Computer Vision and Image Understanding 159 (2017) 105–115

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Soccer player recognition using spatial constellation features and jersey number recognition

Sebastian Gerke*, Antje Linnemann, Karsten Müller

Fraunhofer HHI Einsteinufer 37, 10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 24 March 2016
Revised 22 April 2017
Accepted 24 April 2017
Available online 27 April 2017

ABSTRACT

Identifying players in soccer videos is a challenging task, especially in overview shots. Face recognition is not feasible due to low resolution, and jersey number recognition suffers from low resolution, motion blur and unsuitable player pose. Therefore, a method to improve visual identification using spatial constellations is proposed here. This method models a spatial constellation as a histogram over relative positions among all players of the team. Using constellation features might increase identification performance but is not expected to work well as a single mean of identification. Thus, this constellation-based recognition is combined with jersey number recognition using convolutional neural networks. Recognizing the numbers on a player's shirt is the most straight-forward approach, as there is a direct mapping between numbers and players.

Using spatial constellation as a feature for identification is based on the assumption that players do not move randomly over the pitch. Players rather follow a tactical role such as central defender, winger, forward, etc. However in soccer, players do not strictly adhere to these roles, variations occur more or less frequently. By learning constellation models for each player, we avoid a categorical assignment of a player to one single tactical role and therefore incorporate each player's typical behaviour in terms of switching positions.

The presented player identification process is expressed as an assignment problem. Here, an optimal assignment of manually labeled trajectories to known player identities is calculated. Using an assignment problem allows for a flexible fusion of constellation features and jersey numbers by combining the respective cost matrices. Evaluation is performed on 14 different shots of six different Bundesliga matches. By combining both modalities, the accuracy is improved from 0.69 to 0.82 when compared with jersey number recognition only.

1. Introduction

Soccer is one of the most popular sports in the world. Interest in automatic soccer analysis tools grew significantly in recent years. Soccer analysis results can be used for new ways of storytelling on TV, for match preparations or statistical evaluation. One fundamental analysis is the identification of players to individually associate actions and statistics. But due to the fact that no reliable automatic identification technologies exist at present, this task is typically carried out by human annotators, using respective tools and standards. The identification of players in broadcast sports videos is of utmost interest and accordingly, a number of researchers addressed this problem in the recent past as reviewed in Section 2.

However, identifying players in broadcast soccer videos automatically (and even manually) is challenging. Especially for the overview camera this task is difficult due to the low resolution per player, which makes face recognition impossible. And even jersey numbers are often hard to recognize, especially in standard definition (SD) resolutions. Only with the rise of widely available high definition (HD) content in recent years, jersey number recognition became feasible. The spatial constellation of players supports manual annotators in the identification process, as players do not move randomly over the pitch. However, this feature is left unexploited in most automatic identification approaches.

Therefore, the main contributions of this work are spatial team constellation features and models that are suitable for soccer player identification. Furthermore, a combination of these spatial constellation models with jersey number recognition is contributed.

This paper is organized as follows. First, related work in the area of player identification in sports video is presented in

* Corresponding author.
E-mail address: sebastian.gerke@hhi.fraunhofer.de (S. Gerke).

http://dx.doi.org/10.1016/j.cviu.2017.04.010
1077-3142/

Section 2. Within this work, player identification is posed as an assignment problem. The exact formulation and how to solve an assignment problem is described in Section 3. A way to model spatial constellation of player positions is explained in Section 4. The jersey number recognition using convolutional neural networks (CNN) is presented in Section 5. Then, both modalities are combined, as described in Section 6. In Section 7, the conducted experiments are described in detail. Both modalities are evaluated separately and then compared to the combined method. Finally, a conclusion is drawn in Section 8 including possible further research directions.

2. Related work

Existing work on identification of players in team sport broadcasts mostly rely on visual features only and two subcategories stand out in particular: One performing face recognition, while the other investigates jersey number recognition.

As an example for the first group Ballan et al. (2007) performed face recognition for soccer close-up shots. Specific problems that may occur in sport video close-ups such as high variation in pose, illumination, scale and occlusion are addressed by employing scale-invariant feature transform (SIFT) features. In Mahmood et al. (2014), a sports entertainment application is presented which identifies players in various smart phone videos and images of baseball games to superimpose relevant statistics. For player and subsequent face detection an AdaBoost algorithm with haar-like features is used. Detected faces are matched with a database of 80 player faces by using an LDA-based boosting algorithm. It is shown that images with fewer number of players, which implies that they are taken from a smaller distance and thus are presented in a higher resolution, result in a significantly better recognition performance.

For the second subcategory of visual player identification a typical work flow is presented in Delannay et al. (2009) and Lu et al. (2013a). Despite the different input data, the basic procedure of jersey number recognition is identical. First, explicit localization of jersey numbers is performed, followed by a digit segmentation and a single digit classification. In Delannay et al. (2009) basketball players are identified in a multiple camera setup. Here, the digits are classified by an SVM using multiple features, such as second order moments and color histograms of regions. An optical character recognition (OCR) scheme is adopted in Lu et al. (2013a) to identify jersey number digits of basketball players. The sports videos which are processed in this work are captured by a single pan-tilt-zoom camera from the court view. In Andrade et al. (2003), Andrade et al. propose a jersey number recognition system based on segmentation of number regions, color and size normalization, and optical character recognition (OCR). Messelodi and Modena describe a more general system for athlete identification in sport videos in Messelodi and Modena (2011). It not only identifies numbers, but also athletes' names on the jersey. In their work, they first locate text regions by highlighting regions with a manually given background color, then apply filtering for connected components around actual text. Finally, these regions are fed into an OCR system.

All above mentioned approaches either operate on other sports than soccer, where the resolution per player is higher (e.g. basketball), or they are restricted to process close-up and medium shots only. These are not expected to work well in soccer overview shots where the resolution per player is low. However, overview shots are shown most of the time during broadcast transmission. Therefore, Yamamoto et al. (2013) propose a system to recognize jersey numbers of American football players holistically, i.e. not by employing OCR or single digit recognition but whole numbers. They use SIFT features to provide robustness against viewpoint changes. Similarly, an advanced system for soccer jersey number recognition is introduced in Gerke et al. (2015). A deep convolutional neural network is trained which handles the complete pixel-to-jersey number recognition process and works well for soccer overview (i.e. wide-angle) shots. In Lu et al. (2013b), Lu et al. avoid the problem of low resolution of jersey numbers and faces by recognizing the entire player by employing a combination of maximally stable extremal regions (MSER), SIFT and color features.

In Couceiro et al. (2014), Couceiro et al. demonstrate that a player's position can be predicted using his tactical role. They also show that the predictability varies for different roles by performing an entropy analysis on heat maps. A novel player identification approach is introduced in Gerke and Müller (2015). This method operates independently of visual features and is based on players' spatial constellation within their team. However, the experimental results presented in Gerke and Müller (2015) represent a feasibility study, as they are applied to a test data set consisting of complete 1-min average player positions, rather than regular soccer broadcast material. In particular, the information of all players for every time point is available. This, however, is not the case for regular broadcast material, where only a subset of players is visible simultaneously and tracks are disrupted by shot cuts. Accordingly, the approach from Gerke and Müller (2015) yields a lower identification accuracy when applied to regular broadcast material.

Some further works are identifying players by predicting activities, where one part of the whole system is an automatic assignment of detections to tactical roles (Bialkowski et al., 2013; Lucey et al., 2013). A manual assignment from tactical roles to player identities is then used to identify players. Additionally, they provide statistical insights into high-level tactical behavior of teams like differences between a team's behavior in home and away matches. These approaches pose player identification as an assignment problem (from detections to player identities), similar to what is proposed in this paper. However, the assignment is performed on player positions directly. But the absolute position of a player on the pitch is less descriptive than describing a player's position relative to his teammates' positions. Therefore, in our proposed approach, these spatial arrangements are explicitly modeled in order to mainly capture relative player positions rather than absolute positions. Additionally, we directly assign player identities to observed trajectories, whereas Lucey et al. assign tactical roles to trajectories and rely on a manual assignment of roles to player identities. In Beetz et al. (2005), Beetz et al. (similarly to Bialkowski et al., 2013) use player positions in their ASPOGAMO system to infer tactical roles.

More spatio-temporal analysis on player trajectories from team sports is described in a survey by Gudmundsson and Horton (2016). However, they focus on extracting knowledge and insights from a spatio-temporal analysis rather than using this information to identify players. E.g. in Cintia et al. (2015), Cintia et al. derive scores for teams based on their passing behaviour. Using these scores, they predict team success in a simulation when compared to real championship results. Other applications for spatio-temporal analysis in team sports are individual player ratings (Duch et al., 2010), predicting goal events, as in Frencken et al. (2011) or predicting ball ownership (Wei et al., 2015).

Overall, the presented state-of-the-art in player identification has been either applied to specific content different from soccer or specifically on close-up shots with visible face or jersey number features. For the most commonly used overview setting in soccer broadcast transmission, player identification has been rarely performed. Therefore, our proposed approach extends the player detection in particular in overview shots by explicitly modeling spatial team arrangements, including non-visible (hidden) players, and

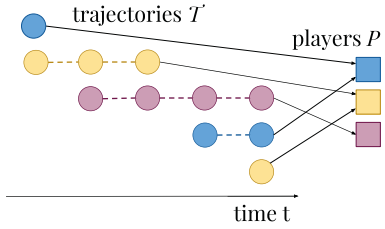

Fig. 1. Assignment problem of observed player trajectories T in the video. Here, a correct solution requires assigning more than one trajectory T to some known players P, as there are more trajectories than known players on the pitch.

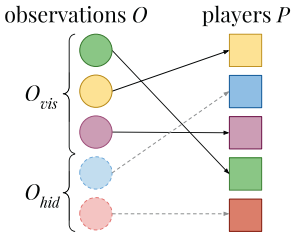

Fig. 2. Assignment problem of observed visible players in a video at a single point in time. To solve the assignment problem, dummy observations of hidden players O_{hid} are inserted to allow an one-to-one mapping between observations O (circles) and players P (rectangle).

combining them with additional jersey number recognition based on convolutional neural networks.

3. Player identification as an assignment problem

Within this work, player observations are given for certain time periods. Consecutive observations of the same player are aggregated into a trajectory. We further assume that the team assignment of each trajectory is already known or easily ascertained due to distinguishable jersey colors. Player trajectories are separated according to their team and the following conditions apply for identifying soccer players only within the same team.

3.1. Assignment problem

A sequence of player observations over a time period $t_i, \ldots, t_{i+n_j}$ is defined as a trajectory

$$T_j = \{o_j^{t_i}, \ldots, o_j^{t_i+n_j}\} \tag{1}$$

where $o_j^{t_i}$ is a single observation at time t_i and n_j is the length of a trajectory T_j. Then, $T = \{T_1, T_2, \ldots, T_m\}$ is the set of all m trajectories in a given video shot.

Our goal is to assign a player identity p from the set of all known players P to every trajectory in T, i.e. we seek an assignment function $\hat{f} : T \to P$. However, as shown in Fig. 1, there might be different trajectories assigned to the same single player p due to players appearing and disappearing from the camera field-of-view. In that case, the number of trajectories differs from the number of known players ($|T| \neq |P|$). The required computation time for solving such assignment problems might be feasible for a single short video sequences however drastically increases for more extensive assignment problems, i.e. analyzing complete soccer match recordings. To obtain a set of single manageable linear assignment problems, the trajectory assignment is decomposed into multiple assignment problems, one for each point in time $t_i \ldots t_{i+n_j}$.

3.2. Assignment of single player observations

The number of players in a set O, where O represents all observable unknown players at a single time point t_i, are of the same size n as the number of player identities in a set P that are known to be on the pitch at the same time t_i. The player identification process at a single point in time can therefore be expressed as a linear assignment problem.

We define a weight function

$$C : O \times P \to \mathbb{R} \tag{2}$$

which is represented by an $n \times n$ matrix $C = (c_{ij})$ containing the assignment costs between all elements of O to all elements of P.

Please note that in a soccer broadcast video, not all players are always visible. As shown in Fig. 2, the set of all observable, unknown players is composed of $O = O_{vis} \cup O_{hid}$, where O_{vis} represents the set of visible and O_{hid} the set of non-visible (hidden) players. We do not assign hidden players, however for a linear assignment problem, a square cost matrix is beneficial. Therefore, costs between elements of O_{hid} and elements of P are set to zero. This means, that assigning a hidden player to any identity does not increase the overall costs.

For optimal assignment, a bijective function f: $O \to P$ that minimizes the cost function is sought:

$$\sum_{o \in O} C(o, f(o)) \tag{3}$$

The minimization can be efficiently solved using the Hungarian algorithm (Munkres, 1957). It allows for solving the assignment in $O(n^3)$ complexity instead of $O(n!)$ for a brute force approach.

Within this paper we present two individual solutions for modeling the weight function; the first solution assigns players by recognizing the players' spatial constellation within their team, the second identifying their jersey numbers. The set of observed players O_{vis} and known players P are the same for both approaches. A linear combination of both weight functions is used for optimal assignment. In the following three sections, the two approaches (Sections 4 and 5) and their combination (Section 6) are described in detail.

3.3. Temporal consistency

To obtain temporal consistency within player trajectories, the solutions of the single bijective assignment problems $\{f_t | t = t_i..t_{i+n_j}\}$ need to be combined appropriately. We use a majority vote

$$\hat{f}(T_j) = \underset{p \in P}{\operatorname{argmax}}(|\{o_j^t |\, f_t(o_j^t) = p, \forall o_j^t \in T_j\}|). \tag{4}$$

This formulation allows to assign a single player to each trajectory. Fig. 3 shows an example for assigning $m = 6$ trajectories to four known players at individual successive time points.

However, temporally overlapping trajectories of different players, which are assigned to the same identity, may occur. An example of this invalid ambivalence is displayed in Fig. 3, where T_1 and T_2 are both assigned to player p_1 at time t_{i+1} after majority voting. There are two ways of handling this type of error: Either neglecting it and thereby reducing accuracy, or by post processing the results whenever this case occurs. In this work, no method to resolve this error is presented, as it does not occur in the dataset used here.

4. Player constellation

One method to identify players in a soccer match is based on the spatial constellation of players within their team. In this section, we describe how these constellations can be used for the representation of observed players and for the trained models of known players. Therefore the feature representation $\mathbf{x}_o$ of observations $o \in O_{vis}$, the trained models M_p of known players $p \in P$,

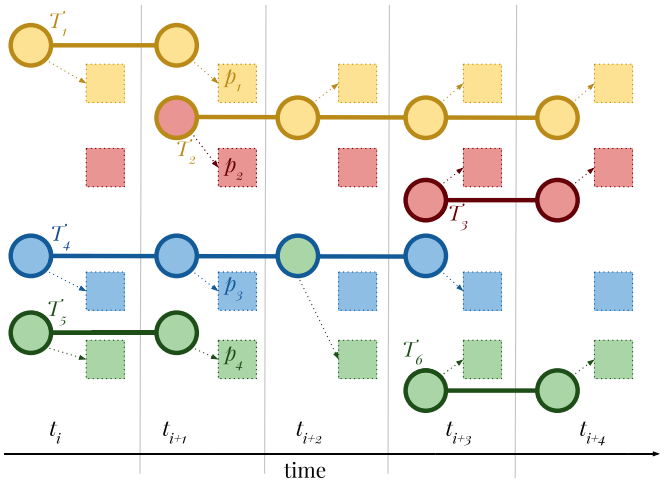

Fig. 3. Decomposing the trajectory assignment into single assignment problems, where connected circles represent a trajectory T_j and rectangles the known player identities p_i. A majority voting scheme is used to obtain the final trajectory assignments, i.e. T_1, $T_2 \rightarrow p_1$, $T_3 \rightarrow p_2$, $T_4 \rightarrow p_3$, T_5, $T_6 \rightarrow p_4$. A potential invalid ambivalence in trajectory assignment might occur at time point $t_i + 1$, where both trajectories T_1 and T_2 at the same time point are assigned to player p_1 after majority voting.

and a distance (or similarity) function $d(\mathbf{x}_o, M_p)$ between feature representations and models are described. Player constellations are determined in world coordinates on the pitch. Therefore, a camera calibration is needed which enables mapping of image coordinates to world coordinates.

4.1. Camera registration

In order to approximately map the position of detected players from their image coordinates $\mathbf{c}_i$ to world coordinates on the pitch $\mathbf{c}_w$, as shown in Fig. 4, a perspective transformation $\mathbf{H}$ between camera plane and world pitch plane is sought, with:

$$\mathbf{c}_i = \mathbf{H}\mathbf{c}_w \tag{5}$$

Since it is sufficient to describe the pitch as a plane, the regular 3D to 2D camera calibration with projection matrix can be simplified here as plane-to-plane homography. Moreover, while for computer graphics approaches, e.g. free view point video as presented in Angehrn et al. (2014) and Yao et al. (2016), usually very precise calibration is needed, for the purpose within this work it is sufficient to obtain an approximate estimate. In Farin et al. (2005), Linnemann et al. (2013) and Homayounfar et al. (2016) it is shown, that the line markings on the pitch serve as an adequate calibration pattern.

Within this paper, we used an extended version of the fully automatic approach presented in Linnemann et al. (2013). Here, a mean calibration error of only 2.53 m is achieved by finding a homography matrix $\mathbf{H}$ which projects a given pitch model in such a way that it best matches a monochrome rendering of the visible line markings. Therefore, a set of feasible homography matrices is obtained by locating specific point correspondences of detected lines and circles in the image to the given pitch model. In case of insufficient point correspondences the homography matrix is predicted from global camera motion.

Instead of searching for point correspondences to calculate the homography directly, a new simplified algorithm is used in this paper. Intrinsic and extrinsic camera parameters $\mathbf{K}$ and $\mathbf{R}$ are estimated and refined in a generic approach to calculate the set of feasible homographies with $\mathbf{H} = \mathbf{KR}$. The key advantage is the possibility to limit the range of camera positions and focal lengths to common values used in soccer broadcasts, i.e. the camera is in line with the half-way-line. Additionally, the camera position can assumed to be static within one soccer match.

4.2. Player constellation features

While the absolute player positions on the field vary significantly, certain patterns for the relative positions are prominent. This is shown in Fig. 5 with different examples for a 1-min average position of players of a single team within a single match. The significant patterns result from common soccer player positioning according to a tactical lineup. Each player mostly plays one or two tactical roles during a match. In order to capture these tactical roles, only features based on the difference vectors between an investigated player o_i and his teammates o' are considered:

$$\begin{aligned} &\mathbf{X}_o^{diff} = \{o' - o_i | o' \in O_{tm} \setminus o_i\} \\ &\text{with} \quad O_{tm} = O_{vis} \cup O_{dis} \quad \text{and} \quad O_{dis} \subseteq O_{hid} \end{aligned} \tag{6}$$

In most realistic soccer broadcasts not all 10 teammates are visible at all points in time. Therefore, proper difference vectors can only be computed between the player and his visible teammates O_{vis}. Teammates who never appear in the image are ignored. In contrast, difference vectors to teammates disappearing from the screen O_{dis} can be calculated by estimating their position. Here, we only use the last known position of a disappearing trajectory as a prediction. Trajectories are not smoothed in a classical way, as they use the obtained player positions directly. Temporal consistency is achieved by employing a majority vote.

In order to describe a position on the field, a coordinate system centered in the middle of the pitch, i.e. at the kickoff point, is used. Its extent is normalized between -1 and 1 in both the x axis (backward/forward direction) and the y axis (left/right direction). Absolute coordinates are not used because pitch sizes are not generally standardized and thus might vary. An example is displayed in Fig. 6 showing the relative vectors of four different players (player 27, 31, 9, 7) with their positioning given in Fig. 5.

In order to simplify handling of these difference vectors, for which no natural order exist, histograms over the difference vectors are calculated and used as a feature for describing a player position. For our approach we use a 2 × 2 histogram, with each

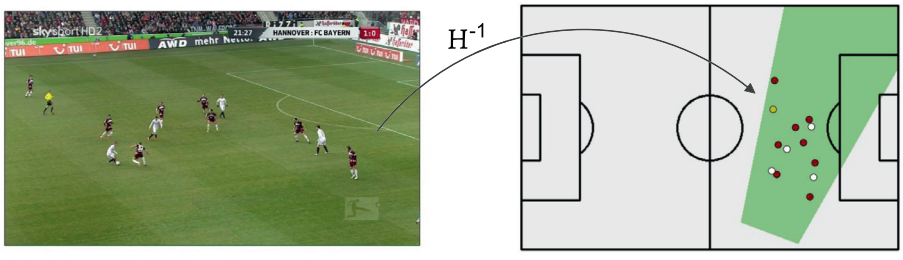

Fig. 4. Transformation of player positions from image coordinates to real-world coordinates on the pitch using the inverse of the homography H^{-1}.

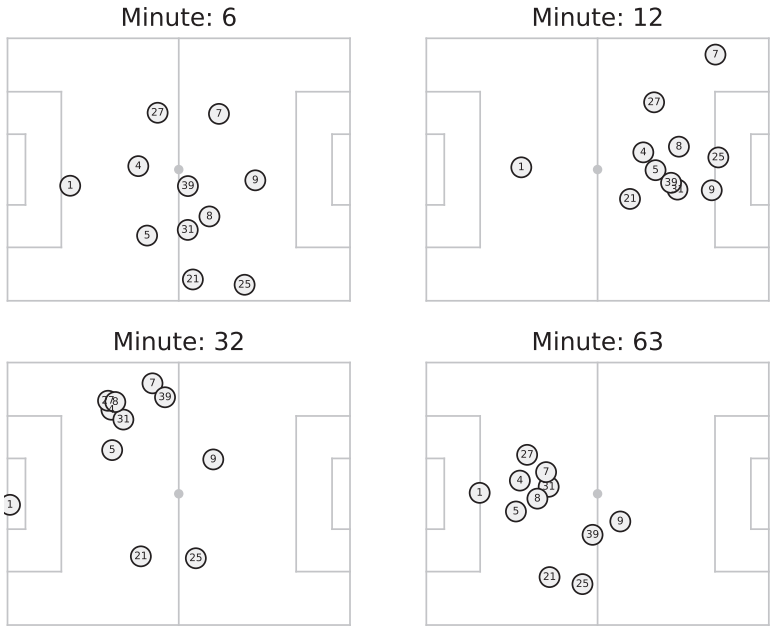

Fig. 5. Typical team lineup in a soccer match for four random minutes of play. While absolute positions per player vary significantly, certain tactical roles emerge.

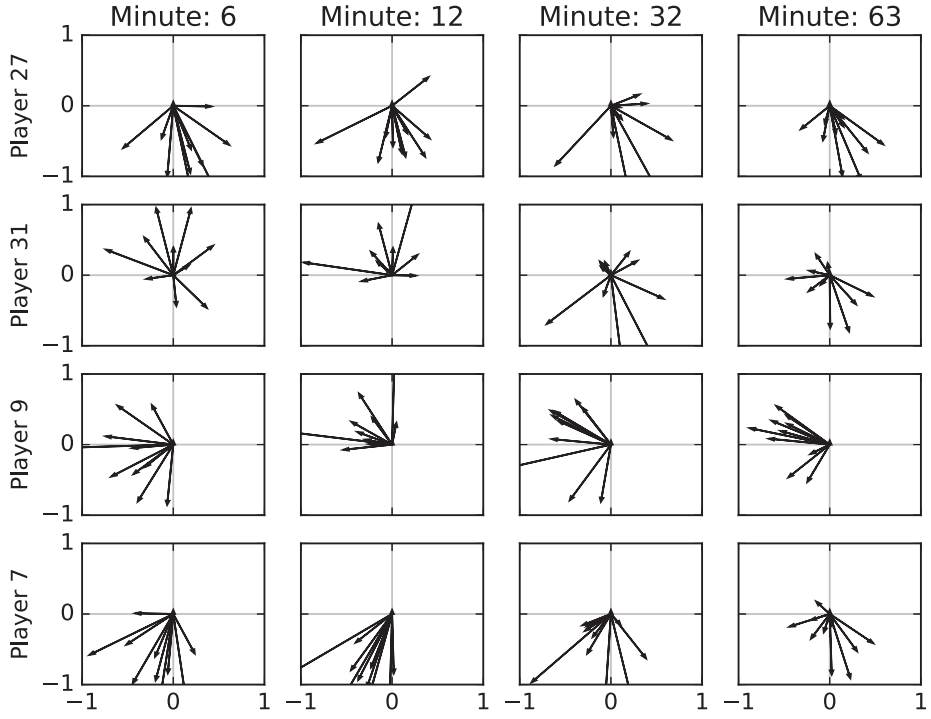

Fig. 6. Relative vectors of players 27, 31, 9, 7.

bin representing the number of players located in the stated direction. Although in previous experiments (see Gerke and Müller, 2015 for details) a 3 × 3 histogram yielded better results in a training data set, we found out that for actual broadcast content, a 2 × 2 histogram configuration works better. In Fig. 6 the edges of the histogram bins are plotted with grey lines. The resulting two-dimensional histograms are shown in Fig. 7. It depicts the histogram representation of the four players at the same four time points as in the previous figures. For each histogram bin, an arrow is drawn, whose extent along both the x and y axis represents the number of teammates positioned in that region. E.g. an arrow from (0, 0) to (5, 5) indicates that there are five teammates in that relative direction of the actual player.

4.3. Position models

To identify players based on their pitch position, individual models for each player are trained. For training we use the dataset described in Gerke and Müller (2015). It consists of 1-min average player positions on the pitch from the first 17 matchdays of the German Bundesliga 2012/2013 season.

For each player $p \in P$, a set of previously described 2 × 2 histogram features $\mathbf{X}_p^{diff}$ is used to train the model. As there are at most 11 players on the pitch, the set of histogram features is discrete and finite. Therefore, it is feasible to build a model consisting of the relative frequencies q_i of all possible *unique* training features

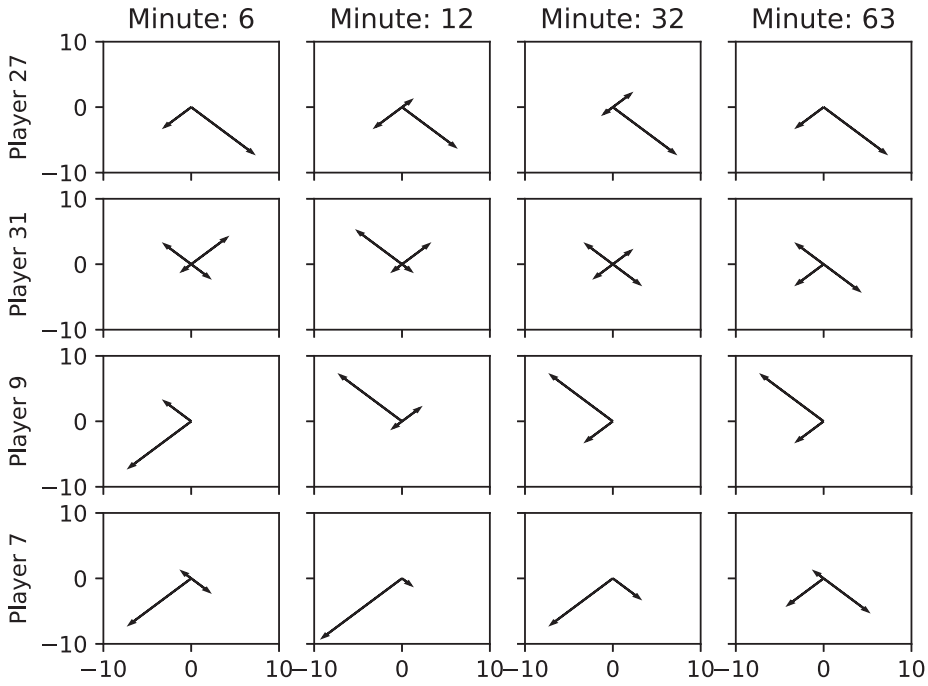

Fig. 7. 2 × 2 histograms over the relative vectors in Fig. 6. Histogram bins point into the direction of their relative position. Please note that the maximum of 10 in each direction relates to the maximum of 10 other players excluding the actual player.

$\hat{\mathbf{X}}_p^{diff}$ for each player:

$$M_{freq}(\mathbf{X}_p^{diff}) = \left\{(q_i, \mathbf{x}_i') | \mathbf{x}_i' \in \hat{\mathbf{X}}_p^{diff}\right\} \tag{7}$$

This relative frequency is then treated as the probability that a player p is represented by feature $\mathbf{x}_i'$, i.e. training consists of calculating relative frequencies of features for all players.

4.4. Distance functions between features and models

The last modeling option in the assignment problem framework is the weight function $\mathbf{C} : O \times P \rightarrow \mathbb{R}$, i.e. a cost matrix that represents the costs of assigning an observed feature to different models and therefore players. Its choice is not arbitrary, as it is usually strongly tied to the type of model and the type of feature used. For histogram features, the inverse relative frequency of the histogram is used as a distance measure between features and models:

$$d_f(\mathbf{x}_o, M_{freq}) = \begin{cases} 1 - q_i & \text{if } \exists (q_i, \mathbf{x}_i') \in M_{freq} : \mathbf{x}_o = \mathbf{x}_i' \\ 1 & else \end{cases} \tag{8}$$

Additionally, where the sum of bins is less than 10 (e.g. some players are not visible in the image sequence), the sum of all histogram probabilities of the valid model $V(\mathbf{x}_o, M_{freq})$ is used. A model histogram $\mathbf{x}_i'$ is considered valid for a given histogram $\mathbf{x}_o$ if the element-wise minimum vector between the two histograms $\min(\mathbf{x}_o, \mathbf{x}_i')$ equals the histogram $\mathbf{x}_o$:

$$V(\mathbf{x}_o, M_{freq}) = \left\{(q_i, \mathbf{x}_i') | (q_i, \mathbf{x}_i') \in M_{freq} : \min(\mathbf{x}_o, \mathbf{x}_i') = \mathbf{x}_o\right\} \tag{9}$$

The distance function is then computed as the sum over the likelihoods of all valid model histograms:

$$d_h(\mathbf{x}_o, M_{freq}) = 1 - \sum_{(q_i, \mathbf{x}_i') \in V(\mathbf{x}_o, M_{freq})} q_i \tag{10}$$

With this distance function, the previously presented feature representation and player model, the Hungarian method is applied to solve the assignment problem.

5. Jersey number recognition using convolutional neural networks

For player identification, recognizing jersey numbers is the most straightforward approach. In professional soccer leagues, every player has a unique jersey number throughout a complete season, allowing an unambiguous mapping from jersey number to player and vice versa.

A convolutional neural network is trained to recognize these numbers in a grayscale image region of 40 × 40 pixel. Grayscale images are used to avoid that the neural network erroneously learns color as a feature for certain numbers, as there might be only few distinct players that share rare jersey numbers. In those cases, the network might learn that for a certain number, the jersey color is a discriminative feature. Only image regions where a jersey number may occur are considered, i.e. the upper half of a player bounding box, scaled to the stated size. Within this upper half of the bounding box, the precise location of the jersey number is not annotated. In the dataset, bounding box size varies between 48 and 110 pixels. They are scaled to 64 × 64 pixels and cropped (from the top, left and right) to 40 × 40 to reduce the influence of background noise.

For the convolutional neural network, all occurring jersey numbers are modeled as separate classes. In our case, this results in a 40-class classification problem (as there are 40 different jersey numbers in the entire dataset). Also, not all one- or two-digit numbers appear in the dataset. Additionally, three classes are introduced for cases when the jersey number is either *not visible* (class 41), *not readable* (e.g. due to motion blur, class 42) or a *box error* occurs (class 43) for those image regions where it does not depict the upper body of a player. The latter case occurs when automatic player detection fails or delivers imprecise results. That means, that the classifier $c(z)$ assigns exactly one class (number) y to each input image region z:

$$c(z) = y, \qquad y \in \{1, 2, 3, \ldots, 40, 41, 42, 43\} \tag{11}$$

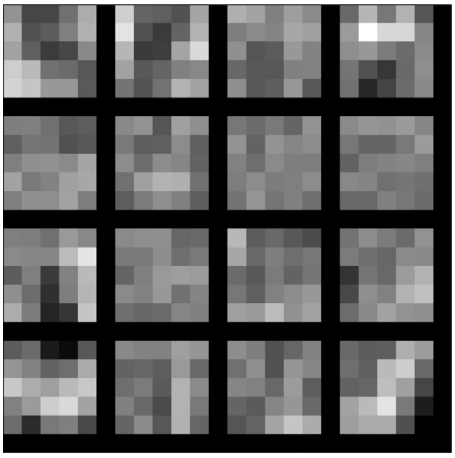

Fig. 8. 16 5 × 5 filter weights of the first convolutional layer.

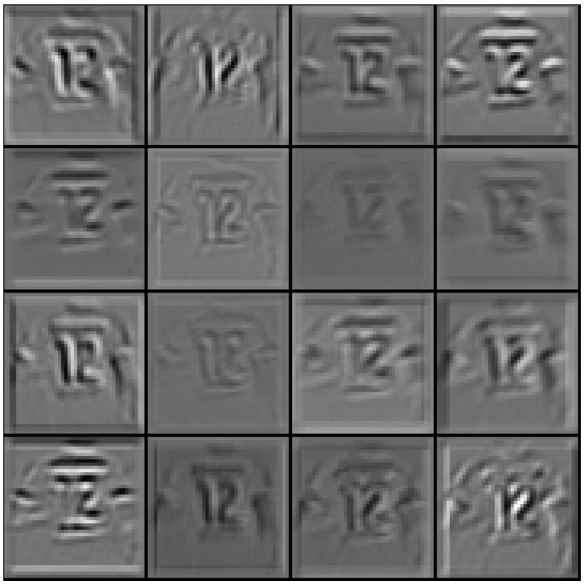

Fig. 9. Output of the first convolutional layer for a sample image.

The Chollet (2015) library for deep neural networks was used for the experiments. The architecture of the convolutional neural network is inspired by models for generic image classification (similar to a model for the CIFAR-10 Krizhevsky, 2009 dataset) and house number recognition in street view images (using the street view house number dataset (SVHN)).

The network architecture consists of three convolutional layers (with 16 × 5 × 5/30 × 7 × 7/50 × 3 × 3 parameters), each with rectified linear units (ReLU) as activation function, followed by a max-pooling layer. Then, three fully connected layers with ReLU activation follow, with a final softmax activation layers.

Samples for the learned convolutional filters are given in Fig. 8 and outputs for those filters for a sample image are shown in Fig. 9. For the fully connected layers, dropout (Srivastava et al., 2014) is used for regularization.

Table 1 gives the details of the network architecture. The convolutional stride is always set to one pixel, while pooling size and stride is set to two pixels for the first convolutional layer and three pixels for the remaining convolutional layers. In contrast, the network architecture in Srivastava et al. (2014) for the SVHN dataset uses more filter channels ((96, 128, 256) instead of (16, 30, 50) used here) for the convolutional layers. The two fully connected layers in Srivastava et al. (2014) consists of 2048 units each, while only 34 units are used in this work, The reason for reducing the number of units is mainly due to a smaller dataset. The SVHN dataset is two orders of magnitude larger (as an extended training corpus of the SVHN dataset is used) than the jersey number dataset used here. In order to compensate for the relatively small dataset, data augmentation is performed. For every training sample in the training set, five modified variants of the original sample are artificially generated by inverting the image luminance, performing random scaling and translation. Sample classification results of the network described here are shown in Fig. 10.

To embed the output of the convolutional neural network for jersey number recognition into the assignment problem, not all outputs are utilized. Instead, only the outputs for the jersey numbers of players P which are known to be present on the pitch are used. The outputs for all bounding boxes within a single trajectory are averaged to obtain a single 11-element vector $\mathbf{v}$ (for eleven known players). All vectors of observed visible players O_{vis} at a single time point t_i are combined to a matrix $\mathbf{C}$, which is then used as cost matrix for the assignment problem:

$$\mathbf{C} = [\mathbf{v_1}, \ldots, \mathbf{v_k}, \mathbf{0}, \ldots, \mathbf{0}]^T \tag{12}$$

As described in Section 3, costs from hidden players O_{hid} to known players are set to $\mathbf{0}$.

6. Classifier fusion

In order to combine recognition based on positional features with jersey number recognition, both modalities are fused. There are mainly three types of fusion possible: Feature fusion, cost matrix fusion and classifier (i.e. assignment) fusion. In our approach, cost matrix fusion is used. This allows for different kinds of cost functions for each modality, which would be more difficult to implement when performing feature fusion. And it has the advantage that the combined solution represents an optimal solution in terms of the assignment problem. This does not hold if two separate assignment problems are solved and the results are combined. Especially contradicting assignments from two separate solutions are hard to resolve correctly in an optimal way.

Both modalities each yield an $n \times n$ cost matrix $\mathbf{C_{JN}}$ and $\mathbf{C_{PF}}$ for jersey number and positional features respectively. In order to obtain comparable cost matrices, both matrices are row-normalized to represent probability distributions.

For jersey number recognition, the quality of the estimation is incorporated into the cost matrix. Therefore, the probability $p_{invalid}$ that the jersey number is not even visible is obtained from the CNN output. It is estimated as the ratio of class probabilities for non-number classes (one of *not visible*, *multiple players*, *not readable*), over all class probabilities:

$$p_{invalid}(o_i) = \frac{\sum_{c' \in \{\mathrm{nv,mp,nr}\}} \mathrm{CNN}(o_i, c')}{\sum_{c \in C} \mathrm{CNN}(o_i, c)} \tag{13}$$

where $\mathrm{CNN}(o_i, c)$ is the output of the convolutional neural network given input o_i and desired output class c. This is incorporated into the cost matrix by weighting each row of the jersey number cost matrix with $m_i = 1 - p_{invalid}(o_i)$. The weight matrix $\mathbf{M_{JN}}$ is composed as

$$\mathbf{M_{JN}} = \begin{bmatrix} m_1 & \ldots & m_1 \\ \vdots & & \vdots \\ m_n & \ldots & m_n \end{bmatrix} \tag{14}$$

For the constellation model, there is no way of predicting the quality of the estimation. Therefore, its weight matrix is set to ones:

$$\mathbf{M_{PF}} = \mathbf{1} \tag{15}$$

Table 1
Deep convolutional network architecture.

Stage	1	2	3	4	5	6
Layer type	conv + max	conv + max	conv + max	full	full	full (out)
# channels	16	30	50	34	34	45/15
Filter size	5 × 5	7 × 7	3 × 3	–	–	–
Conv. Strides	1 × 1	1 × 1	1 × 1	–	–	–
Pooling Size	2 × 2	3 × 3	3 × 3	–	–	–
Pooling Str.	2 × 2	3 × 3	3 × 3	–	–	–
Input Size	40 × 40	20 × 20	6 × 6	2 × 2	–	–

Fig. 10. Sample classification results using the best configuration. Each column shows random results for the classes 2, 3, 4, 6, 8, 10, 13, 15, 16, 21, 20 and 25.

Table 2
Correlation between the optimal weight (for jersey number recognition) and different properties of the shots and its trajectories.

Property	# players s	$\bar{x}$	σ_x	$\bar{y}$	σ_y	length l
Correlation	−0.10	0.22	−0.35	0.02	0.04	−0.49

Then, weighted cost matrices $\mathbf{C'_{JN}} = \mathbf{M_{JN}} \odot \mathbf{C_{JN}}$ and $\mathbf{C'_{PF}} = \mathbf{C_{PF}}$ (according to (15)) are used to calculate the final cost matrix C. Here, a weighted average is used as

$$\mathbf{C} = w_j \cdot \mathbf{C'_{JN}} + (1 - w_j) \cdot \mathbf{C'_{PF}} \tag{16}$$

for solving the assignment problem. w_j is then globally optimized for all shots in a leave-one-out cross validation.

Additionally, methods for predicting shot-wise optimal weights w_j are evaluated. Therefore, a correlation analysis between different shot and trajectory attributes and the optimal weight w_j is performed. This includes the number s of players visible during the shot (# players), the shot length l, the average over the distances to the middle line in x and y direction ($\bar{x}$, $\bar{y}$), and the standard deviations (σ_x, σ_y) of all x and y positions over all observations. The correlations are shown in Table 2. As there are several non-zero correlations, predicting the optimal weight given these properties seems promising.

The prediction of the optimal weight w_j, given a vector of these shot and trajectory attributes

$$\mathbf{a} = (l, s, \bar{x}, \bar{y}, \sigma_x, \sigma_y), \tag{17}$$

is posed as a regression problem. The task is to predict the jersey number weight w_j given the attribute vector a using the prediction function f:

$$w_j = f(\mathbf{a}) \tag{18}$$

Support vector regression with a linear kernel is used to train a regression model. A leave-one-out cross-validation over 14 shots is performed to train on 13 shots and predict the optimal weights for the respective hold-out shot. Finally, the weighted cost matrix is used to solve the assignment problem.

7. Experimental results

In order to assess the presented player identification approach independently of previous steps in a video processing pipeline (e.g. automatic player tracking), it is based on manually created player trajectories. This avoids that errors in an automatic tracking process are propagated to the player identification. For this manual labeling, automatic player detection on regularly sampled video frames (every second) was performed. Then, human annotators were asked to associate player bounding boxes in consecutive frames to generate the trajectories.

The remainder of this section is organized as follows. First, the test set and its properties are described in detail. Then, details on the evaluation metric are presented. Subsequently, the training corpus and the results for recognition based only on constellation features are shown, followed by those descriptions for only jersey number recognition. Finally, combined results are presented.

7.1. Test set

As there are no publicly available datasets of soccer videos and corresponding player trajectories, six different soccer videos were recorded and player trajectories were manually annotated for 14 shots. These trajectories were sampled at 1 Hz. The videos were recorded full-length broadcasts from the 2012/2013 German Bundesliga season. The lengths of these shots vary between 12 s and 62 s. They contain in total 8122 different player appearances in 654 different trajectories. The goal is to identify all player identities, whenever visible.

Table 3 gives a more detailed overview over the different video shots of the test set. The fact that even overview shots in live TV broadcast are usually at most one minute long makes the collection of long-term trajectories intractable. However, longer trajectories are supposed to improve identification accuracy. They would increase the probability for each trajectory that a jersey number is visible in some frames; and it would allow for a less variable position model, as short-term deviations from tactical roles are com-

Table 3
Test set used for evaluating player identification.

ID	*Duration*	*# Frames*	*Trajectories*	*Player App.*
1	0:14	14	25	188
2	0:34	34	51	421
3	0:37	37	54	585
4	0:50	50	84	733
5	0:12	12	17	218
6	0:21	21	25	489
7	0:44	44	42	592
8	0:32	32	39	567
9	0:38	38	43	472
10	1:02	62	70	1154
11	0:48	48	56	723
12	0:50	50	63	940
13	0:21	21	32	464
14	0:33	33	53	576

pensated. Another property of the dataset is the fact that there are usually more trajectories than players on the field. This is caused by players leaving and re-entering the camera's field-of-view. Having more trajectories than players makes the task even harder, for the same reasons mentioned above for the limited shot lengths.

7.2. Evaluation

The identification score for one video shot is calculated as the weighted accuracy for all its trajectories T. For each trajectory, it is evaluated if the predicted player assignment function $\hat{f}(T_i)$ is correct according to a ground truth assignment $f^*(T_i)$. To calculate the accuracy for all assignments within one shot, the set of correctly assigned trajectories T_{corr} is defined as:

$$T_{corr} = \{T_i | T_i \in T : \hat{f}(T_i) = f^*(T_i)\} \tag{19}$$

The accuracy for each trajectory is weighted by its length, i.e. the number of single observations it contains:

$$acc(\hat{f}) = \frac{\sum_{T_j \in T_{corr}} |T_j|}{\sum_{T_i \in T} |T_i|} \tag{20}$$

This reduces the influence of short trajectories, such that a trajectory that appears for one second has a smaller influence on the overall accuracy than a significantly longer trajectory.

7.3. Player recognition using constellation features

When performing player recognition by only using constellation features, training is performed on the dataset described in Gerke and Müller (2015). As described in Section 7.1, the training and test set have different temporal resolution (1 min and 1 s respectively). Additionally, not all players on the pitch are always visible. This is handled by assigning players that are not visible to their last known position. For the experiments, a 2 × 2 spatial bin specification is used. While additional bins might increase the classifiers capability, we found that it is more robust for non-visible players if only 2 × 2 bins are used.

Using only constellation features, a length-weighted accuracy of 0.49 is obtained. For the different shots in the test set, the accuracy varies between 0.13 and 0.86, i.e. the accuracy variation is quite large. Table 4 gives the weighted accuracy for all shots within the testset. For very short shots of up to 600 frames (i.e. 24 s), the constellation based recognition performs rather poor. There are basically two reasons for this behaviour. The shorter the shot, the higher the probability that not all teammates are visible within the duration of the shot. Furthermore, a longer shot describes a more precise description of a player's spatial behaviour. Outliers, where players move away from their typical position, have a higher (negative) impact in shorter shots. Other conclusions (e.g. the existence of a lower bound in shot length for good constellation feature results) cannot be easily drawn from these results. It is e.g. not possible to predict the accuracy from the shot length. There is a positive correlation of 0.52 between the length of a shot and the accuracy of constellation based identification, but accuracies oscillate significantly even among shots with similar lengths.

Table 4
Shot-wise accuracy (mean over all frames) for player identification based on constellation features and jersey numbers (JN) respectively, with different weights for jersey numbers given. The best result for each shot is highlighted.

ID	Accuracy @ JN weight				opt. weight
	0.0	1.0	0.9	opt.	
1	0.24	0.67	**0.72**	–	0.9
2	0.41	0.67	**0.85**	–	0.9
3	0.37	0.72	0.80	**0.81**	0.95
4	0.35	0.48	**0.62**	–	0.9
5	0.18	0.70	**0.77**	–	0.9
6	0.13	0.77	0.79	**0.82**	0.95
7	0.56	0.75	0.95	**0.95**	0.85
8	0.58	0.76	**0.90**	–	0.9
9	0.15	0.53	0.57	**0.61**	0.85
10	0.44	0.71	**0.94**	–	0.9
11	0.59	0.69	0.81	**0.83**	0.7
12	0.84	0.79	0.93	**0.94**	0.6
13	0.59	0.78	0.91	**0.91**	0.8
14	0.86	0.64	0.83	**0.98**	0.7
Weighted Mean	0.49	0.69	0.82	0.83	

7.4. Player recognition using jersey number recognition

As shown in Table 4, using jersey number recognition alone yields better results compared to using constellation features only (weighted accuracy: 0.69 vs. 0.49). Additionally, Table 4 shows the result for each shot separately. The variations between different shots are smaller than those for constellation features, they range between 0.48 and 0.79. For most shots, they even vary only between 0.65 and 0.75, making the performance of the jersey number recognition more predictable. We believe that this is mostly due to the weighting scheme that incorporates the confidence of the CNN that a jersey number is visible at all. Interestingly, there is no significant positive correlation between the length of a shot and the accuracy for this shot, it is even slightly negative (−0.09). One could expect better results for longer shots, as longer shots increase the probability that jersey numbers are visible.

7.5. Combined player recognition

As described in Section 6, a weighted combination of the two cost matrices, one for each modality, is used. Therefore, the weight for constellation features $w_c = 1 - w_j$ is calculated as the counter weight to the jersey number recognition weight w_j. Relative weights are varied between $w_j = 0.0$ (which means that only constellation features are used) and $w_j = 1.0$ (only jersey numbers are used). The accuracy per shot depending on the weight distribution between constellation features and jersey numbers is visualized in Fig. 11. Here, each regular line represents one shot of the test set. The black bold line is the weighted (by shot length) average over all shots. The average accuracy reaches its optimum of 0.82 with jersey number weight $w_j = 0.9$ (and accordingly constellation feature weight $w_c = 1 - w_j = 0.1$). However, single shots have slightly higher accuracies at different weights. Optimizing the weights for each shot separately only slightly increases the accuracy to 0.83, even though accuracy is used as the optimization criterion. However, accuracy is usually not available a priori without ground truth labels. Thus, this result serves as an upper bound of

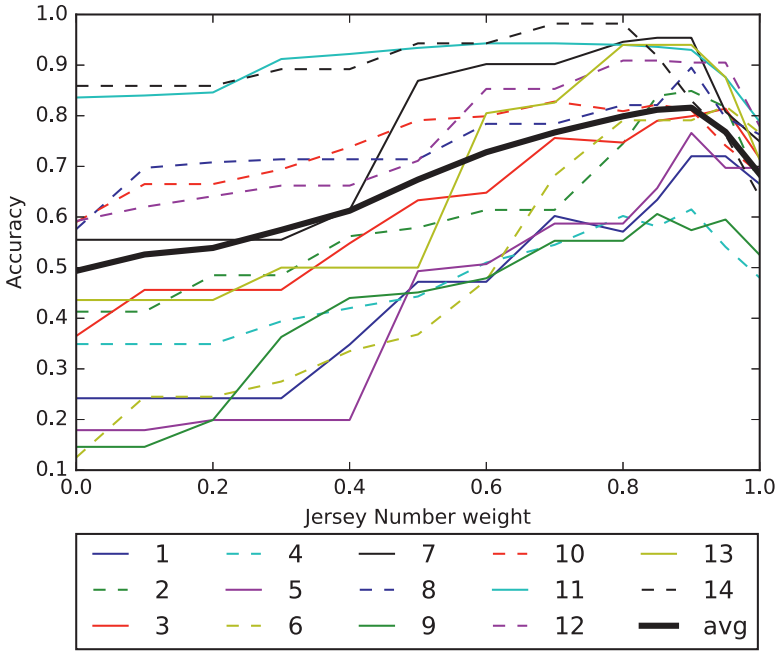

Fig. 11. Accuracy for varying weights for jersey numbers and constellation features. Each line represents one shot. The black bold line is the weighted (by shot length) average over all shots.

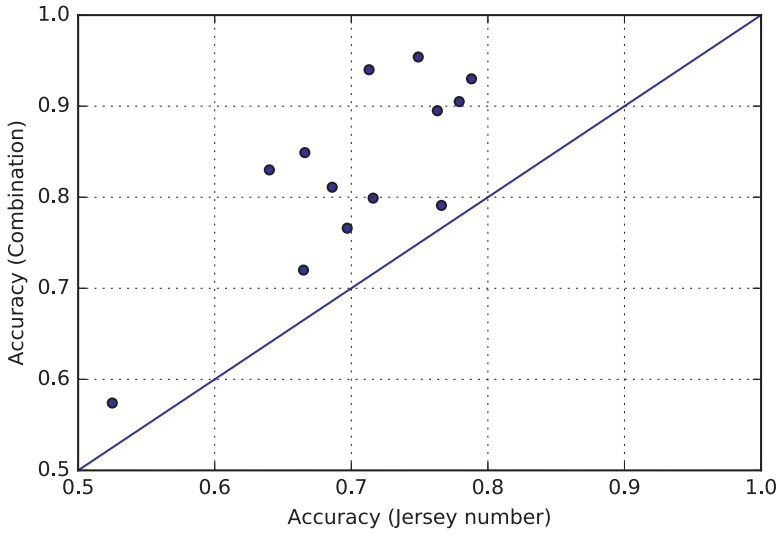

Fig. 12. Accuracy for each shot with jersey number recognition alone (x axis) and combined with constellation features (y axis) with weight $w_j = 0.9$ for jersey numbers and $w_c = 0.1$ for constellation features. Each shot is represented by one dot.

the expected gain when predicting better weights based on available shot and trajectory attributes such as shot length and spatial properties of trajectories.

The correlations presented in Section 6 give an indication that it might be feasible to predict optimal weights using the mentioned attributes. However, as stated above, the expected improvement is quite low, as we found an accuracy of 0.83 as an upper bound, giving only a slight improvement over a fixed $w_j = 0.9$ (accuracy 0.82).

As described in Section 6, a support vector regression model is trained that predicts a weight distribution per shot, given its attributes. In order to predict weights for a single shot, leave-one-out cross validation is used. When applying this regression model to predict weights, the accuracy (0.81) over all shots does not reach that of a constant weight $w_j = 0.9$ (0.82). Accordingly, a constant weight $w_j = 0.9$ for jersey numbers and $w_c = 0.1$ for constellation features is recommended.

Fig. 12 plots the accuracy for using only jersey numbers (x axis) over the accuracy for using both features in combination (y axis, with $w_j = 0.9$). Each dot represents the results for a single shot, while the blue line indicates equal accuracy between jersey numbers and fused results. All dots are above the blue line, meaning that for all shots, an improvement is obtained by using a combination of constellation features and jersey numbers over the exclusive usage of jersey numbers. As shown in Table 4, combining both modalities improves the average accuracy from 0.69 with jersey numbers only to 0.82 using both jersey numbers and constellation features.

8. Conclusion

In this paper, a method for player identification in broadcast soccer videos is presented. For this, a combination of jersey number recognition and spatial constellation of players on the pitch is described. The general problem of identifying players in overview shots of TV soccer broadcast is posed as a piece-wise assignment problem. Within this assignment problem, both modalities are integrated by combining cost matrices for jersey number recognition and constellation features respectively. Using spatial constellation as a single means of identifying soccer players does not work very well, yielding an accuracy of only 0.49. This is expected, as soccer players do not strictly follow tactical patterns. They rather temporarily switch positions in order to create confusion among the opponent's players. This is especially true for more offensive players which are supposed to be more creative and unpredictable in their behaviour.

When evaluating jersey number recognition using convolutional neural network, an accuracy of 0.69 is obtained. Here, the problem is that often not all jersey numbers are visible or the visible ones are hard to identify due to motion blur. By combining these two modalities, an improved identification accuracy of 0.82 is reached.

While the results encourage the application of the proposed combined approach towards a fully automatic metadata generation for broadcast soccer videos, they leave further room for improvement. The fact that overview shots are often interrupted by e.g. close-up shots, re-initialization is required after each shot break. Here, approaches that link players over shorter periods of non-overview shots could improve player motion models. Another challenge is the handling of players that are not visible for certain periods, which is common for soccer broadcast material. Predicting the position of invisible players could improve the quality of spatial constellation features.

References

Andrade, E., Khan, E., Woods, J., Ghanbari, M., 2003. Player identification in interactive sport scenes using region space analysis prior information and number recognition. In: International Conference on Visual Information Engineering (VIE). IEE, pp. 57–60.

Angehrn, F., Wang, O., Aksoy, Y., Gross, M., Smolic, A., 2014. MasterCam FVV: robust registration of multiview sports video to a static high-resolution master camera for free viewpoint video. In: IEEE International Conference on Image Processing, ICIP, pp. 3474–3478.

Ballan, L., Bertini, M., Bimbo, A.D., Nunziati, W., 2007. Soccer players identification based on visual local features. In: 6th ACM International Conference on Image and Video Retrieval, pp. 258–265.

Beetz, M., Kirchlechner, B., Lames, M., 2005. Computerized real-time analysis of football games. IEEE Pervasive Comput. 4 (3), 33–39. doi:10.1109/MPRV.2005.53.

Bialkowski, A., Lucey, P., Wei, X., Sridharan, S., 2013. Person re-identification using group information. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–6.

Chollet, F., 2015. Keras. https://github.com/fchollet/keras.

Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., Malvaldi, M., 2015. The harsh rule of the goals: data-driven performance indicators for football teams. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10.

Couceiro, M., Clemente, F., Martins, F., Machado, J., 2014. Dynamical stability and predictability of football players: the study of one match. Entropy 16 (2), 645–674. doi:10.3390/e16020645.

Delannay, D., Danhier, N., De Vleeschouwer, C., 2009. Detection and recognition of sports(wo)men from multiple views. In: ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), pp. 1–7.

Duch, J., Waitzman, J.S., Amaral, L.A.N., 2010. Quantifying the performance of individual players in a team activity. PLoS ONE 5 (6), 1–7.

Farin, D., Han, J., de With, P.H.N., 2005. Fast camera calibration for the analysis of sport sequences. In: IEEE International Conference on Multimedia and Expo, ICME, pp. 1–4.

Frencken, W., Lemmink, K., Delleman, N., Visscher, C., 2011. Oscillations of centroid position and surface area of soccer teams in small-sided games. Eur. J. Sport Sci. 11 (4), 215–223.

Gerke, S., Müller, K., 2015. Identifying soccer players using spatial constellation features. KDD Workshop on Large-Scale Sports Analytics.

Gerke, S., Müller, K., Schäfer, R., 2015. Soccer jersey number recognition using convolutional neural networks. In: IEEE International Conference on Computer Vision Workshops.

Gudmundsson, J., Horton, M., 2016. Spatio-temporal analysis of team sports - a survey. CoRR abs/1602.06994.

Homayounfar, N., Fidler, S., Urtasun, R., 2016. Soccer field localization from a single image. In: Computer Vision and Pattern Recognition, pp. 1–16.

Krizhevsky, A., 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report.

Linnemann, A., Gerke, S., Kriener, S., Ndjiki-Nya, P., 2013. Temporally consistent soccer field registration. In: IEEE International Conference on Image Processing (ICIP), pp. 1316–1320.

Lu, C.-W., Hsu, C.-Y., Kang, L.-W., Lin, C.-Y., Weng, M.-L., Liao, H.-Y.M., 2013. Identification and tracking of players in sport videos. In: International Conference on Internet Multimedia Computing and Service (ICIMCS).

Lu, W.-L., Ting, J.-A., Little, J.J., Murphy, K.P., 2013. Learning to track and identify players from broadcast sports videos. IEEE Trans. Pattern Anal. Mach. Intell. 35 (7), 1704–1716.

Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., Sheikh, Y., 2013. Representing and discovering adversarial team behaviors using player roles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2706–2713.

Mahmood, Z., Ali, T., Khattak, S., Hasan, L., Khan, S.U., 2014. Automatic player detection and identification for sports entertainment applications. Pattern Anal. Appl. 18 (4), 971–982.

Messelodi, S., Modena, C.M., 2011. Scene text recognition and tracking to identify athletes in sport videos. Multimed. Tools Appl. 63 (2), 521–545. doi:10.1007/s11042-011-0878-y.

Munkres, J., 1957. Algorithms for the assignment and transportation problems. J. Soc. Ind. Appl.Math. 5 (1), 32–38.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Wei, X., Sha, L., Lucey, P., Carr, P., Sridharan, S., Matthews, I., 2015. Predicting ball ownership in basketball from a monocular view using only player trajectories. In: IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 780–787.

Yamamoto, T., Kataoka, H., Hayashi, M., Aoki, Y., Oshima, K., Tanabiki, M., 2013. Multiple players tracking and identification using group detection and player number recognition in sports video. In: Industrial Electronics Conference (IECON), pp. 2442–2446.

Yao, Q., Nonaka, K., Sankoh, H., Naito, S., 2016. Robust moving camera calibration for synthesizing free viewpoint soccer video. In: IEEE International Conference on Image Processing, ICIP, pp. 2–6.