Multi-target tracking in team-sports videos via multi-level context-conditioned latent behaviour models

Jingjing Xiao[1]
shine636363@sina.com

Rustam Stolkin[2]
r.stolkin@bham.ac.uk

Ales Leonardis[3]
a.leonardis@cs.bham.ac.uk

[1] School of Electronics, Electrical and Computer Engineering, University of Birmingham, Birmingham, UK

[2] School of Mechanical Engineering, University of Birmingham, Birmingham, UK

[3] School of Computer Science, University of Birmingham, Birmingham, UK

Abstract

Multi-target tracking techniques increasingly exploit contextual information about group dynamics. However, approaches established in pedestrian tracking make assumptions about features and motion models which are often inappropriate to sports team tracking, where motion is erratic and players wear similar uniforms with frequent inter-player occlusions. On the other hand, approaches designed specifically for sports team tracking are predominantly aimed at detecting game-state rather than using game-state to enhance individual tracking. We propose a multi-level multi-target sports-team tracker, which overcomes these problems by modelling latent behaviours at both individual and player-pair levels, informed by team-level context dynamics. At the player-level, targets are tracked using adaptive representations, constrained by probabilistic models of player behaviour with respect to collision avoidance. At the team-level, we exploit an adaptive meshing and voting scheme to predict regions of interest, which inform strong motion priors for key individual players. Thus, latent knowledge is derived from team-level contexts to inform player-level tracking. To evaluate our approach, we have developed a new data-set with fully ground-truthed team-sports videos, and demonstrate significantly improved performance over state-of-the-art trackers from the literature.

1 Introduction

Multi-target tracking remains a significant open problem in computer vision. Many applications involve video surveillance and pedestrian tracking, but there is a growing interest in the automatic tracking of players in team sports, e.g. for automated sports commentary [7].

Initially, progress in pedestrian tracking was mostly due to improved target models: generic appearance models or detectors, dynamic motion models, and better optimization

strategies [3]. More recently, pedestrian tracking has advanced considerably by formulating multi-target tracking in terms of data association. Independent motion models have been implemented to limit the complexity of solutions [4], while inter-correlations have also been studied between pedestrians using context [2].

Team sport tracking poses three main difficulties [7] which distinguish it from conventional pedestrian tracking: 1) players in team uniforms share similar appearance; 2) motion changes are abrupt and erratic in contrast to comparatively smooth and linear pedestrian motions; 3) the motions of two distant players may be highly correlated. Appearance models alone are not usually sufficient to distinguish team-sport players. Player motion itself can be discriminating, however players often undergo abrupt motion changes when gaining/losing possession of the ball. Fortunately, the latent team strategy can provide additional hints for tracking, i.e. seemingly erratic motion of a single player may actually be consistent with the overall motion strategy of the team.

Some pedestrian tracking literature, [3], has exploited models of dynamic social behaviour to enhance tracking. However, [9] does not explicitly consider the problem of person-person occlusion which is of central importance to team-sports tracking. Recent work on sports videos [6] uses multi-player motion fields to enable dynamic scene analysis in football games. However, this work is aimed at extracting a high level understanding of the game-play from player trajectories, rather than using this knowledge to enhance the tracking itself. Additionally, while the flow-field approach of [6] works well for football, we find that it breaks down in other sports such as volley-ball, where player-player correlations are more complex, yielding highly non-smooth and dis-continuous flow fields. Other recent work, [7], combines tracklet analysis with higher-level context-conditioned team dynamics to achieve impressively robust data-association. However, this approach of best-fitting a series of tracklets to a segment of video-footage is designed for offline video-analytics, rather than real-time, online frame-by-frame tracking. Related works, [2], [12], also exploit various kinds of contextual constraints of object motions, however these do not extend to models of player decision-making and its influence on trajectory priors.

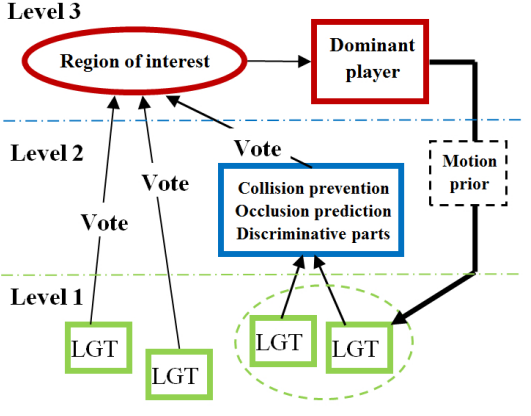

Figure 1: Multi-level tracking algorithm. Level 1: each player tracked by [3]. Level 2: player-player occlusions handled by player-pair behaviour model. Level 3: group or team-level context-dynamics gives dominant player trajectory prediction.

We propose a new multi-target tracker which functions on multiple levels or scales of representation, shown in Fig.1. At the lowest level (Level 1), we track individual players

using the state-of-the-art LGT "Local-Global" tracker [3].This, itself involves two "layers" of tracking: a parts-based set of "local" patches (based on intensity distributions), and a "global" target model (incorporating motion, shape and colour distributions).

The LGT player models (Level 1) are next augmented by an additional model at the player-pair level (Level 2), which encodes the motion preferences of two or more players in close proximity, in the form of a probability distribution representing their tendency to avoid collisions. The pair-wise collision-avoidance model is used to modify the local patch models and global target models of a target pair: the global motion model is modified by the collision avoidance model, providing a stronger motion prior; a prediction is made about which local patches will be occluded during the pair-wise player interaction; and remaining patches are weighted according to their predicted discriminative power during such interactions.

We next examine the motion of multiple players at the team-level (Level 3). Based on player positions, provided by the lower tracking levels, we propose an adaptive approach to meshing the playing area in which the mesh resolution scales appropriately with player density. A player-voting method is then proposed which computes a region of interest (ROI), based on the distribution of player locations and their individual velocities. The region of interest does not necessarily indicate the ball position, but may equally indicate the future ball position, or some other position of strategic importance, as predicted by the players. Using this information, it is possible to select one or more "dominant" players, who tend to move with a clearly identifiable trajectory towards the ROI, with a high degree of confidence. Since most instances of pair-wise inter-player occlusions involve such a dominant player, the dominant player trajectories can be used to re-learn the probabilistic motion models used by the intermediary pair-wise tracking level, as described above. The adaptive mesh-scaling ensures that the resolution of such trajectory predictions scales proportionately to the predictive confidence. The main contributions of this paper are:

1) A probabilistic model of collision-avoidance motion strategies at the player-pair level (Level 2), continuously re-learned online from latent information, context-conditioned by large-scale motion at the team-level. This is used to augment the global-layer motion-models of the LGT, and also to modify the local-layer of the LGT trackers, by predicting occluded parts, and adaptively weighting each target part according to its discriminating power.

2) A method for identifying an ROI, and one or more dominant players, whose trajectories can be predicted with confidence relative to the ROI. These dominant player trajectories are then used to condition the collision-avoidance models.

Our method does not rely on training data, background models, any camera calibration, or "tracking-by-detection". Instead it is initialised merely from a bounding box for each target in the first frame.

2 Multi-level multi-target tracker

Our multi-target tracker comprises three main levels of reasoning (Fig.1): a player-level based on a state-of-the-art tracker LGT [3] (itself comprising two sub-levels called local and global "layers"); a local group-level which models the behaviour of player-pairs; and a global group-level which reasons about team dynamics.

2.1 Individual player level (Level 1)

At the level of individual targets, our tracker is based on LGT [3] which is a high-performance state-of-the-art tracker, with publicly available code and evaluation data [1]. LGT uses an adaptive two-layer target representation. These local and global layers each provide constraints for re-learning the other, which enables stable adaptation.

An individual tracker, T_k, represents a target as a set of global properties: location L_k, motion M_k, and appearance ζ_k. The target location is represented by a bounding box $L_k = (x_k, y_k, w_k, h_k)$, where x_k, y_k are coordinates of the top-left corner, and w_k, h_k denote the width and height respectively. Motion $M_k = (\dot{x}_k, \dot{y}_k)$ denotes velocity of the player's centroid.

The global appearance model of an individual target can be re-learned from an image region, defined by set of local parts or patches, which form the local layer of the model. We later show how to modify these local patches to overcome player-player occlusions. The global layer provides a 2D distribution over image locations for continually allocating and learning new local patches:

$$p(T_k|L_k, M_k, \zeta_k) \propto p(T_k|L_k)p(T_k|M_k)p(T_k|\zeta_k) \tag{1}$$

The set of N_p local patches $U_k = \{u_k^{(i)}\}_{i=1:N_p}$ are local distributions of image measurements, ζ_k, with weights:

$$W_k^{(i)} = p(\zeta_k, u_k^{(i)}|U_k) = p(\zeta_k|u_k^{(i)})p(u_k^{(i)}|U_k) \tag{2}$$

2.2 Local group-level (Level 2)

A pair of sports players may share similar appearance and interact in close proximity with each other, making data association very difficult [12]. Our method addresses these problems by using latent player behaviour models to improve the accuracy of motion models and the discriminative power of appearance models.

2.2.1 Collision avoidance motion model

Humans typically plan future movements that avoid collisions. Even during contacts, two players do not physically coalesce, so that a collision avoidance assumption remains effective. Our collision-avoidance model detects instances of player-pairs, defined by proximity:

$$\ell(T_k^m, T_k^n) = \begin{cases} 1, \text{if } \left\|L_k^m - L_k^n\right\|_2 < l_p \\ 0, \quad \text{otherwise} \end{cases} \tag{3}$$

where l_p is scaled according to the size of the region around each target that the tracker searches at each frame. Thus, a pairing detection indicates a high risk of the two player's trajectories intersecting in the image plane, resulting in occlusion. However: in all sports, two trajectories can never physically merge on the 2D ground plane; in many sports, players often deliberately modify their motion to avoid collisions, especially "dominant" players (e.g. in possession of the ball) who wish to avoid interceptions.

We use a second order motion model, which assumes: i) each player predicts that the other player will continue with a uniform velocity; and, ii) each player will reduce that component of his/her own velocity which lies in the other player's direction to zero before an expected collision at future time $k + \Delta k$. Thus, the n^{th} player will modify their present

speed V_k^n by a deceleration, a_k^n, to arrive at a modified speed $\hat{V}_k^n = V_k^n - a_k^n$, in response to the m^{th} player. This yields two constraints:

$$\Delta k V_k^m + \left(\Delta k V_k^n - \frac{1}{2} a_k^n \Delta \mathrm{k}^2\right) = L_k^n - L_k^m, \qquad V_k^n + V_k^m - \Delta k a_k^n = 0 \tag{4}$$

which are solved for two unknowns, Δk and a_k^n, giving:

$$a_k^n = (V_k^n + V_k^m)^2 / (2L_k^n - L_k^m) \tag{5}$$

Predicted player decisions, a_k^n, a_k^m for each member of a player-pair enable enhanced motion predictions, $\hat{V}_k^n$, $\hat{V}_k^m$ for potential occlusion situations. We use these predicted velocities to estimate the future position of each patch of each interacting player. Collision avoidance behaviour can now be represented probabilistically as a motion prior:

$$p\left(V_k | u_k^{(i)}\right) = \begin{cases} 1 - e^{\left(-\lambda_M d_k^{(i)} / l_p\right)}, & \text{if } d_k^{(i)} < l_p \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

which can be incorporated into Eq.(2) via a product as: $\hat{W}_k^{(i)} = W_k^{(i)} p\left(V_k | u_k^{(i)}\right)$. This motion prior leaves probabilities unmodified for patches which are predicted to remain un-occluded, but devalues the weighting assigned to likely occlusion patches as a function of their predicted distance $d_k^{(i)}$ from the occluding player's centroid. l_p is the same parameter which is used in Eq.(3). λ_M is a constant.

2.2.2 Predictive enhancement of appearance models

The augmented motion model can be used to predict future positions of each target by: $\hat{L}_k = L_k + \hat{V}_k$. We use this to detect imminent occlusion situations and predict which target-parts will become occluded. Predicting occluding and occluded states of player-pairs is relatively easy in team-sports for two reasons: i) sports are predominantly played on flat surfaces, with homographic mapping to the camera's image plane; ii) television cameras, regardless of view-point, pan or tilt-angles, never have an appreciable roll-angle, i.e. the image-plane x-axis is always parallel to the plane of play. Therefore relative proximity to the camera of two targets can be determined merely by comparing their y-coordinates. The occluded parts of the more distant target can now be computed as: $\hat{R}_k^{mn} = \hat{R}_k^m \cap \hat{R}_k^n$, where $\hat{R}_k^m$ and $\hat{R}_k^n$ is the occupied 2D region of player m, n respectively. $\hat{R}_k^{mn}$ denotes the overlap between the player-pair, which ranges from 0-100%.

If part $u_k^{(i)}$ is occluded, its weight $\omega_k^{(i)}$ is set to zero. Once more than 60% of a target becomes occluded, appearance features become unreliable, so we rely solely on the motion model for target propagation. During occlusions, it is critical to focus attention on whichever parts of the player models are most mutually discriminating. We therefore compute a dissimilarity score $p\left(S_k | u_k^{(i)}\right)$ for each visible player-patch, $u_k^{(i)}$, compared to all patches $\left\{v_k^{(j)}\right\}_{j=1:N_v}$ of the other interacting player:

$$p\left(S_k | u_k^{(i)}\right) = \frac{\psi\left(u_k^{(i)}\right)}{N_v} \sum_{j=1}^{N_v} \left[1 - \rho\left(c\left(u_k^{(i)}\right), c\left(v_k^{(j)}\right)\right)\right] \tag{7}$$

where $c(.)$ is the normalized colour histogram of a patch; $\rho(.)$ is the Bhattacharyya distance [10]; $\psi\left(u_k^{(i)}\right)$ labels occluded and un-occluded parts:

$$\psi\left(u_k^{(i)}\right) = \begin{cases} 0, & \text{if } u_k^{(i)} \in \hat{R}_k^{mn} \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

2.2.3 Combining appearance and motion predictions

We now explain how to combine the predictively enhanced appearance model, Sec.2.2.2, with the collision-avoidance motion-model, Sec.2.2.1. We do this by replacing the 2D target distribution $p\left(\zeta_k|u_k^{(i)}\right)$ of Eq.(2) by:

$$p\left(\zeta_k, S_k, M_k|u_k^{(i)}\right) = p\left(\zeta_k|u_k^{(i)}\right) p\left(S_k|u_k^{(i)}\right) p\left(M_k|u_k^{(i)}\right) \tag{9}$$

where $p\left(M_k|u_k^{(i)}\right)$ is computed from Eq.(6) and $p\left(S_k|u_k^{(i)}\right)$ is computed from Eq.(7).

2.3 Global group-level (Level 3)

The performance of single-target trackers can be improved by making use of high-level knowledge of target behaviour. Analogously, multi-target tracking in sports videos can be improved by utilising semantic-level understanding of the game. In our global player-level model, we extract latent information about team dynamics and use it to inform motion-priors for dominant players.

2.3.1 Detecting a region of interest

We now present a concept which we call the "region of interest", or ROI. Unlike other work [2][8] this does not mean "focus area" or the estimated ball position in the current frame. Instead it is a more general concept, referring to a location which the players believe will be of strategic importance in the near future. This may include the present ball position, predicted future ball position, or some other region of imminent strategic importance according to a semantic-level understanding of the game. Players tend to move towards, or distribute close to, the ROI. Therefore, the ROI can provide prior information about future player motions. To detect the ROI, we first form a player flow-field for the image, defined as: $F_k = \left\{\bar{V}_k^1, \ldots \bar{V}_k^j\right\}$ where (to reduce noise)$\bar{V}_k^j = \frac{1}{m}\sum_{t=k}^{k-m} V_t^j$ is the m-frame moving average of the velocity V_k^j of player j at frame k.

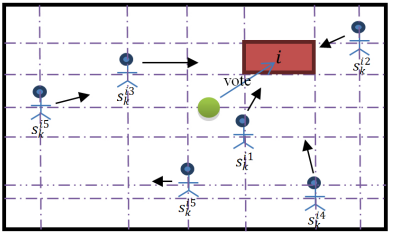

Figure 2: forming mesh according to players' distribution. Green circle: centre of players' distribution; Red region: potential region of interest.

Next, we adaptively mesh the image according to the current set of player positions, Fig. 3, by intersecting each player with gridlines in each axis. This simple mesh construction has a useful adaptive property: when players are spread far apart, the mesh is coarse, and will yield a low resolution estimate for the ROI position, which reflects the lower accuracy ROI estimate that can be achieved from a loose player distribution. In contrast, a high local player density produces accurate ROI estimates, reflected by the fine resolution mesh.

It is observable, in many sports, that a player's motion is less correlated with the ROI if their position is far away from it. While players local to the ROI usually move towards it, more distant players may choose to move away from it to ensure that gaps in the court or playing field are filled. We model this behaviour by assigning weights for each player's vote for each possible cell of the mesh, according to their distance: $\omega_k^{ij} = e^{-\lambda d^{ij}}$, where ω_k^{ij} is the weight of the j^{th} mesh cell; d^{ij} is the distance from the player to the cell; λ is a constant. The j^{ij} player's weighted vote for the i^{th} cell is now set proportional to its velocity towards that cell:

$$s_k^{ij} = \omega_k^{ij} \bar{V}_k^i \left(L_{cell}^i - L_k^j \right) / \left| L_{cell}^i - L_k^j \right| \tag{10}$$

where L_{cell}^i is the location of the i^{th} cell and L_k^j is the location of player j at frame k. Since players tend to cluster around the ROI, their distribution provides an additional cue to the ROI location. Therefore, we also allow the player's group centroid to contribute a vote: $s_k^{ic} = e^{-\lambda^c d^{ic}}$, where d^{ic} is the distance between the i^{th} cell and the players' centroid and λ^c is a constant parameter. The overall ROI vote for the i^{th} cell from N_p players is: $S_k^i = s_k^{ic} \sum_{j=1}^{N_p} s_k^{ij}$. The ROI is now selected as that cell with the largest vote.

2.3.2 Detecting and modelling dominant players

We now introduce the concept of dominant players, defined as those closest to and/or moving rapidly towards the ROI. We observe that dominant player motion is more stable and predictable than that of other players, who tend to adjust their own motions in response to dominant player actions. We also observe that most player-player occlusions involve a dominant player. Hence, dominant players yield high confidence motion-priors, which we use to overcome such occlusions. We detect dominant players as those with high dominance score:

$$D(i) = \begin{cases} e^{\left(\bar{V}_k^{iROI} - \lambda_{ds} d_i\right)}, \text{if } \overline{V}_k^{iROI} > 0 \\ 0, \qquad \text{otherwise} \end{cases} \tag{11}$$

where $\bar{V}_k^{iROI}$ is the component of player velocity $\bar{V}_k^i$ in the ROI direction, d_i is distance between player and ROI, and λ_{ds} is a constant parameter.

Once detected, dominant player motions are modelled probabilistically:

$$p\left(C_k^{dom} | u_k^{(i)}\right) = 1 - e^{-\left(\lambda_d / \left| V_k^{p\_i} - \bar{V}_k \right|\right)} \tag{12}$$

where C_k^{dom} represents confidence in a candidate location for a dominant player's i^{th} patch, $u_k^{(i)}$, given the patch-velocity $V_{k+1}^{p\_i}$ which would have caused this location, and its discrepancy with the overall player motion. $\bar{V}_k$ is smoothed velocity. λ_d is a constant which controls sensitivity to motion discrepancies.

An important additional consideration is the persistence of a player's dominant status, and the reliability of the dominant motion model, which decreases with time. We handle this with an auto-regressive persistence factor $T_d \propto \sum_{i=k-m}^{k} \delta(i)$, where $\delta(m) = 1$ if the player is detected as dominant at frame i and zero otherwise. The ideal constant of proportionality may depend on the characteristics of a particular sport, but we find that a value of 2 works well for volleyball and football. T_d decreases with time, so that the impact of the dominant motion prior disappears at $T_d = 0$. We can now modify Eq.(9) to include dominant player dynamics, to evaluate candidate patch locations:

$$p\left(\zeta_k, S_k, M_k, C_k^{dom} | u_k^{(i)}\right) = \left[p\left(C_k^{dom} | u_k^{(i)}\right)\right]^{w_j} p\left(\zeta_k, S_k, M_k | u_k^{(i)}\right) \tag{13}$$

where ω_j adjusts the impact of the dominance component of motion prediction as: $\omega_j = 1 - e^{-\lambda_t T_d}$, where λ_t is a constant parameter.

3 Experiments

We evaluate our method on ten different volleyball videos, to show its robustness to extremely challenging scenes. We also evaluate it on a football sequence, to demonstrate the generality and transferability of the models to different kinds of sports.

We have chosen BPF, [8], for comparison with our method, because it is an award-winning tracker designed specifically for team-sports videos. It combines a state-of-the-art Adaboost detection method with a Particle Filter to add robustness against environment clutter (e.g. players from opposition team), and (like our method) is a true online tracker. Additionally we show the different effects of the various innovations described in this paper by comparing: 1) multiple instances of the original LGT tracker [3]; 2) LGT enhanced by a "discriminative parts" method inspired from state-of-the-art pedestrian tracker [11], as described in Sec. 2.2.B; 3) further enhancement by our $local group - level$ models; 4) enhancement by $local - and - global group - level$ models.

The comparison methods, [3], [11] and [8] are a state-of-the-art single object tracker, multiple people tracker, and dedicated team-sports tracker respectively.

Accuracy ratio φ_k is defined as the overlap between the tracker-output bounding box TT_K and ground-truth box GT_k:

$$\varphi_k = (TT_k \cap GT_k)/(TT_k \cup GT_k) \tag{14}$$

Once each tracker has been run on our ground-truthed video dataset, we evaluate as follows. For a range of values of accuracy ratio, φ_k , we evaluate overall tracking performance according to the following four metrics [5]: False Negative Ratio (FNR); False Positive Ratio (FPR); Miss Match Ratio (MMR); Multiple Object Tracking Accuracy ($MOTA = 1 - FNR - FPR - MMR$). We then plot $MOTA$ against φ_k for each method. All experiments use the following parameter values: λ_M=0.03, λ=0.001, λ^c=0.001, λ_{ds}=0.01, λ_d=0.1 λ_t=0.1.

Fig.3 shows examples of our latent behaviour models detecting the ROI and the dominant player. Our method accurately locates the ROI in football video, showing similar performance to the motion fields approach of [6]. However, motion field approaches are not suitable for volleyball, where fewer players with more erratic motions cause non-smooth flow-fields. In contrast, our voting method can reliably detect ROI in volleyball videos also. The dominant player is identified successfully throughout the football video, with only very occasional failures (e.g. frame 30, Fig.3) in the much more challenging volleyball sequences.

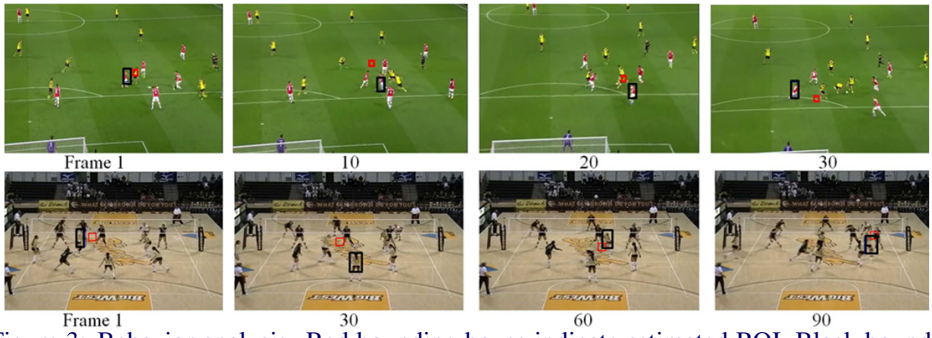

Figure 3: Behavior analysis. Red bounding boxes indicate estimated ROI, Black bounding boxes show a dominant player.

Table 1: Volleyball

Metric	LGT	LGT+DP	LGT+DP+LM	LGT+DP+LM+GM	BPF [8]
FNR	5.26%	4.66%	3.90%	**2.90%**	32.49%
FPR	1.27%	1.27%	1.27%	**1.27%**	1.27%
MMR	16.89%	15.37%	8.69%	**7.25%**	41.03%

Fig.4 shows the trade-off curves between *MOTA* and accuracy for each of the methods. Note, we do not show BPF, [8], on these trade-off curves as its performance is disproportionately poor (see following analysis in Tab.1 and Tab.2).

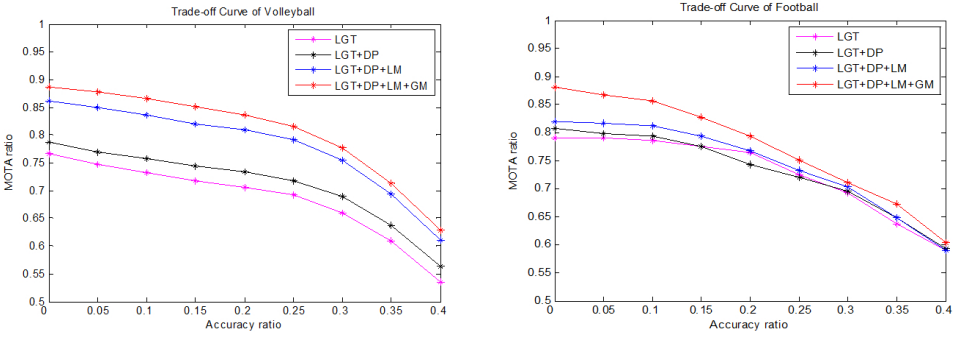

Figure 4: MOTA vs accuracy for volleyball (left) and football (right). LGT, discriminant parts (DP), local group-models (LM) and global group models (GM)

Tab.1 and Tab.2 show how the BPF method, [8], performs very poorly on volleyball and football videos. We believe this is because BPF relies on simple colour features which are easily distracted by background clutter (e.g. spectators). Tab.1 and Tab.2 show how *MOTA* breaks down into contributing factors: FNR, FPR, MMR. The *MMR* mismatches arise from two situations: i) when nearby trackers swap targets and ii) when the lost trackers re-acquire wrong targets. Therefore, we also analyze the relative values of FNR and *MMR*.

In volleyball, Tab.1, *MMR* obtained with the original LGT tracker is much larger than FNR, indicating that the principle failure mode arises from player-player occlusions. This in turn suggests that the improved performance of our approach mainly arises from the local group-level (player-pair) models.

In contrast, for the football game, *MMR* is much smaller than FNR (see Tab. 2) suggest-

Table 2: Football

Metric	LGT	LGT+DP	LGT+DP+LM	LGT+DP+LM+GM	BPF [8]
FNR	11.38%	9.79%	9.52%	**3.97**%	47.35%
FPR	6.88%	6.88%	6.88%	**6.88**%	6.88%
MMR	2.65%	2.65%	1.59%	**1.06**%	12.43%

ing that the main failure mode is due to losing targets - this means that *MMR* failures arise mainly from the re-acquisitions of wrong targets. This suggests that we can expect fewer improvements from the local-player-level models (Level 2) in football, with an improved performance predominantly due to the global group-level model - because football involves fewer player-player interactions than volleyball.

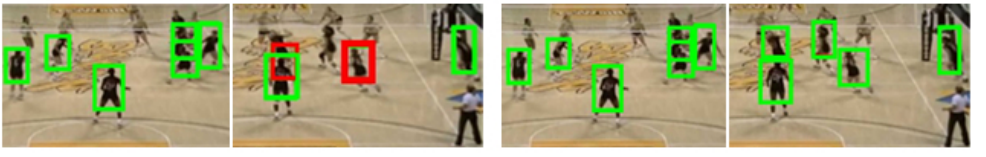

Figure 5: Frames 34, 81 of volleyball sequence: LGT (left pair) and our multi-level tracker (right pair). Green/red bounding boxes denote correct/erroneous tracking respectively.

Fig.5 illustrates the behaviour of our multi-level multi-target tracker in comparison to multiple instances of the original LGT tracker in the case of interactions among the players. The group-level models enable successful tracking of interacting/occluding player-pairs where LGT fails (see the right-most player-pair in the right-most image).

4 Conclusion

Sports-team tracking poses difficult challenges, which require high-level knowledge of team dynamics and player-player interactions. We have shown how a single tracker, that represents individuals as combinations of local and global layers, can be extended to exploit latent information about local and global group behaviours. Multi-level models are powerful, since different levels of representation show advantages in different tracking contexts that arise in different kinds of sporting activity.

Acknowledgments

Xiao is supported by the China Scholarship Council and also by a University of Birmingham school scholarship. Stolkin is supported by a senior University of Birmingham fellowship.

References

[1] Visual tracking using global and local visual information. Available at: http://www.vicos.si/Research/LocalGlobalTracking.

[2] William Brendel, Mohamed Amer, and Sinisa Todorovic. Multiobject tracking as maximum weight independent set. In *CVPR*, pages 1273–1280. IEEE, 2011.

[3] Luka Cehovin, Matej Kristan, and Ales Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(4):941–953, 2013.

[4] Robert T. Collins. Multitarget data association with higher-order motion models. In *CVPR*, pages 1744–1751. IEEE, 2012.

[5] Bernardin Keni and Stiefelhagen Rainer. Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008, 2008.

[6] Kihwan Kim, Matthias Grundmann, Ariel Shamir, Iain Matthews, Jessica Hodgins, and Irfan Essa. Motion fields to predict play evolution in dynamic sport scenes. In *CVPR*, pages 840–847. IEEE, 2010.

[7] Jingchen Liu, Peter Carr, Robert T Collins, and Yanxi Liu. Tracking sports players with context-conditioned motion models. In *CVPR*, pages 1830–1837. IEEE, 2013.

[8] Wei-Lwun Lu, Kenji Okuma, and James J Little. Tracking and recognizing actions of multiple hockey players using the boosted particle filter. *Image and Vision Computing*, 27(1):189–205, 2009.

[9] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, pages 261–268, 2009.

[10] Patrick Pérez, Carine Hue, Jaco Vermaak, and Michel Gangnet. Color-based probabilistic tracking. In *ECCV*, pages 661–675. Springer, 2002.

[11] Bo Yang and Ram Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. In *ECCV*, pages 484–498. Springer, 2012.

[12] Ting Yu and Ying Wu. Collaborative tracking of multiple targets. In *CVPR*, volume 1, pages I–834. IEEE, 2004.