Available online at www.sciencedirect.com

SciVerse ScienceDirect

AASRI Procedia 00 (2014) 000–000

AASRI Procedia

www.elsevier.com/locate/procedia

2014 AASRI Conference on Sports Engineering and Computer Science (SECS 2014)

Online Multi-player Tracking in Monocular Soccer Videos

Michael Herrmann[a,*], Martin Hoernig[a], Bernd Radig[a]

[a]*Technische Universität München, Image Understanding and Knowledge-Based Systems, Boltzmannstr. 3, D-85748 Garching, Germany*

Abstract

The tracking of players in monocular soccer videos is a challenging task because of numerous difficulties that can occur especially in TV broadcasts, such as camera motions, severe occlusion of players, or inhomogeneous lightning conditions. We propose a new robust method for multi-player tracking, which is based on finding local maxima on a confidence map. This map represents an ensemble of visual evidences, such as colors of the team outfits, responses of a HOG human detector, and grass regions in images. This combination of features allows for a robust online tracking procedure that does not require any further information about the camera calibration or other user input. In the evaluation using four representative datasets, our algorithm shows remarkable accuracy and outperforms a state-of-the-art pedestrian tracker.

Selection and/or peer review under responsibility of American Applied Science Research Institute

Keywords: computer vision, soccer analysis, player tracking

1. Introduction

We aim to develop a system for an automatic analysis of soccer matches that extracts match statistics and performs tactical analysis from monocular recordings, such as TV broadcasts, which usually exhibit numerous difficulties. One important task of such a system is the 2D tracking of players in the video images. For this purpose, we propose a new robust unsupervised online method that provides three main contributions:

* Corresponding author. Tel.: +49-89-289-17779; fax: +49-89-289-17757.
E-mail address: michael.herrmann@tum.de

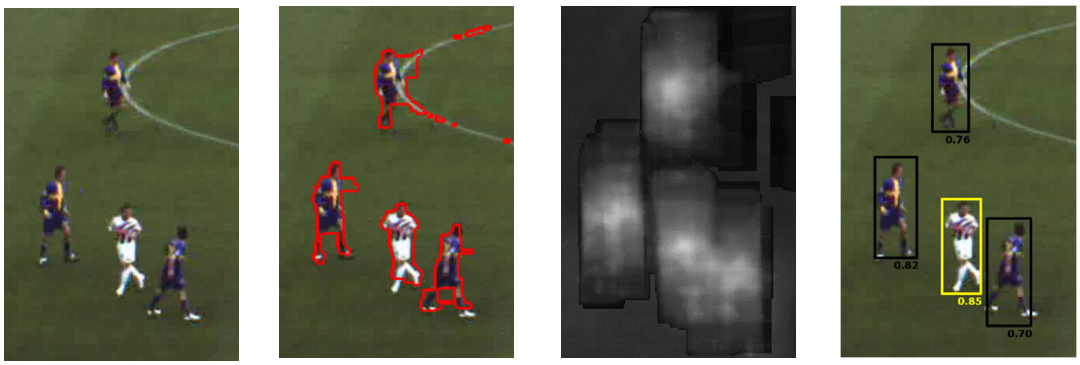

Figure 1: From left to right: original image, segmentation results, confidence map, and detection results (with confidence values).

- An observation model that is based on a combination of soccer-specific features (grass color), match-specific features (colors of team outfits), and universal features (HOG detector) resulting in a robust confidence map for player positions (see sections 2 and 3 and Figure 1).
- An efficient measurement model that finds optimal locations on a confidence map starting at predicted positions of single-target Kalman filters [1] (see section 4).
- An efficient player detection, avoiding time-consuming tracking-by-detection (see section 4).

In human tracking, the tracking-by-detection approach is widely used for multi-target tracking (see i.e. [2], [3]). Human detectors often show low accuracy and poor computational performance for small objects, while in soccer videos, especially in low-resolution records and wide-angle scenes, player heights of 40 pixels and less are quite common. In contrast, kernel-based approaches, such as the mean-shift tracking [4], often require an initialization of the templates and risk drifting away, due to object-specific template adaption.

Zhang et al. [5] proposed an online method with a combination of tracking-by-detection and kernel-based tracking and achieved state-of-the-art results using standard pedestrian tracking data sets.

In principle, our approach is similar, but we use soccer-specific knowledge to automatically detect players without an exhaustive sliding-window approach. Our observation model is partially based on the ideas of ASPOGAMO [6], but incorporates more universal features, such as the HOG detector, and does not need any human-guided initialization. It is similar to the method proposed by [7], which is, however, a stand-alone player detection and does not incorporate tracking information over time.

2. Preliminaries

Player positions. In our approach, the position of a player p_i is modeled within an image $\boldsymbol{I^t}$ at time t using an axis-aligned bounding box $\boldsymbol{B_i} \coloneqq \{(x,y) \in \mathbb{N}^2 | (x_i \leq x < x_i + w_i) \wedge (y_i \leq y < y_i + h_i)\}$, where (x_i, y_i) is the upper-left corner and w_i and h_i are width and height, respectively. A bounding box describes the extent of a player in the image and in our method the aspect ratio is fixed to $w_i \coloneqq 0.41 h_i$ (see [8]).

Player segmentation. Our confidence map is based on segmented player regions in the image and we use a robust grass segmentation method following [9]. This procedure takes as input a RGB image $\boldsymbol{I^t}$ at time t and returns an image region $\boldsymbol{H^t}$, describing the enclosing hull of the playing field, and a foreground region $\boldsymbol{F^t}$, which, under ideal circumstances, covers the image region of all players in the visible part of the playing field (see Figure 1). Due to noise and movement artefacts usually not all players are segmented properly, whereas parts of other foreground objects, such as field lines, can be included.

Unsupervised generation of color templates. According to the FIFA rules [10], the colors of the jerseys have to be chosen so that the players from different teams, the referees, and the goal keepers are all pair wise

distinguishable. We make use of this fact by initially determining color templates for each type of outfit, based on color histograms.

Given the first frame $\boldsymbol{I^0}$ of a video sequence, our automatic detection procedure (see section 5) returns player positions represented by a set of n_0 bounding boxes $\boldsymbol{B^0} := \{\boldsymbol{B_1^0}, \dots, \boldsymbol{B_{n_0}^0}\}$. We extract the set of RGB color vectors of $\boldsymbol{I^0}$ for the pixels within $(\boldsymbol{B_1^0} \cup \dots \cup \boldsymbol{B_{n_0}^0}) \cap \boldsymbol{F^0}$ and cluster this set by applying the k-means++ algorithm [11] using the Euclidean distance. The cluster centers are the k_C dominant color vectors that represent the k_C bins of our color histograms. A color vector is assigned to the bin that is represented by the nearest dominant color. We divide each bounding box $\boldsymbol{B_i^0}$ to its equal-sized top, mid, and bottom parts $^{\boldsymbol{T}}\boldsymbol{B_i^0}$, $^{\boldsymbol{M}}\boldsymbol{B_i^0}$, and $^{\boldsymbol{B}}\boldsymbol{B_i^0}$ and calculate three histograms with respect to $^{\boldsymbol{T}}\boldsymbol{B_i^0} \cap \boldsymbol{F^0}$, $^{\boldsymbol{M}}\boldsymbol{B_i^0} \cap \boldsymbol{F^0}$, and $^{\boldsymbol{B}}\boldsymbol{B_i^0} \cap \boldsymbol{F^0}$. This takes into account that a player's outfit changes from top to bottom, regarding jersey, shorts and socks. Stacking and normalizing result in a feature unit vector with $3k_C$ entries for each bounding box.

To calculate the distance of two normalized histogram vectors $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$, we use the Hellinger distance as proposed by [4] and define $d_H(\boldsymbol{h_1}, \boldsymbol{h_2}) := \sqrt{d_H^2(\boldsymbol{h_1}, \boldsymbol{h_2})}$ with $d_H^2(\boldsymbol{h_1}, \boldsymbol{h_2}) := 1 - \sum_{i=1}^{3k_C} \sqrt{\boldsymbol{h_1}(i)\boldsymbol{h_2}(i)}$.

We try to estimate the number of different outfits among the given bounding boxes by clustering the set of histogram vectors using k-means++ with $k = 1, \dots, 5$ (there are five different possible outfits: 2 x players, 2 x goalkeepers, 1 x referees) using the distance d_H. We choose k_{min} as the smallest k for which the result of the clustering fulfills our decision criterion: for a given set of l histogram vectors $\boldsymbol{h_1}, \dots, \boldsymbol{h_l}$ and their k cluster centers $\boldsymbol{o_1}, \dots, \boldsymbol{o_k}$, each vector $\boldsymbol{h_i}$ is assigned to its cluster center $\boldsymbol{o}(\boldsymbol{h_i})$ and we define the minimum distance as $d_{min} := \min_i d_H(\boldsymbol{h_i}, \boldsymbol{o}(\boldsymbol{h_i}))$ and the maximum distance as $d_{max} := \max_i d_H(\boldsymbol{h_i}, \boldsymbol{o}(\boldsymbol{h_i}))$. With the threshold parameters t_{min} and t_{max}, our decision criterion is $(d_{min} \leq t_{min}) \wedge (d_{max} \leq t_{max})$. During tracking each tracked bounding box is assigned to the outfit class with nearest distance d_B at the moment of detection.

HOG-based human detection. We trained a human detector based on histograms of oriented gradients (HOG) according to [12] with some slight modifications to their default detector. We use a 64×128 pixel detection window with a human size of 41×100 pixel (see [8]). To allow for an efficient calculation, we do not apply a Gaussian spatial window during the accumulation of the histograms. To avoid soccer-specific overfitting, the classifier is trained using the INRIA pedestrian data set [12].

3. Confidence map for player positions

We construct a function that describes a degree of quality of the image evidence $\boldsymbol{I^t}$, given a set $\boldsymbol{B^t} := \{\boldsymbol{B_1^t}, \dots, \boldsymbol{B_{n_t}^t}\}$ of bounding boxes at time t. We assign each bounding box $\boldsymbol{B_i^t} \in \boldsymbol{B^t}$ to the connected component of the foreground $\boldsymbol{F^t}$ with the greatest overlap. If $\boldsymbol{B_i^t} \cap \boldsymbol{F^t} = \emptyset$, we assign $\boldsymbol{B_i^t}$ to an empty region. Hence, the result is a set of image regions $\boldsymbol{S^t} := \{\boldsymbol{R_1^t}, \dots, \boldsymbol{R_{m_t}^t}\}$, where each $\boldsymbol{R_j^t} \in \boldsymbol{S^t}$ has a set of assigned bounding boxes ${}_j\widetilde{\boldsymbol{B}}^t := \{{}_j\widetilde{\boldsymbol{B}}_1^t, \dots, {}_j\widetilde{\boldsymbol{B}}_{n_R}^t\}$. In the following, fractions are only calculated if the denominator is not zero. Otherwise, they are omitted. We perform all calculations for each region $\boldsymbol{R_j^t}$ and its assigned bounding boxes independently. For the purpose of clearness we omit the index j and set $n_R := n(\boldsymbol{R_j^t}) \in \mathbb{N} \setminus \{0\}$. In most cases $n_R = 1$. Cases with $n_R > 1$occur for example if one player is partly occluded by another player in the image.

Region-based evidence. Our region-based features are inspired by the compactness constraint and the size constraint of [6] and are based on the assumption that an image region of one or more players should be fully covered by bounding boxes. The coverage feature c_1 describes the degree of coverage of the connected component $\boldsymbol{R^t}$ by the union of the bounding boxes and is defined as

$$c_1(\widetilde{\boldsymbol{B}}_1^t, \dots, \widetilde{\boldsymbol{B}}_{n_R}^t, \boldsymbol{R^t}) := \text{area}\big((\widetilde{\boldsymbol{B}}_1^t \cup \dots \cup \widetilde{\boldsymbol{B}}_{n_R}^t) \cap \boldsymbol{R^t}\big)/\text{area}(\boldsymbol{R^t}). \quad (1)$$

The over-coverage feature c_2 penalizes regions inside the bounding boxes that are not covered by the foreground region. To favor positions where $\boldsymbol{R^t}$ is evenly spread along the horizontal axis, each bounding box $\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}$ is divided to the upper half ${}^{\boldsymbol{u}}\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}$ and the lower half ${}^{\boldsymbol{l}}\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}$ and we define

$$c_2(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t}) \coloneqq 1 - \max_{\mathrm{i}} \max_{\mathrm{o}\in\{\mathrm{u,l}\}} \mathrm{area}({}^{o}\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}} \setminus \boldsymbol{R^t}) / \mathrm{area}({}^{o}\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}). \tag{2}$$

The overlap feature c_3 takes into account that in general the bounding boxes of two players do not fully overlap (except in the rare case of full occlusion). It is defined as

$$c_3(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t}) \coloneqq 1 - \max_{\mathrm{i,j}} \mathrm{area}(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}} \cap \widetilde{\boldsymbol{B}}_{\boldsymbol{j}}^{\boldsymbol{t}}) / \mathrm{area}(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}} \cup \widetilde{\boldsymbol{B}}_{\boldsymbol{j}}^{\boldsymbol{t}}). \tag{3}$$

Color-based evidence. The color-based confidence c_4 is based on the dominant colors and the outfit classes (see section 2). The normalized color histogram vector $\boldsymbol{h}(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}) \in \mathbb{R}^{3kc}$ is calculated with respect to $\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}} \cap \boldsymbol{R}_{\boldsymbol{c}}^{\boldsymbol{t}}$. We set $\boldsymbol{R}_{\boldsymbol{c}}^{\boldsymbol{t}} \coloneqq \boldsymbol{R^t}$, if $\mathrm{area}(\boldsymbol{R^t}) > 0$. Otherwise, $\boldsymbol{R}_{\boldsymbol{c}}^{\boldsymbol{t}}$ is the largest axis-aligned inner ellipse of $\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}$. Each bounding box has an assigned outfit class with histogram vector $\boldsymbol{o}(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}) \in \mathbb{R}^{3kc}$ and we define

$$c_4(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t}) \coloneqq 1 - \sqrt[n_R]{\Pi_{i=1}^{n_R}\, d_H\big(\boldsymbol{h}(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}), \boldsymbol{o}(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}})\big)}. \tag{4}$$

HOG-based evidence. We use a human detector, trained as described in section 2. A resized sub-image with size 64×128 pixel is generated for each bounding box $\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}$, so that the height of the bounding box corresponds to 100 pixels in the sub-image, the sub-image has the same center point as the bounding box, and the aspect ratio of the pixels remains unchanged. This sub-image is used to perform a classification of the human detector. The result of the classifier is a decision value $d_D(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}) \in \mathbb{R}$, which is mapped using the parameters u_D and v_D to $d_D'\big(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}} | u_D, v_D\big) \coloneqq (d_D(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}) - u_D)/(v_D - u_D)$. The confidence is defined by the geometric mean

$$c_5\big(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t} | u_D, v_D\big) \coloneqq \sqrt[n_R]{\Pi_{i=1}^{n_R} \mathrm{max}(1, \mathrm{min}(0, d_D'\big(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}} | u_D, v_D\big))}. \tag{5}$$

Gating. During tracking we additionally incorporate a gating term, based on the distance to the predicted location of the tracker (if available). For each bounding box $\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}$, we calculate the Euclidean distance of its center point to the center point of the corresponding predicted bounding box $\widetilde{\boldsymbol{P}}_{\boldsymbol{i}}^{\boldsymbol{t}}$, normalized by the length of the diagonal of $\widetilde{\boldsymbol{P}}_{\boldsymbol{i}}^{\boldsymbol{t}}$ and trimmed to a maximum value of 1. The gating confidence c_6 is defined by one minus the mean distance of the bounding boxes $\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}$ and their corresponding predictions.

Ensemble averaging. The single features $c_i, i \in \{1, \dots 6\}$ are combined by weighted ensemble averaging using the weights $w_i, i \in \{1, \dots 6\}$, resulting in the overall confidence function $c(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t})$, which we define as follows:

$$c(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t}) \coloneqq \Sigma w_i c_i(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t}) / \Sigma w_i. \tag{6}$$

Note that $c(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t}) \in [0; 1]$ because $\forall i: c_i\big(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}^{\boldsymbol{t}}, \dots, \widetilde{\boldsymbol{B}}_{\boldsymbol{n_R}}^{\boldsymbol{t}}, \boldsymbol{R^t}\big) \in [0; 1]$. For $\boldsymbol{R^t} = \emptyset$, we set $w_1 \coloneqq 0$ and $w_2 \coloneqq 0$ and if there is only one object ($n_R = 1$), we omit the overlap feature and set $w_3 \coloneqq 0$.

Maximization of the confidence. For every new time step, we search for positions with a local maximum on the confidence map near the Kalman predictions. To keep the number of calculations low, we propose a

simple greedy heuristic: we iteratively optimize the confidence with respect to the positions of the bounding boxes, while their sizes and the region $\boldsymbol{R^t}$ remain fixed, thus for n_t bounding boxes, we have $2n_t$ variables. In each iteration of the maximization procedure, we evaluate the confidence map at $2n_t + 1$ positions near the current position and determine the position with the highest confidence value. If this value is higher than the value at the current position, the iteration continues, starting at the best position. Otherwise, the iteration stops with the current position as the result. Usually, our efficient greedy approach stops after 2-3 iterations and shows, however, satisfactory results.

The $2n_t + 1$ positions in the neighborhood arise as follows: first the gradient of c is approximated using a finite forward difference (with step size $\Delta x \coloneqq 8$), which results in $2n_t$ evaluations of the confidence map (instead of $4n_t$ using central difference). Then one step (likewise with step size Δx) in the direction of the gradient is taken for the remaining evaluation.

Measurement of object size. Mainly due to perspective projection the size of a player in the image depends on the image position. To enforce consistency, we normalize the sizes of the predicted bounding boxes before the maximization step. For this purpose, we fit a linear model using a least-squares approach, where the height h is the dependent and the coordinate y is the independent variable. Afterwards, the height of each bounding box is adapted according to this model and the width is determined by $w \coloneqq 0.41h$ (see [8]).

After the maximization step, for each bounding box $\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}$ we iterate several steps of scale with respect to the original size and take the size with the best confidence value c_{best} as a resulting measurement, which results in the final measurement quality $q\left(\widetilde{\boldsymbol{B}}_{\boldsymbol{i}}^{\boldsymbol{t}}\right) \coloneqq c_{best}$.

4. Player tracking

We perform multi-target tracking by applying a single-target Kalman filter [1] for each tracked bounding box. A new measurement is generated by an optimization step with respect to the confidence map using the Kalman prediction as the starting position. The detection of new and lost targets is performed with the help of deterministic heuristics, which are guided by the confidence map.

State and measurement model. The state of player p_i is represented by a six-tuple $o_i \coloneqq (x_i\ y_i\ \dot{x}_i\ \dot{y}_i\ w_i\ h_i)^T$, where x_i, y_i, w_i, and h_i are the parameters of the bounding box, and $\dot{x}_i$ and $\dot{y}_i$ are the velocity in x and y, respectively. We apply a constant size and constant velocity model. Thus, the changes in size and the acceleration of the objects are implicitly modeled by the process noise. Measurements are performed for the position and the size variables.

Loss and detection of players during tracking. During tracking a tracked bounding box $\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{t}}$ is removed if its measurement quality $q(\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{t}})$ is below a threshold t_{loss} in two frames in a row or if it has a significant overlap with another tracked bounding box (with higher measurement quality) in more than four frames in a row.

In return, connected components that fulfill some size constraints according to the linear size model are considered as single player candidates and for these positions the confidence map is evaluated and maximized. Single player candidates that have a measurement quality $q(\boldsymbol{B}_{\boldsymbol{i}}^{\boldsymbol{t}})$ greater than or equal to a threshold t_{detect} in two frames in a row are added to the tracked bounding boxes.

Initial player detection. For the initial player detection, we select connected components of the foreground region $\boldsymbol{F^0}$ with an appropriate orientation and aspect ratio as single player candidates. In the neighborhood of each candidate, we perform a sliding-window human detection with a scale space which depends on the size of the candidate region. The positive detections are used to fit our linear size model. The human detection is repeated in the neighborhood of all connected components of $\boldsymbol{F^0}$ with a restricted scale space depending on the position of the region in the image. Again, the positive detections are used to fit the size model and to estimate the team histograms (see section 2). These detections are the starting point for the first maximization step resulting in the initial player positions.

5. Implementation details and evaluation results

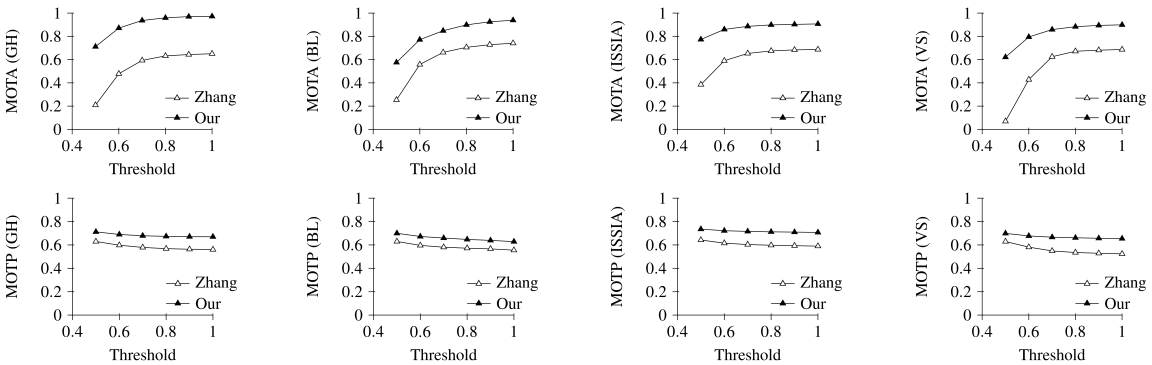

Figure 2: MOTA (top row) and MOTP (bottom row) results (in dependence of the overlap threshold) of the proposed tracking procedure in comparison with the procedure of Zhang et al. [5]. Results are given for each data set (GH, BL, ISSIA, VS) from left to right.

In our implementation we set the number of dominant colors $k_C := 64$ as an estimated upper bound for five outfits with each three four-colored parts (jersey, short, socks). The other parameters are determined empirically and set as follows: $t_{min} := 0.1$, $t_{max} := 0.25$, $u_D := -4$, $v_D := 0.5$, $w_1 := 0.75$, $w_2 := 0.4$, $w_3 := 0.05$, $w_4 := 1.0$, $w_5 := 0.5$, $w_6 := 0.05$, $t_{loss} := 0.35$, and $t_{detect} := 0.7$. We evaluate our system with the help of MOTA and MOTP [13] and use manually annotated ground truth of the following video sequences:

- German – Holland, TV broadcast, SD, 37s, 925 frames (**GH**)
- Bayern Munich – OSC Lille, TV broadcast, HD, 15s, 752 frames (**BL**)
- ISSIA-CNR Camera 3, static camera, Full-HD, 120s, 3000 frames, online available [14] (**ISSIA**)
- VS-PETS Camera 3 Test, static camera, 720x576, 100s, 2500 frames, online available [15] (**VS**)

Ground-truth and tracked bounding boxes are fixed to an aspect ratio 0.41:1 as proposed in [8]. Some annotated targets that are outside the field (like coaches and linesman), as well as ground-truth bounding boxes that are cropped by the image boundaries, are added to an ignore list, i.e. they don't need to be matched, but it is not an error if they are matched.

Our baseline is the publicly available tracker of Zhang et al. [5], which achieves state-of-the-art tracking results on pedestrian tracking datasets. We used the standard parameter values which come with the source code, except the HOG_DETECT_FRAME_RATIO, which we set to 2 for BL and ISSIA, and to 3 for GH and VS allowing to detect small players. As the detector of this tracking procedure generates a lot of false positive detections in the audience area, we ignore all its tracked objects outside our field hull H^t at time t.

Figure 2 shows the comparison of the MOTA / MOTP of our proposed tracking procedure and the procedure of Zhang et al. depending on the overlap threshold for the assignment of the tracked bounding boxes to the ground truth. Our system achieves satisfactory MOTA scores for the standard overlap threshold of 0.5. With increasing overlap threshold, the accuracy becomes remarkable (up to 0.9 and more). In all cases, our results outperform the results of the baseline. Our non-optimized implementation processes about 1-2 frames per second on an Intel Core2 Quad Q9650. Because of its highly parallel structure, we believe that a near real-time performance could be possible. In contrast, the method of Zhang et al. processes 0.1 frames per second.

We provide a publicly available visualization of our tracking results for ISSIA and VS (see [16] and [17]).

6. Conclusion

We proposed an unsupervised online 2D tracking procedure for players in monocular soccer videos that applies an efficient determination of local maxima in a confidence map. This map is based on a robust combination of soccer-specific (grass color), match-specific (team outfit colors) and general (HOG detector) image features. Avoiding a time-consuming sliding-window approach our system allows for a fast player tracking that in addition does not require any further input, such as user input or camera parameters. Our tracking results achieve high accuracy and outperform a state-of-the-art pedestrian tracker.

Acknowledgements

The funding for this research was provided by German Research Foundation (DFG) grant no. RA359/12-1.

References

[1] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems.," *Transactions of the ASME - Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
[2] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.
[3] H. Izadinia, I. Saleemi, W. Li, and M. Shah, "(MP)2T: Multiple People Multiple Parts Tracker," in *Computer Vision – ECCV 2012*, vol. 7577, Berlin, Heidelberg: Springer, 2012, pp. 100–114.
[4] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based Object Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.
[5] J. Zhang, L. L. Presti, and S. Sclaroff, "Online Multi-person Tracking by Tracker Hierarchy," in *IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 379–385.
[6] M. Beetz, S. Gedikli, J. Bandouch, B. Kirchlechner, N. von Hoyningen-Huene, and A. C. P. Perzylo, "Visually Tracking Football Games Based on TV Broadcasts," in *20th International Joint Conference on Artificial Intelligence*, 2007, pp. 2066–2071.
[7] S. Gerke, S. Singh, A. Linnemann, and P. Ndjiki-Nya, "Unsupervised Color Classifier Training for Soccer Player Detection," in *Visual Communications and Image Processing*, 2013, pp. 1–5.
[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[9] M. Hoernig, M. Herrmann, and B. Radig, "Real Time Soccer Field Analysis from Monocular TV Video Data," in *11th International Conference on Pattern Recognition and Image Analysis (PRIA-11-2013)*, Samara, 2013, vol. 2, pp. 567–570.
[10] FIFA, "Laws of the Game," 2014. [Online]. Available: http://www.fifa.com/aboutfifa/footballdevelopment/technicalsupport/refereeing/laws-of-the-game/. [Accessed: 30-Apr-2014].
[11] D. Arthur and S. Vassilvitskii, "k-means++: the Advantages of Careful Seeding," in *18th annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, 2007, pp. 1027–1035.
[12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 2005, vol. 1, pp. 886–893.
[13] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, May 2008.
[14] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P. L. Mazzeo, "A Semi-automatic System for

Ground Truth Generation of Soccer Video Sequences," in *6th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 559–564.
[15] University of Reading, "VS-PETS Football Dataset," *Third IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2002. [Online]. Available: http://www.cvg.reading.ac.uk/VSPETS/vspets-db.html. [Accessed: 30-Apr-2014].
[16] M. Herrmann, M. Hoernig, and B. Radig, "Player Tracking Results ISSIA-CNR," 2014. [Online]. Available: http://www.youtube.com/watch?v=L9t7ei6gAjk. [Accessed: 30-Apr-2014].
[17] M. Herrmann, M. Hoernig, and B. Radig, "Player Tracking Results VS-PETS," 2014. [Online]. Available: http://www.youtube.com/watch?v=SL1LjRAbPgI. [Accessed: 30-Apr-2014].