tripods. The playing field in AFL is oval shaped and large in comparison to other sports: Adelaide Oval is 167m long and 124m wide[2]. Five full high definition 25fps Axis Q1755 cameras[3] were mounted in a corporate box at the top of the grandstand (Figure 3). The elevation provided by the grandstand is beneficial for occlusion handling as it allows a greater ability to see over and behind players which, when viewed from ground level, would be occluded.

We were constrained to set up all cameras at the same location, providing a single point-of-view overlooking the entire field. The large field of play meant players on the far side only filled a small number of pixels, with heights of around 40 pixels in the worst case. Lower resolution players have fewer descriptive pixels and are hence more difficult to detect and classify, especially in crowded scenarios where players tend to visually merge and meld together. An ideal camera arrangement would be cameras surrounding the field, allowing for multiple angles to overcome occlusion, and play to be relatively close to at least one camera even when far from others.

A range of different camera set-ups were experimented with using different zoom levels, orientations, and focusing on different parts of the ground. However, we found the best approach was to have the cameras in a zoomed out horizontal panoramic formation in order to capture the entire field (Figure 4).

B. *Annotating Footage*

Most object detectors include a training stage to learn the appearance of the object of interest (in this case an AFL

[1] http://chyronhego.com/sports-data/player-tracking
[2] http://www.afc.com.au/news/2014-02-04/oval-retains-unique-size
[3] http://www.axis.com/products/cam_q1755/

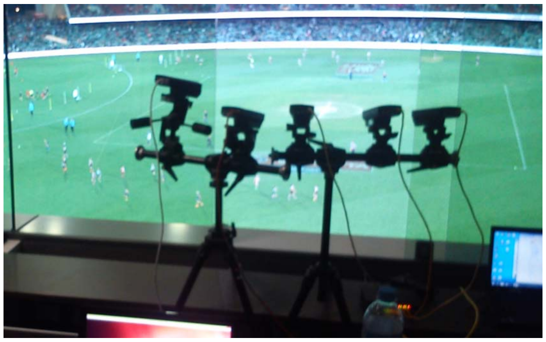


Fig. 3. The five cameras setup on rig with two tripods overlooking field.

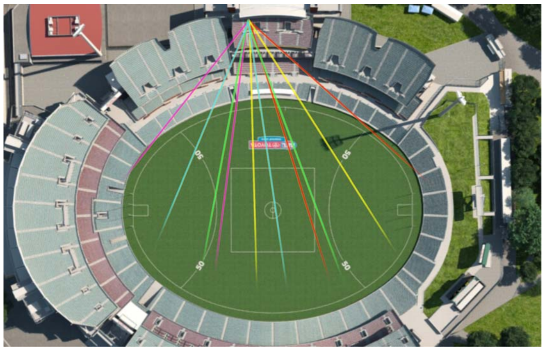


Fig. 4. Camera field of view representation over Adelaide Oval, the five cameras represented by the five coloured triangular shapes. They are generally aligned to overlap slightly to allow for the creation of a panoramic sequence of the entire field to be built.

player). This requires that many examples of players are manually annotated by drawing a box around them. Rather than simply marking players, we also labelled them with their team identity, and whether they were partially occluded. Based on the shape of the bounding box, extreme pose variations such as sitting or lying down were also annotated. This extra information was useful for improving performance in the AFL scenario, where occlusion and pose variation is far more common than for pedestrians. Some examples of how annotations (training samples) are marked are shown in Figure 5.

Training and testing data was gathered from 651 and 188 frames respectively, totalling 13,001 (1,046 occluded) positive training samples and 5,620 (423 occluded) positive testing samples. Negative training samples are collected by randomly selecting areas of the training frames which aren't marked as a positive sample, allowing positive sample overlap of 30% to ensure some difficult negatives.

III. Player Detection

The detection framework implemented in the project is that of Dollár and Appel [17]. This is a sliding window detector that learns an appearance model for the object of interest using the AdaBoost algorithm. A Matlab implementation of the framework is available in their vision toolbox[4] and includes Histograms of Oriented Gradients (HOG) [5] as the default feature. We also use HOG as our feature, as it is the base feature of choice for pedestrian detection problems.

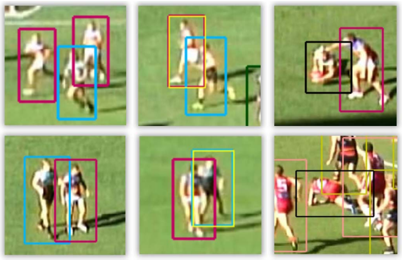


Fig. 5. Annotation examples: Left: Normal annotations where bounding box height is constrained to double the width, players are centred; Middle: Occluded annotations (yellow boxes) when significant part of player is occluded by foreground player; Right: Extreme pose variation (black boxes) where boxes can be any size and ratio to cover a player.

Since the appearance model is learnt from training data, the performance of this detector depends critically on the quality and quantity of available data. We iteratively trained and evaluated using a range of data, with the goal of improving results at each iteration. Over 30 different detectors were trained, each varying in factors such as the number of bootstrapping rounds, number of positive and negative samples, inclusion of positive samples marked with occlusion, feature choice, and other small parameter changes.

Bootstrapping is a means of obtaining a set of informative negative samples [18] during the training process. It involves running the classifier on a new image or sequence and adding all of the false positives to the negative training set. Doing this for one or more iterations helps eliminate the likely false positives. Walk et al. [19] note that the number of bootstrapping rounds is a key component to a detector's performance with at least two rounds necessary for Dalal and Triggs' [5] HOG with linear SVM to achieve its full performance.

A. Evaluation

A number of pedestrian detection and tracking benchmarks are available, including INRIA [5], ETH [20], Caltech [21]. Detection and classification methods are typically evaluated by measuring precision and recall [22] for varying detector thresholds, and we follow this procedure in our evaluations below.

1) INRIA, CALTECH, AFL: The importance of building an AFL training set and using it to build a specialised AFL classifier within the detector is highlighted in Figure 6. Two models, each trained on the INRIA training set [5] and CALTECH training set [21], were compared with a model trained on the AFL data. It is clear from the results that neither of the other models are suitable for the AFL case, with the AFL classifier achieving substantially better results on AFL test data. The INRIA and CALTECH models are inadequate for the AFL case for a number of reasons including:

- the pedestrians are in a distinctly different environment, with large amounts of background and foreground variability;
- the pedestrians have much more limited pose variation than AFL players; and
- the pedestrians are captured from side-on, whereas in the AFL case the camera is elevated.

[4] http://vision.ucsd.edu/~pdollar/toolbox/doc/

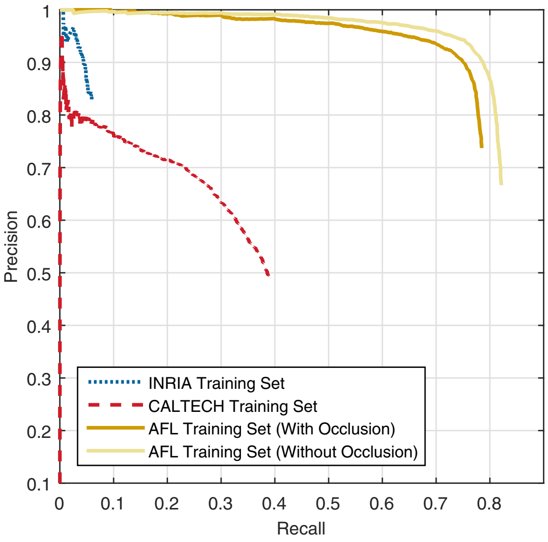


Fig. 6. Precision-Recall curve comparing models trained on different datasets and in the AFL case trained with and without occluded training samples. All results are from testing AFL test data.

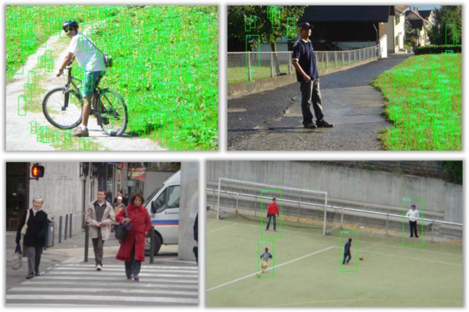
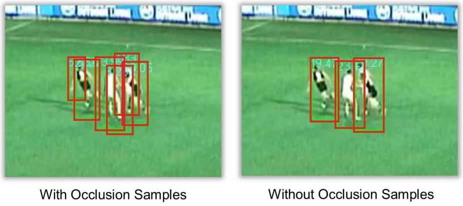


Fig. 7. The AFL detection results on some selected INRIA test images. Fails on all cases except for the soccer image which is reflective of the AFL problem.



With Occlusion Samples

Without Occlusion Samples

Fig. 8. False positives is crowded scenarios when using a detector with occluded samples included in the training data, compared to detector with them excluded.

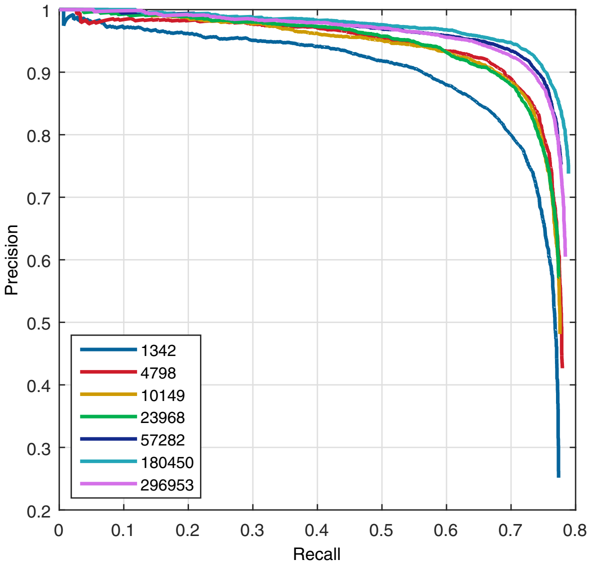


Fig. 9. Precision-Recall curve showing the comparison between accuracy of models trained with different number of negative samples. Number in legend is number of negative samples, number of positive samples kept constant at 11952.

Performing the opposite experiment and evaluating the AFL detector on the INRIA test set further highlights the need for specifically trained detectors. On the INRIA test set, the INRIA trained detector achieves a log-average miss rate of 12.93% but the AFL detector fails almost completely with a miss rate of 99.39%. In almost all test images in the INRIA dataset the AFL detector misses the person completely, and often in green noisy areas of the frame, such as grass and trees it generates many false positives (Figure 7). There is one image however, that the AFL detector handles successfully, and that is of people playing soccer. The soccer image is very similar to the AFL data, with a mostly uniform background, similar pose situations, and the camera position at a similar elevated angle.

2) Occluded Samples: As previously mentioned, during the annotation process certain samples were flagged as being partially occluded. Two detector models were trained each with four rounds of bootstrapping, one with occluded samples included, one with them excluded. Figure 6 shows that excluding occluded samples provides a more accurate classifier. Further observation shows that the classifier trained with occluded samples detects a greater number of false positives in crowded and heavy occluded scenarios (Figure 8). This suggests the model has a lessened ability to distinguish individual players from general patches with high pixel intensity variation.

3) Number of Negative Samples: Positive training samples were manually annotated, whereas the negative samples were chosen randomly from image areas not containing positive samples. We tested the effect of increasing the number of randomly selected negatives, but as seen in Figure 9, results can be varied. Sample numbers over $300,000$ were evaluated however they resulted in no further improvements to accuracy than that of $180,000$.

4) Number of Bootstrapping Rounds: The number of bootstrapping rounds has been shown to alter the performance of classifiers in general [19]. Figure 10 shows that for the AFL case, performing at least one round of bootstrapping improves the accuracy of the classifier, however undertaking more than one round achieves no further significant increases. Further inspection reveals that despite more than one round decreasing the number of false positives provoked by on-field signage, it also causes many partially occluded positive samples to be missed. This result is likely due to the combination of these

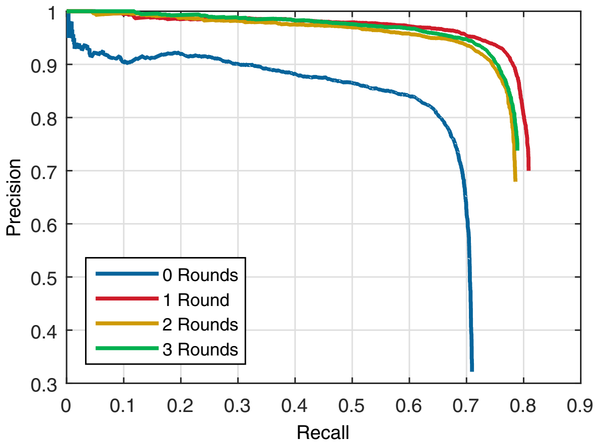


Fig. 10. Precision-Recall curves comparing the effects of differing the number of bootstrapping rounds on the classifier model.

classifiers being trained without occluded ground truth and the overlap degree of 30% which was allowed between positive ground truth and randomly sampled negatives from the same frames.

IV. Team Classification

The task of the team classification step is to assign each detected player to one of five classes: team A, team B, umpire, runner, or other. This is typically done by calculating a fixed length feature vector for each detection, and then using the vector as input for a classifier.

AFL teams (18 currently), as well as umpires, runners, and other officials, all have their own specific uniforms, each comprised of certain colours and patterns. However, variation in player pose as well as the low resolution of players on the far side of the field leads to great variation in the appearance of uniform patterns. Therefore, the classification feature needs to be robust to partial and low quality images of a player. We used a colour histogram feature, as it can operate on low resolution images and does not require the entire uniform to be visible.

It was observed that most of the pixels in a typical detection bounding box don't represent the player, with even less representing the uniform. In fact, it was estimated that the area of the detection box containing pixels representing the uniform can range from only 5%-15% (Figure 11). This suggested a weighted mask to focus the descriptor vector on the area of the bounding box corresponding to the uniform. Numerous different masks were evaluated, but all have a similar structure with the main focus on the upper middle section of the detection box, around where the uniform is expected to appear. Each weight is simply one or two 2D Gaussian functions:

$$f(x,y) = A\exp\left(-\left(\frac{(x-x_o)^2}{2\sigma_x^2} + \frac{(y-y_o)^2}{2\sigma_y^2}\right)\right) \quad (1)$$

with different values for the amplitude (A), the centre point (x_o, y_o), and spread in each direction (σ_x, σ_y).

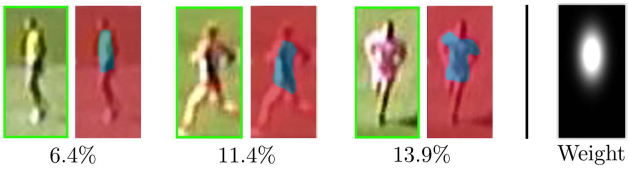


Fig. 11. Three examples of the percentage area size of the pixels representing uniforms relative to entire bounding boxes and the spatial weight found to achieve the best accuracy.

The descriptor vectors were constructed by histogramming the pixel intensities into 64 bins for each colour channel. Histogramming was used to lower the dimensionality of the descriptor vector and also to apply some form of smoothing over the individual colour values by grouping similar values together. Both the RGB and HSV colour formats were tested. Often HSV is used in colour classification tasks because it separates luma (image intensity), from chroma (colour information), generally providing greater discriminative power. When histogramming into the bins, the weights are applied based on pixel position, with pixels that likely refer to the uniform having a greater impact on the histogram.

A support vector machine (SVM) was used as the classifier, because of its speed, simplicity and discriminative power. The same training data used for the detector module was also used to train the team classifiers, with occlusion training samples excluded. A quadratic kernel was used, as a linear kernel was observed not to converge. Each descriptor vector is plotted into a $b * c$ dimensional space, where b is the number of bins, and c is the number of colour channels ($64 * 3 = 192$). SVM models were trained for each individual team, that is, for each model all training samples are labelled as negative except for those from one team. A team is assigned to a test sample by evaluating the sample with all of the SVM models related to the match the sample is from. If a single model produces a positive result then the sample is classified as the team corresponding with that model, however if more than one, or no models give a positive result then the sample is classified as unknown.

A. Evaluation

1) RGB versus HSV Colour Formats: RGB and HSV classification models were trained and tested each achieving a mean average precision of 84.31% and 90.37% respectively. As expected, the HSV colour format has significantly more discriminative power for the AFL scenario. The separation of the value channel from the hue and saturation channels provides more robustness to lighting changes and shadows, with value changing while hue and saturation remain rather constant.

2) Different Teams versus Different Environments: Figure 12 presents the classification testing results for each of the individual teams. While there is some difference between the success rate for each team, it is clear that the environmental conditions have more of an effect than the actual team uniform. Although teams with greater contrasting colours are classified more accurately than teams with similar uniforms, the environmental conditions are more significant. It can be seen for all of the night captured teams, although there is some variation in accuracy, all classifiers perform better than any used during the sunny conditions where overexposure can have negative effects.

3) Teams or Match Based Classifiers: The impact the environmental and lighting conditions have on the classification accuracy for any team is significant, and suggests that performance may improve with models tailored for specific lighting conditions. Experimenting with this hypothesis, two sets of models were trained. Firstly models trained on a per team basis with training data from any match the team appeared in, and secondly models trained on a per team per match basis with training data from only the match which is being tested. The mean average precision results for the per team models were 85.98% and for the per team per match models were 90.37%. Even though the training data for the match specific team classifiers was only a subset of the data used to train the more generalised classifiers, the classifiers were more suited for the particular match conditions. This highlights the need for classifiers trained on appropriate training data rather than just more training data.

V. Player Tracking

Two tracking approaches were compared in this work, a local Kalman Filter implementation and a global energy minimisation algorithm proposed by Milan et al. [11]. For both trackers, each track is assigned a team based on the maximum number of classifications related to the detections attributed to that track.

For a given frame, the Kalman Filter algorithm estimates expected position of all current tracks based on track velocities estimated from past frames. Unassigned detections are associated to tracks greedily using Euclidean distance from the predicted position within a 20 pixel radius. Tracks that are not assigned to detections from the previous frame are tested against detections from up to 5 frames earlier. If after this process a detection isn't associated with a track, it may be either a new target entering the frame, a previously lost target, or false positive detection, and it is used to initialise a new track. At the end of the process tracks lasting for less than 20 frames, expected to be false positive detections, are removed from the solution.

The global approach uses a discrete continuous energy minimisation technique to perform the data association and trajectory calculation. This approach was shown to have some promising results on the challenging benchmarks PETS 2009/2010 [23] [24] and has some desirable properties for this project. The global approach explicitly handles partial and full inter-object occlusion, and has natural inclusion of per-frame detection evidence, appearance, dynamics, persistence, and collision avoidance. A Matlab implementation is available online for download[5].

A. Evaluation

Only recently have quantitative evaluation methodologies for multi-target tracking been proposed and used in literature to benchmark methods. One technique, CLEAR MOT, was developed by Stiefelhagen et al. [25] [26] and is composed of two metrics, Multi-Object Tracking Accuracy (MOTA) and Multi-Object Tracking Precision (MOTP). However, the AFL problem is very different from pedestrian tracking, and also very specific in its goals. Therefore it is not evaluated on pedestrian tracking benchmarks, but rather empirically only in the AFL domain.

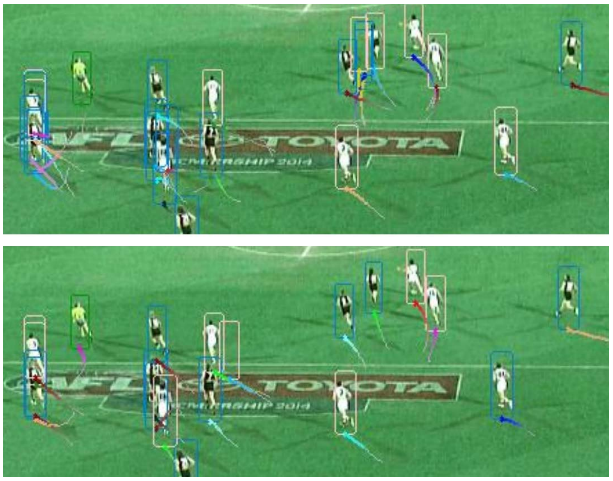


Fig. 13. Empirical comparison between frames from the original energy minimisation configuration (top) against frames from the tuned configuration (bottom).

1) Parameter Tuning: The energy minimisation approach has fifteen independent parameters to set and little documentation of the purpose and effect of each. Each parameter was modified independently and the effects were systematically analysed. The final global tracking configuration contains five modified parameters which penalise high numbers of tracks and attempts to separate close tracks. Firstly the spatio-temporal threshold was increased from 10 to 15 to allow for greater movement of targets between frames, as often for the AFL scenario players move relatively fast when compared to pedestrians. Secondly the outlier cost and label cost were changed from 200 to 100 and 200 to 250 respectively, to restrict the number of tracks yet still allow reasonable outliers to be captured. Specifically the increase in label cost prevents large numbers of false positives being proposed in the highly congested areas. Lastly the unary factor and pairwise factor were both modified from .1 to .14 and 100 to 15 respectively, to separate close detections into different individual tracks, which is necessary to handle the regular close proximity of players. The greatest difficulty was finding the balance between having one track for each player while still maintaining two tracks for nearby and crossing players without erroneous merging. Figure 13 provides comparisons between frames from the original configuration with frames from the tuned configuration. The effects of the parameter tuning are minor and dependent on the particular scenario. For the case below there are many less false positives, 10 in the original and 4 in the tuned, both have 15 true positives.

2) Kalman Filter VS Energy Minimisation VS Combination: The Kalman Filter tracker was adequate for tracking multiple isolated targets. However it failed often in circumstances where the detection boxes became more noisy and spatially spread over time, such as when players moved quickly or

[5] http://www.milanton.de/files/software/dctracking-v1.0.zip

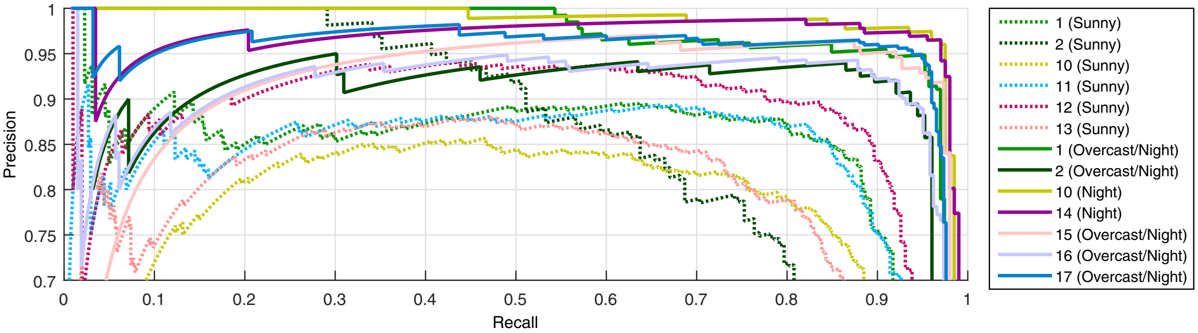


Fig. 12. Precision-Recall curve of team classification accuracy for each team in each condition. Numbers represent different teams, dotted lines represent sunny conditions, and solid lines represent overcast or night conditions.

when boxes disappear or cover multiple mutually occluding players. These factors meant that the Kalman Filter approach regularly terminated and initiated relatively short tracks for targets. Using the global energy minimisation approach to refine the Kalman Filter merged many of the short tracks into longer continuous tracks. It also enabled better recovery from occlusion, maintaining tracks across frames where detections disappear temporarily. The energy minimisation method was also able to be utilised on its own, but in congested parts of the scene many false positives would ‘float’ erratically over the busy areas latching on to various detections representing many different players. This erroneous behaviour is partly attributed to the tracker having difficulty distinguishing close targets, and also partly to the very noisy detections resultant from the detectors inability to handle extremely congested areas.

Figure 14 is an example of some of these scenarios for each variation of the tracker. At position (a) in the energy minimisation approach the erratic and unrealistic switching and sliding of tracks can be seen. However with the use of the initial Kalman Filter solution these problematic cases don’t arise and don’t form part of the combined solution. At position (b) in the Kalman Filter approach, the track for the darker player on the right has only just been initialised as that player just passed behind the umpire to his left in previous frames, causing his track to be terminated and re-initialised. In the combined tracker the two shorter tracks get joined into a longer track that exists constantly as the player passes behind the umpire. The players in (c) present the general effect of the initial Kalman Filter solution on the energy minimisation approach. The Kalman Filter solution is restricted to zero or one track per player at any point in time, whereas the energy minimisation approach uses one or more tracks. Using the Kalman Filter as an initialisation to the energy minimisation method restricts the latter from building too many false positive tracks.

B. Runtime Analysis

The runtimes for each of the modules in the pipeline are shown in Table I. All runtime analysis was performed on a 64-bit desktop running Windows 7 with a Intel i7-4790U CPU at 3.6GHz and 16GB of RAM. The detector module runtime depends on the number of sample windows evaluated in the classifier, which is dependent on the frame dimensions, as well as the specified window padding and scaling steps. With the use of the fast SVM model the majority of the time in the team classification stage is spent on feature calculation. For the tracker options, the Kalman Filter is a clear winner in regards to runtime efficiency, also aiding in speeding up the energy minimisation approach when used as an initialiser.

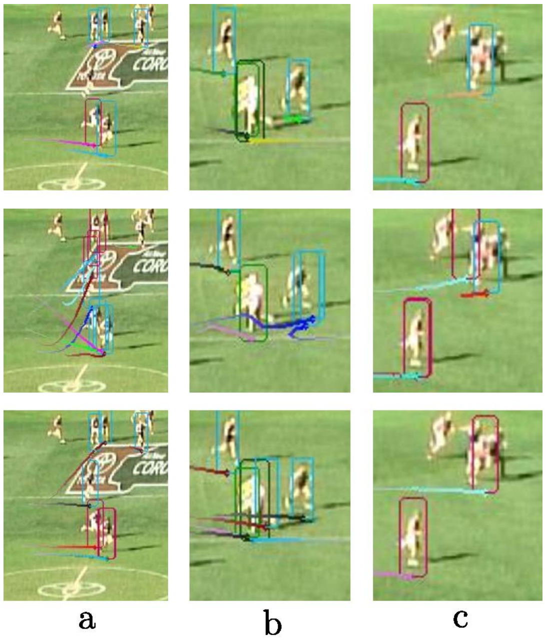


Fig. 14. Empirical comparison between frames from the local Kalman Filter approach (top), the global energy minimisation approach (middle), and the combination of both (bottom).

VI. Conclusion

By systematically experimenting, refining and evaluating different detection, classification and tracking approaches, this paper has adapted techniques from pedestrian detection and tracking to work for the more challenging AFL scenario. The method is tailored to handle many difficult AFL situations not common in pedestrian tracking, including heavy occlusion, varied pose, fast motion and low image resolution. This framework provides a good foundation for future development of higher level information extraction, such as collection of statistics and match analysis.

		Detector			Team Classifier						Tracker								
Frames: 1000					**Feature Calculation**			**SVM Classification**			**Kalman Filter**			**Energy Minimisation**			**KF then EM**		
Dataset	**Detections**	**Time**	**FPS**	**DPS**	**Time**	**FPS**	**DPS**	**Time**	**FPS**	**DPS**	**Time**	**FPS**	**DPS**	**Time**	**FPS**	**DPS**	**Time**	**FPS**	**DPS**
R03Q3C5	**25232**	210	4.76	120	1828	.547	13.8	6.21	161	4063	16	62.5	1577	32459	.031	.777	216	4.63	117
R04Q3C2	**21737**	208	4.81	105	1156	.865	18.8	5.51	181	3945	12.9	77.5	1685	13552	.074	1.6	174	5.75	125
R07Q4C2	**2696**	197	5.08	13.7	203	4.93	13.3	.518	1931	5205	.759	1318	3552	334	2.99	8.07	17.2	58.1	157
R12Q3C1	**22507**	207	4.83	109	1307	.765	17.2	5.25	190	4287	13.2	75.8	1705	20766	.048	1.08	181	5.52	124
R14Q4C4	**5840**	288	3.47	20.3	329	3.04	17.8	.993	1007	5881	2.12	472	2755	1359	.736	4.3	37.1	27	157

TABLE I. RUNTIMES OF THE EACH MODULE IN SECONDS, FRAMES PER SECOND (FPS) AND DETECTIONS PER SECOND (DPS). ALL DATASETS ARE 1000 FRAMES LONG. NOTE BOTH STEPS OCCUR IN THE TEAM CLASSIFIER, WHEREAS ONLY ONE OF THE TRACKER METHODS NEED TO BE RUN.

We found that it was vital to train the detector using manually labelled AFL ground truth data, with off-the-shelf pedestrian detectors failing to achieve reliable performance. Team classification based on HSV colour using spatial weights and an SVM classifier was found to be sufficient to separate teams with over 90% accuracy for most environmental conditions, although the prevailing condition (sun/cloud/night) was found to have a greater effect on accuracy than the actual team uniforms. The two very different tracking approaches each have their own strengths and weaknesses, with the fast local Kalman Filter approach creating relatively accurate short tracks while the slow global energy minimisation technique creates longer tracks but requires initialisation. A combination of the two approaches performed better than either individually.

We will continue to work on this project in collaboration with the AFL. Future work includes improving the efficiency of individual components, with a view to real-time performance so that match statistics can be live delivered during a game.

ACKNOWLEDGMENT

This work was partially funded by ARC Linkage Grant LP130100154 and Champion Data.

REFERENCES

[1] N. Schmook, "Swans' surge drives new AFL club membership record," http://www.afl.com.au/news/2015-08-26/swans-surge-drives-new-afl-club-membership-record, 2015, [Online; last accessed 28-August-2015].

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2. IEEE, 2000, pp. 142–149.

[3] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 878–885.

[4] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[6] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1515–1522.

[7] ——, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1820–1833, 2011.

[8] S. Avidan, "Ensemble tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 2, pp. 261–271, 2007.

[9] Y. Guan, X. Chen, D. Yang, and Y. Wu, "Multi-person tracking-by-detection with local particle filtering and global occlusion handling," in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–6.

[10] A. Milan (né Andriyenko), K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1926–1933.

[11] A. Milan (né Andriyenko), S. Roth, and K. Schindler, "Continuous energy minimization for multi-target tracking," *Pattern Analysis and Machine Intelligence*, 2013.

[12] P. Nillius, J. Sullivan, and S. Carlsson, "Multi-target tracking-linking identities using bayesian network inference," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2187–2194.

[13] K. Okuma, D. G. Lowe, and J. J. Little, "Self-learning for player localization in sports video," *arXiv preprint arXiv:1307.7198*, 2013.

[14] R. Hamid, R. K. Kumar, M. Grundmann, K. Kim, I. Essa, and J. Hodgins, "Player localization using multiple static cameras for sports visualization," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 731–738.

[15] R. Hamid, R. Kumar, J. Hodgins, and I. Essa, "A visualization framework for team sports captured using multiple static cameras," *Computer Vision and Image Understanding*, vol. 118, pp. 171–183, 2014.

[16] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1830–1837.

[17] P. Dollár, "Piotr's Image and Video Matlab Toolbox (PMT)," http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html.

[18] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 39–51, 1998.

[19] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 1030–1037.

[20] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 304–311.

[22] "Precision and recall," https://en.wikipedia.org/wiki/Precision_and_recall.

[23] A. Ellis, A. Shahrokni, and J. M. Ferryman, "Pets2009 and winter-pets 2009 results: A combined evaluation," in *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*. IEEE, 2009, pp. 1–8.

[24] J. Ferryman and A. Ellis, "Pets2010: Dataset and challenge," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. IEEE, 2010, pp. 143–150.

[25] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," in *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 1–44.

[26] B. Keni and S. Rainer, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2008.