Person Re-Identification Using Group Information

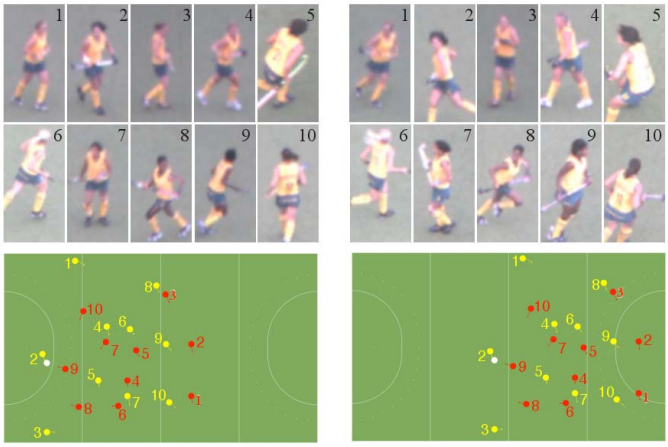
Alina Bialkowski*, Patrick Lucey†, Xinyu Wei* Sridha Sridharan*
*Image and Video Laboratory, Queensland University of Technology, Australia
†Disney Research, Pittburgh, USA
{a.bialkowski, xinyu.wei}@connect.qut.edu.au, patrick.lucey@disneyresearch.com, s.sridharan@qut.edu.au

***Abstract*—After first observing a person, the task of person re-identification involves recognising an individual at different locations across a network of cameras at a later time. Traditionally, this task has been performed by first extracting appearance features of an individual and then matching these features to the previous observation. However, identifying an individual based solely on appearance can be ambiguous, particularly when people wear similar clothing (i.e. people dressed in uniforms in sporting and school settings). This task is made more difficult when the resolution of the input image is small as is typically the case in multi-camera networks. To circumvent these issues, we need to use other contextual cues. In this paper, we use "group" information as our contextual feature to aid in the re-identification of a person, which is heavily motivated by the fact that people generally move together as a collective group. To encode group context, we learn a linear mapping function to assign each person to a "role" or position within the group structure. We then combine the appearance and group context cues using a weighted summation. We demonstrate how this improves performance of person re-identification in a sports environment over appearance based-features.**

I. Introduction

In a surveillance setting, the difficulty of re-identifying a person exponentially increases proportional to the time the person was last seen. This is because the number of possible options a person has increases over time, which in terms of permutations, quickly increases to infinity. For example, a person seen on a street corner at time t_0 has numerous options in terms of what they are doing next. They could potentially walk down the adjacent streets, hop on a train/bus/taxi, change clothing, or even just wait on the corner. After ten seconds, we would have a good estimate of where that person could be due to the limited options a person could take in that time, in addition to the scene remaining somewhat familiar. However, one minute later the number of options increases and makes this task much more difficult. Now, ten minutes later this task is near impossible (unless the person has not moved) as the number of possible permutations explodes. However, if that person was part of a large group of people, the likelihood of that group being confused as another group is much lower. As such, we can greatly reduce the search space of re-identifying that person by just choosing the most likely candidate within that group of people - instead of the huge number of possibilities if the person was alone. As humans are social beings, most realistic settings have people moving as a group rather than an individual. In this paper, we leverage off the group dynamic to improve person re-identification.

Although the above assumption is obvious, the big bottleneck which has restricted research in this area is the collection and annotation of large amounts of group data. As such, nearly all the work in person re-identification has solely focussed on modelling a single person (apart from the work of Zheng et al. [1]). However, with the influx of vision-systems being applied in the sporting domain due to the commercial applications, sporting datasets provide a perfect test-bed for investigating different approaches for person re-identification. They provide similar scenarios to surveillance environments, where resolution is generally low, and people may have similar appearances (e.g. a group of school children in uniform).



(a) (b)

Fig. 1. The players of a sports team are represented at two time instants, (a) and (b). It can be seen that the player appearances vary significantly between the two time instants, in terms of illumination, viewing angle and pose, and it is difficult to distinguish between the players because they are wearing the same uniform. While appearance alone is ambiguous, the structure of the team, represented in the bottom half of the figure, remains similar. This group context can be fused with appearance features to improve person re-identification.

In this work, we consider the task of recognising the identity of people with very similar appearances, from team sports video data (see Figure 1). This provides a constrained environment to evaluate person re-identification models as there are a fixed number of subjects, and a well defined area where they can be observed. However, it is also extremely challenging, as the video is captured in an outdoor environment by multiple cameras, which results in significant illumination variations between observations of each subject, the image resolution of players is low, which makes digit recognition on the jerseys near impossible, and player poses vary significantly. Such a dataset allows us to evaluate person re-identification models in real-life conditions, and allows us to evaluate the use of group context. Using this dataset, we first explore how existing appearance-based features perform. We then propose the use of group context in the form of player roles to help

disambiguate between people with similar appearances, and demonstrate how this contextual information can be used to improve person re-identification performance.

II. Related work

The majority of existing person re-identification methods rely solely on visual information to identify individuals. In such *appearance-based* re-identification methods, approaches seek to extract a variety of global and local features from the whole-body that are distinctive and robust to viewpoint, pose and illumination changes. Colour, texture and interest point features have been extracted and classified in a number of ways. Interest points have been used to locate and compare local patches between individuals. Gheissari et al. [2] segmented a person into regions using a triangulated graph model to compare colour and edgel information between corresponding parts of individuals, while Hamdoun et al. [3] used interest points based on a variant of SURF [4].

Other approaches extract a large number of features, and learn the weightings and most discriminative components from a training set. Gray and Tao [5] proposed an Ensemble of Localized Features (ELF), and used AdaBoost to learn the most discriminative colour and texture-based features, while Bąk et al. [6] learned the most discriminative Haar-like features and dominant colour descriptors using AdaBoost. Prosser et al. [7] reformulated the person re-identification problem as a ranking task instead of distance calculation/absolute scoring, using RankSVM to learn discriminative features. Schwartz et al. [8] proposed Partial Least Squares (PLS) reduction to project a large feature set consisting of colour, texture, and edge information into a low-dimensional discriminant latent space. Bąk et al. [9] considers the multi-shot scenario where several frames of each individual are available, and proposed the Mean Riemannian Covariance Grid (MRGC) to represent people. A person is split into a grid of overlapping cells, for which covariance features [10] are extracted, and the most relevant patches to describe each individual are learnt based on variance. Unlike the boosting and ranking approaches which learn a global weighting of features across all subjects, Liu et al. [11] distribute weights to different features based on their importance in that image (e.g. colour is more informative when a person wears a textureless bright coloured shirt, while texture can be more important for a person wearing a checkered shirt).

Instead of having a training phase to learn discriminative features or regions, other methods simply extract a collection of features which are view invariant. Bazzani et al. [12] proposed a person descriptor which includes a global HSV colour histogram, an 'average' texture of the person and a set of recurring textural motifs within the subject. Farenzena et al. [13] extended this work, in Symmetry-Driven Accumulation of Local Features (SDALF). They used symmetry to split a person into head, torso and legs, and added Maximally Stable Colour Region (MSCR) [14] features in the models, and achieve good view invariance. Zhao et al. [15] extracted and matched distinctive salient parts based on colour and SIFT features in an unsupervised manner, and outperform SDALF, PLS and ELF. However, this requires there to be unique colour or textural components within the subject set.

When people have very similar appearances, such as when wearing uniform, the intra-person appearance variations may be greater than the inter-person variations, so additional information or context must be used to more accurately re-identify people. In all the above described methods, only appearance is used for matching as the task assumes that the camera placement or paths that people may take between cameras is unknown. Other person re-identification methods have looked at additionally utilising the spatial layout and temporal constraints of the camera network to limit the set of candidates to be matched [16]–[18]. Depending on the camera network, such context is not always available, and instead other contextual information must be used. Zheng et al. [1], showed that associated groups of people instead of individuals can improve person re-identification performance, using a group descriptor which encodes visual words and their spatial relationships.

In team sports, player positions and movement are heavily linked to one another and to game context, and can be used to fill in the gaps of missed tracks caused by poor player detection. Liu et al. [19] made use of such contextual information to improve player tracking. They extracted game context from the global and local distribution of players (to indicate which team is attacking, and situations when opposing players normally follow each other closely) to give a more accurate motion model for tracking players. Lucey et al. [20] used team centroid as a contextual feature to approximate player role in conjunction with a spatiotemporal bilinear model to clean-up noisy data. Lu et al. [21] used a conditional random field incorporating SIFT, MSER and colour histogram features to track and identify individuals in broadcast footage. Their data had sufficient resolution to detect jersey numbers (when visible) which allows for better person identification than low resolution domains, where existing methods have only looked at extracting the team of each player.

Compared to the existing approaches for person re-identification which only consider individuals' appearance, we incorporate group information. However, unlike the method of Zheng et al. [1] in which the appearance of a group is modelled, we utilise group structure, and incorporate context in the form of relative player positions or "roles". Also, unlike the method of Lu et al. [21] who had higher resolution broadcast footage and used conditional random fields incorporating appearance and temporal information, our work looks at identifying players in low resolution footage (i.e. player heights of 40-100 pixels) without temporal information.

III. Evaluation Overview

A. Dataset

To evaluate person re-identification using group context, we use team sports video data. This provides a real-life outdoor environment to evaluate person re-identification models with group context, and a fixed number of subjects visible over a long duration with repetitive behaviours and structure. Typical challenges of person re-identification are present such as variations in subjects' orientation (e.g. viewed from front/side/back), pose (standing straight, crouching), resolution, and illumination. Examples of these variations for a single player are shown in Figure 2.

To provide a large dataset of players and their group context, we recorded several matches of field-hockey using a

Fig. 2. Example image patches of a single player, captured at different times and locations on the field are shown above. A wide degree of appearance variation in terms of illumination, viewpoint, and pose is apparent.

Fig. 3. View from the 8 fixed HD cameras taken from the field-hockey pitch.

test-bed of eight fixed high-definition cameras. An example frame from each camera is shown in Figure 3. While the camera test-bed provides complete coverage of the field, we consider the task of person re-identification instead of tracking, as we wish to observe how group context can assist in person re-identification.

We automatically extracted image patches using a state-of-the-art real-time person detector [22], which detects players by interpreting background subtraction results in terms of 3D geometry, where players are coarsely modelled as cylinders. The images were scaled to a fixed size of 96×50 pixels, and ground truth player positions and their identity were manually labelled.

After detecting the player positions and extracting their image patches, the player locations from all eight cameras were merged based on proximity, and the team or "group" was assigned based on a colour histogram of the player in the LAB colour space. This essentially provides a top down view of the match with player positions and their team. An example of this process, on two camera views is shown in Figure 4.

The roles and person re-identification were evaluated on a single match, where we considered the team dressed in yellow tops and green skirts (See Figure 1 for an example of 10 of these players). Two parts of this match were annotated to evaluate role assignment and person re-identification (consisting of 3893 frames and 8838 frames respectively). The person re-identification evaluation was performed on a set of 94 images automatically extracted for the 13 players of the team.

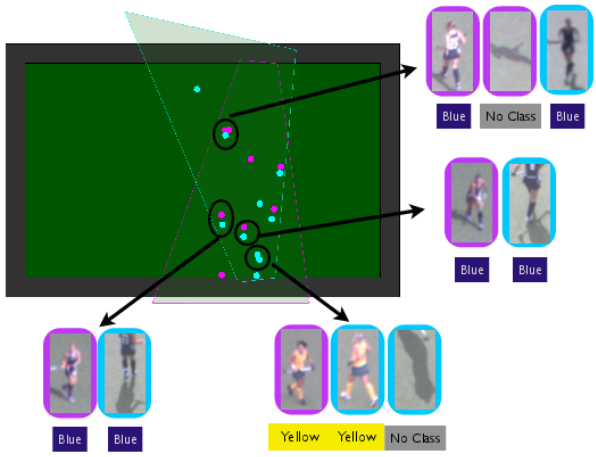


Fig. 4. Player detection and team recognition in a multi-camera environment.

B. Visual Features

In a domain where people are only visible at low resolution, and in different orientations and poses, features which can be extracted from a distance and which are invariant to these variations are desirable for person re-identification. In a sports domain, players within a team have very similar appearances as they wear the same uniform. Given high resolution imagery of each player, they could be uniquely identified based on their facial characteristics or jersey numbers, however these features are not visible due to insufficient resolution and motion blur. Texture, which has been used in a number of existing approaches, does not provide any information for distinguishing between individuals in this context as they all wear the same uniform. Height and body shape varies slightly between players, however due to the unconstrained pose and orientation, it is not possible to accurately extract these. Traits which are visible at long range and which may distinguish between players include hair, skin and shoe colour, and therefore we look at descriptors which encode colour and spatial information. We consider two types of appearance features to describe each person: Symmetry-Driven Accumulation of Local Features (SDALF) and region covariance descriptors.

In the SDALF [13] approach, a person is split into their body and legs based on symmetry, and features are extracted and matched between these parts. Because illumination and colour can vary significantly between observations, we normalise by scaling the mean illumination and applying histogram equalisation. We then extract a weighted histogram and maximally stable colour region (MSCR) features for each body part (we exclude the recurrent highly structured patterns which was used in the method proposed by the authors, as players are dressed in the same uniform). By splitting a person into the body parts, coarse correspondence can be gained in matching features.

In the region covariance descriptor [10], information is encoded about the variances of the features inside a region, their correlations with each other and a spatial layout. We calculate a collection of these descriptors for each player image by splitting the image into a set of regions, and describing each by a covariance matrix, C_R. Each pixel in the region is represented by a point in feature space (e.g. the spatial

coordinates of the pixel, its colour, and gradients), and the covariance matrix of a region R in the image consisting of n pixels, can be calculated using the covariance of the features $\{\mathbf{z}_k\}_{k=1\ldots n}$:

$$\mathbf{C}_R = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{z}_k - \boldsymbol{\mu})(\mathbf{z}_k - \boldsymbol{\mu})^T, \quad (1)$$

where $\boldsymbol{\mu}$ is the mean value of the $\mathbf{z}_k$'s.

Our pixel feature vector, $\mathbf{z}_k$, consists of the x and y coordinates of the pixel and its colour value in each channel (R,G,B),

$$\mathbf{z}_k = [x, y, R_{xy}, G_{xy}, B_{xy}]. \quad (2)$$

Gradient features were not used as they were found to decreased performance, due to large variations in player pose.

Corresponding regions between two images can then be compared using the following distance measure, proposed is [23]:

$$\rho(\mathbf{C}_1, \mathbf{C}_2) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(\mathbf{C}_1, \mathbf{C}_2)}, \quad (3)$$

where $\{\lambda_i(\mathbf{C}_1, \mathbf{C}_2)\}_{i=1\ldots n}$ are the generalised eigenvalues of $\mathbf{C}_1$ and $\mathbf{C}_2$.

To get the overall distance between two images, we perform a weighted sum of the cell region distances:

$$\text{Distance}(Subject_A, Subject_B) = \sum_{m=1}^{M} \frac{\rho(C_{A,m}, C_{B,m})}{\sigma_{A,m} + \sigma_{B,m}}, \quad (4)$$

where m corresponds to the cell region number up to the total number of cells, M, and $\sigma_{A,m}$ and $\sigma_{B,m}$ are the variance for subject A, and B, between region cells in the database.

The variance of subject i for cell m is calculated as:

$$\sigma_{i,m} = \frac{1}{n-1}\sum_{k=1;k\neq i}^{N} \rho^2(C_{i,m}, C_{k,m}), \quad (5)$$

where k corresponds to all the other subjects in the database.

In this way, each cell is weighted according to its variance to corresponding cells within the subject database (similarly to Bąk et al. [9]), in an aim to give greater weight to more discriminative cells (i.e. the regions which differ most or have greater variance from other subjects should represent regions that are more discriminant).

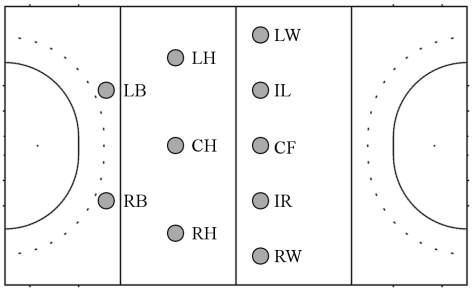


Fig. 5. In field-hockey, players move as a formation, with each player in the team being assigned a role or responsibility. Given that we can sense the locations of all the individuals, we can estimate the roles that each player takes within the formation at an instant in time, and use this to assist in identification.

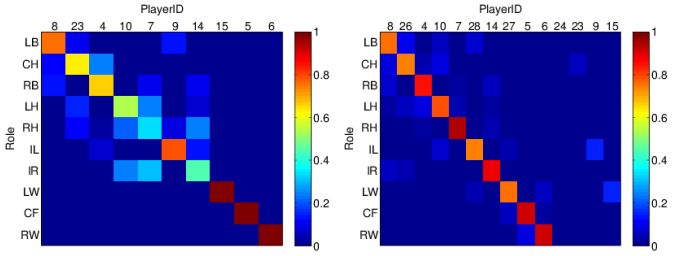


Fig. 6. Distribution of roles to player identities from the manually labelled player roles and identities for part 1 and part 2 of the match. It is apparent that players tend to play one main role, and sometimes swap to similar (neighbouring) roles. It can also be seen that in the second half, players have been substituted so their role is replaced by another player (e.g. 9 → 28, 15 → 27).

C. Role Assignment

When a group of individuals move through a space, such as a crowd moving through a foyer, patterns of motion occur. People generally walk in designated paths, and take similar routes due to the layout of the space and to avoid collisions with other people. This results in structure to the movement of people, and accurately modelling the structure can assist re-identification.

In the majority of team sports, the coach or captain designates an overall structure or system of play for a team. In field hockey, the structure is described as a formation involving roles or individual responsibilities. For instance, the 5:3:2 formation defines a set of roles R = {left back (LB), right back (RB), left halfback (LH), centre halfback (CH), right halfback (RH), inside left (IL), inside right (IR), left wing (LW), centre forward (CF), right wing (RW)}, as shown in Figure 5. While players may swap roles throughout a match, they will predominantly play in one role, and hence roles can be used as a contextual feature for identifying players

For training and evaluation purposes, the player roles were manually labelled in our dataset. To give an indication of how well roles correspond to player identities, we calculated the frequency that each player identity was assigned to the manually labelled roles, and this is presented in Figure 6 as confusion matrices, sorted so that the players most likely to play in each role appear on the diagonal. From these matrices, it is apparent that roles provide information towards player identities.

To assign players to roles automatically, we adopt a similar procedure to that proposed in [20]. We first sense the location of all the players on the team, and then map each player's position in the formation to a role. Given an initial ordering of the 10 field players of a team, $\mathbf{p}_t = [x_1, y_1, x_2, y_2, ..., x_{10}, y_{10}]^T$, at time instant t, the goal is to find the 10×10 permutation matrix, $\mathbf{x}_t$, which re-arranges the players into role order: $\mathbf{r}_t = \mathbf{x}_t\mathbf{p}_t$. The permutation matrix is a binary matrix, and if element $\mathbf{x}_t(i,j) = 1$, it indicates that player i is assigned to role j. By definition, every row and column in this matrix must sum to one (i.e. each player is assigned to one role).

The role assignment task is formed as an optimisation problem where the goal is to minimise the L_2 reconstruction error:

$$\mathbf{x}_t{}^* = \arg\min_{\mathbf{x}_t} \|\hat{\mathbf{r}} - \mathbf{x_t}\mathbf{p_t}\|_2^2. \quad (6)$$

This is a linear assignment problem where an entry $C(i,j)$ in the cost matrix is the Euclidean distance between role locations:

$$C(i,j) = \|\hat{\mathbf{r}}(\mathbf{i}) - \mathbf{p_t}(\mathbf{j})\|_2 \quad (7)$$

To solve the assignment problem, we first find the most similar prototype formation, $\hat{\mathbf{r}}$, from a set of 25,000 labeled frames of field hockey data, based on the mean and covariance of the team's formation in the current frame. Then the optimal permutation matrix is found using the Hungarian algorithm [24].

To evaluate the automatic role assignment, we compare against the manually annotated role labels. The results are presented in Figure 7. It can be seen that we have a major diagonal, however sometimes the roles are wrongly classified due to ambiguity in the formation or roles, particularly in the midfield (RH,IL,IR). The automatic role assignment has an accuracy of 66.0%.

Since we don't know in advance which roles a player will take throughout a match, we assume that each player has an 'assigned role' as shown in Table I. Given an assigned role, we can estimate the most likely player ID. Due to player substitutions, and multiple players being able to take each role, it is evident that using role context alone, we can't tell which players are on the field at a certain time, and so we will have ambiguity in which player is playing. It is expected that appearance features will improve results.

IV. Experiments

To evaluate person re-identification performance and the proposed group context using player roles, we evaluate appearance features alone, role alone, and then combine role and appearance information.

TABLE I. Player IDs assigned to each role

LB	CH	RB	LH	RH	IL	IR	LW	CF	RW
8	23	4	10	7	9	14	15	5	6
	26				28		27		

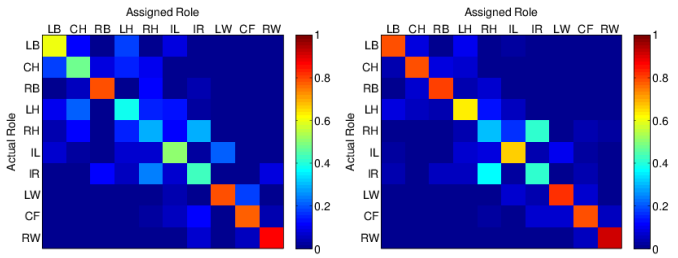
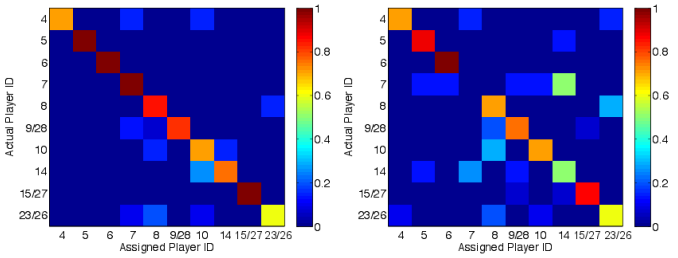


Fig. 7. Accuracy of automatic assignment of roles for Part 1 = 59.0%, Part 2 = 69.0% (overall of 66.0%)



(a) (b)

Fig. 8. Accuracy of person identification using (a) manually labelled roles = 84.5% and (b) automatically assigned roles = 67.4%

A. Identification using roles

Given prior knowledge of which roles corresponds to which players (as in Table I), we can perform identification directly on any member of a group formation without a gallery or training set. In Figure 8, identification results for all 94 images of our evaluation dataset, using role only are displayed. Results are shown for both perfect role extraction (i.e. roles manually annotated by an expert), and automatic role extraction. Note that due to player substitutions, we can not distinguish between the substituted players (23/26, 9/28, 15/27) using role alone.

When the role to player correspondences are not known before-hand, person re-identification can be performed by comparing roles of the testing subjects to those in the gallery set. This requires a distance measure of how similar or different a role is to another role. To get a distance measure between any two roles, we use the average confusion matrix of automatic assignment accuracy (see Figure 7) because roles which are easily confused must be similar. We convert this to a distance measure by subtracting the average assignment probabilities from a matrix of ones (and hence similar roles will be given a lower distance comparison measure).

B. Comparing features for identification

Since we wish to see how well we can perform identification, we present results using Cumulative Matching Characteristic (CMC) curves [5]. Each point is calculated as the cumulative probability that the actual subject of a test measurement is among its k top matches (where k is called the rank). e.g. The Rank-1 value indicates what proportion of the time a player is the closest match to itself, while Rank-5 indicates how often the correct player is within the top 5 matches to itself.

The evaluation was performed on a set of 94 images

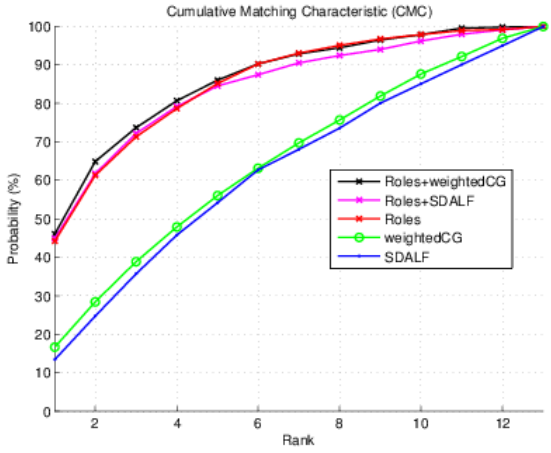


Fig. 9. Cumulative Matching Characteristic curves for each of the person re-identification features, displaying the probability that the correct subject is within the top-k matches to itself, where k is the rank.

automatically extracted for the 13 players of the team. Two randomly selected images of every player were selected from the database (one for the gallery, and the other as part of the test set). The gallery and test set were then matched, and the results of 100 experiments were averaged to produce the results. We compare SDALF features, weighted covariance grid ("weightedCG"), and role assignments (automatically generated from the player position within the formation), and appearance features combined with roles using a weighted summation. The results are presented in Figure 9.

In Figure 9, it can be seen that both appearance features, SDALF and weighted covariance grid features, perform similarly poorly. This is expected as the players are all dressed very similarly, and appearance features in low resolution footage are insufficient to distinguish between the players. The weighted covariance grid slightly outperforms SDALF, and this may be due to the discriminative weighting of the cells. In comparison, roles are able to distinguish players very well, and we gain a minor improvement by fusing roles with appearance.

V. Summary and Future Work

Person re-identification is very difficult in low-resolution video footage, especially when people wear similar clothing which limits the usefulness of traditional appearance-based approaches. To circumvent these issues, we propose the use of "group" information as a contextual feature to aid in the re-identification of a person. To encode group context, we learn a linear mapping function to assign each person to a "role" or position within the group structure. We then combine the appearance and group context cues using a weighted summation. We demonstrate how this improves performance of person re-identification in a sports environment over appearance based-features. In future work, we plan to test our method on more data as well as try to learn salient appearance features of a person.

Acknowledgment

This research was supported by the Queensland Government's Department of Employment, Economic Development and Innovation.

References

[1] W. S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC*, vol. 5, 2009.

[2] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Computer Vision and Pattern Recognition (CVPR)*, 2006.

[3] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *International Conference on Distributed Smart Cameras (ICDSC)*, 2008.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *European Conference on Computer Vision (ECCV)*, 2006.

[5] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *European Conference on Computer Vision (ECCV)*, 2008.

[6] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using Haar-based and DCD-based signature," in *Advanced Video and Signal Based Surveillance (AVSS)*, 2010.

[7] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking." in *BMVC*, vol. 2, 2010, p. 6.

[8] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis," in *International Conference on Computer Vision (ICCV)*, 2009.

[9] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *Advanced Video and Signal-Based Surveillance (AVSS)*, 2011, pp. 179–184.

[10] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *European Conference on Computer Vision (ECCV)*, 2006.

[11] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: what features are important?" in *Computer Vision–ECCV 2012. Workshops and Demonstrations*, 2012, pp. 391–401.

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

[13] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[14] P. Forssén, "Maximally stable colour regions for recognition and matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[15] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[16] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 146–162, Feb. 2008.

[17] G. Lian, J. Lai, and W.-S. Zheng, "Spatial–temporal consistent labeling of tracked pedestrians across non-overlapping camera views," *Pattern Recognition Letters*, vol. 44, no. 5, pp. 1121–1136, 2011.

[18] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognition Letters*, vol. 33, no. 14, 2012.

[19] J. Liu, P. Carr, R. Collins, and Y. Liu, "Tracking Sports Players with Context-Conditioned Motion Models," in *CVPR*, 2013.

[20] P. Lucey, A. Bialkowski, P. Carr, S. Morgan, I. Matthews, and Y. Sheikh, "Representing and Discovering Adversarial Team Behaviors using Player Roles," in *CVPR*, 2013.

[21] W. L. Lu, J. A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[22] P. Carr, Y. Sheikh, and I. Matthews, "Monocular object detection using 3d geometric primitives," in *ECCV*, 2012.

[23] W. Förstner and B. Moonen, "A metric for covariance matrices," *Technical Report, University of Stuttgart*, 1999.

[24] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, 1955.