Real-time Person Detection and Tracking in Panoramic Video

Marcus Thaler, Werner Bailer
JOANNEUM RESEARCH, DIGITAL – Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
{firstname.lastname}@joanneum.at

Abstract

The format agnostic production paradigm has been proposed to offer more engaging live broadcasts to the audience while ensuring the cost-efficiency of the production. An ultra-HD resolution panorama is captured, and streams for different devices and user profiles are semi-automatically generated. Information about person positions and trajectories in the video are important cues for making editing decisions for sports content. In this paper we describe a real-time person detection and tracking system for panoramic video. The approach extends our earlier tracking by detection algorithm by addressing a number of robustness issues that are especially relevant in sports content. The design of the approach is strongly driven by the requirement to process high-resolution video in real-time. We show that we can achieve improvements of the robustness of the algorithm while being able to perform real-time processing.

1. Introduction

Sports is a very important genre for broadcasters, both commercially and terms of audience engagement. In order to keep this position, the broadcast industry is constantly challenged to provide more immersive and interactive user experiences, and realizing this with constant or even shrinking production budgets. The industry can only win this challenge, if the efficiency of the production process can be increased, which requires better tools for automation. This in turn needs metadata for the audiovisual content. Some of this metadata can be captured automatically by specific sensors, but much important information is only in the audiovisual signal. Computer vision has thus an important role for extracting spatiotemporally fine-grained metadata from the visual content, which cannot be annotated manually due to time and cost constraints.

A *format agnostic* approach [10] to broadcast production has been proposed in order to enable a more engaging viewing experience, while avoiding a cost increase. This approach advocates a paradigm shift towards capturing a format agnostic representation of the whole scene from a given viewpoint, rather than the view selected by a cameraman based on assumptions about the viewer's screen size and interests. This system uses the concept of a *layered scene representation*, where several cameras with different spatial resolutions and fields-of-view can be used to represent the view of the scene from a given viewpoint. The views from these cameras provide a base layer panoramic image (obtained in the system by the OmniCam panoramic camera), with enhancement layers from one or more cameras more tightly-framed on key areas of interest. The same concept is used for audio by capturing an ambient sound field together with a number of directional sound sources. The system thus allows end-users to interactively view and navigate around an ultra-high resolution video panorama showing a live event. Sports events (e.g., soccer, track & field) are important target scenarios for this technology. The system includes components called Scripting Engines which take the decisions about what is visible and audible at each playout device and prepare the audiovisual content streams for display. Such components are commonly referred to as a *Virtual Director*. In order to take reasonable decisions, the Scripting Engines need knowledge about what is currently happening in the scene and which camera streams are capturing that action.

The positions and trajectories of persons, e.g. players, are crucial input for the production process, as they are naturally the focus of the viewers' attention in sports. Based on their motion and interactions we get information about where action is going on. This is especially relevant for field sports involving a larger number of persons. Information about person positions in the image is also important in order to observe visual grammar for framing shots and following moving persons. Finally, we also want to enable users to follow specific persons, which they are interested in, and which might not currently be in the view of a camera operator focusing on the main action.

2. Related work

Detection and tracking of persons in field sports is a very popular test scenario for many tracking approaches due to the involving challenges like multiple occlusions, motion blur and uniformly colored jerseys. In [7], tracking of soccer players in a multi-camera environment using color distributions based appearance model and two motion models is discussed. In addition to predicting motion of players in the image plane, a 3D motion model using homographies is established. The 3D model considers players' movements on a common ground plane to handle handover of person IDs and occlusion in the overlapping image areas.

To overcome occlusion problems camera calibration parameters are used in [2] to project detection of persons from different static camera views on a common ground plane. Using these projections, tracking of basketball players is performed on the basis of an established occupancy map and a graph based tracking using player recognition by jersey numbers. The tracking and detection information is then used for an automatic summary of the game captured by an omnidirectional multi-camera environment. In [8], 3D coordinates of soccer players are processed to overcome occlusion on the basis of SIFT features. Matching of SIFT features between multiple views covering the same players from different viewpoints is used to establish the 3D position of each player. Establishing correspondences by feature matching between consecutive frames is then used for tracking.

Detection and tracking of athletes under the presence of motion blur and different body deformations on the basis of a particle filter is presented in [11]. For person detection a vote-based confidence map is used. An appearance model is established for each detected person and optical flow is used for tracking. In [12] and [5], the trajectories of field sport players in sports videos captured by moving broadcast cameras are used to support the analysis of the athletes' movement and behavior. In the first approach the discrimination between soccer players and the playing field and furthermore, the assignment to the appropriate team is derived by color histograms of player regions. Tracking of the players is based on particle filters and to overcome occlusions Haar features are used. In order to track basketball players an adjustable mean shift algorithm based on color histograms is used in [5]. The players' trajectories in the image plane are projected into a real world model representation of the playing field using an automatic estimated homography for further behavior analysis. The homography estimation between the image view of a non-static camera and the model representation of the playing field is based on automatic playing field detection.

In this paper we describe a real-time person detection and tracking system for panoramic video. The approach extends the tracking by detection approach we have proposed in [6] by addressing a number of robustness issues that are especially relevant in sports content (e.g., fast motion, motion blur, occlusions). The design of the approach is strongly driven by the requirement to process this high-resolution data in real-time, thus we propose a number of performance optimizations. Finally, we provide evaluation results for the improved algorithm, using a data set with more detailed annotation than in the previous work.

The rest of this paper is organized as follows. Section 3 summarizes our earlier approach and describes the improvements in detail. In Section 4 we describe the evaluation scores we are using and report the results of the proposed algorithm on a soccer data set. Section 5 concludes the paper.

3. Person detection and tracking

We describe a real-time person detection and tracking system. The approach extends the tracking by detection approach we have proposed in [6] by addressing a number of robustness issues that are especially relevant in sports content (e.g., fast motion, motion blur, occlusions) and optimizes some steps in the process. We first briefly summarize our earlier work. This method performs person tracking in a 180° panorama at a resolution of 6×FullHD and is separated into two main components (see Figure 1). The region tracker is responsible for person detection and tracking in the individual parts called tiles of the panoramic image. The multi-tile tracker is subsequently responsible to solve the handover of assigned person IDs between the different tiles.

For each tile, the algorithm consists of a person detection and a person tracking component. A modified version of the fastHOG [9] person detector is used, implementing HOG [1] on the GPU using CUDA[1]. To achieve real-time performance the scale ratio defining the different sizes of the sliding window applied to calculate the HOG descriptor has been increased, and is locally adapted based on prior knowledge about the scene setup, resulting in a speedup factor of 4. In order to handle missed detections because of low gradients (e.g., motion blur) the OpenCV Blob detector[2] is used. To connect person detections of subsequent frames, a KLT tracker implementation based on CUDA [3] is used. In [6], a further optimization taking advantage of NVIDIA's Fermi[3] architecture is described, resulting in real-time tracking capability of up to 10K feature points in two FullHD streams on a single GPU.

[1] Compute Unified Device Architecture, http://www.nvidia.com/object/cuda_home_new.htm

[2] http://opencv.willowgarage.com/wiki/VideoSurveillance

[3] http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf

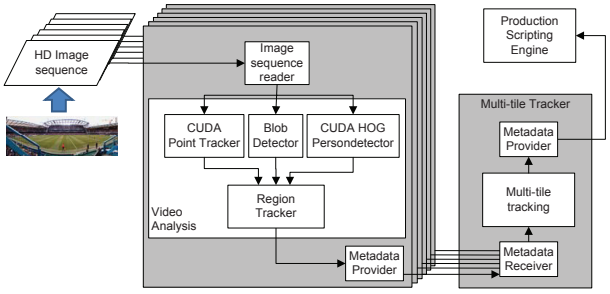

Figure 1. Overview of the tracking by detection algorithm described in [6]

Because of the tile-based processing, the approach can be parallelized in order to achieve overall real-time performance. This can either be done by processing tiles on separate machines or one a single machine, with a multi-core CPU and multiple CUDA enabled GPUs.

In the following, we describe the improvements over previous work. The real-time capability is a key requirement for live sports broadcasting applications. Therefore, several components of the algorithm have been optimized to decrease runtime. To address the requirements of different sports (e.g., degree of dynamics, number of persons involved) we determine two confidence values, a person detection score and a person tracking score. Depending on the particular scenario the algorithm can be parameterized either provide all detected persons and trajectories or provide only stable tracks with high confidence scores (e.g., to have a virtual camera follow a player). In order to deal with highly dynamic sports with fast motion, improvements of the use of the blob detector have been performed. In cases of fast and sudden motion, especially under non-optimal lighting conditions, motion blur will occur. The resulting decrease of gradients poses a problem for both HOG and KLT, and must be compensated for. To increase the long-term stability of tracks motion models from previous tracking results are established. In order to reconnect tracks split by occlusion, motion from previous tracking results is predicted and verification by color features is introduced. Furthermore, prior knowledge is used to enhance the performance of the approach. For instance, the known playing field geometry is used to define the region of interest in form of image masks. Additionally, our approach takes advantage of known number of maximum involved athletes, especially for cases where the assignment of person IDs is ambiguous due to occlusion. The details of these improvements for the detection and tracking components are detailed in the following sections.

3.1. Person detection

Using a blob detector that complements gradient-based methods is very important for our approach because of motion blur caused by rapid movements due to rushing or diving to the ball and rapid turns of players (e.g., in basketball). In the previous implementation, blob appearances in consecutive frames where verified in order to remove false detections. Instead, verification is now only based on the scene scale, resulting in a runtime performance improvement of 25%. The detections from HOG and the person detector are merged by fusing those with at least 50% spatial overlap. To verify the joint person regions a scene scale derived from the players' size is used. Due to the use of a static panorama camera persons at the same distance to the camera centre are represented by approximately the same height (in pixels). The joint results of both detectors provide the regions of the detected athletes each represented by a bounding box (see Figure 2).

We introduce a confidence value called person detection score (pds), determined for a detection in one frame as

$$pds = \alpha_d \left(\beta \frac{w_{BB}}{h_{BB}} + (1-\beta) s_h \right) + \alpha_p \frac{|P \cap BB|}{h_{BB}}, \quad (1)$$

where β is 1, if the detection is a blob detection, w_{BB} and h_{BB} are the width and height of the bounding box BB of the detection, s_h is the detection score from the HOG detector, and P is the set of tracked points in the frame (thus $|P \cap BB|$ are the number of the points inside the detected bounding box). α are the respective weights, with $\alpha_d = 1.2$ and $\alpha_p = 1$ in the experiments.

3.2. Person tracking

The main enhancements are improvements of the region tracking strategy by adding a color histogram comparison and a person tracking score. The verification of color features is used to solve occlusion handling and is also a key component of the person tracking score. For each detected person region a color histogram (RGB, 30 bins) is calculated and updated for each frame in the image sequence and afterwards compared to the color histogram of the corresponding person tracks from the previous frame.

The person tracking score (pts) is introduced as a measure for the long-term stability of the tracks. The score is defined for a recent time window of k frames of the trajectory as

$$pts = \frac{1}{k} \sum_{i=1}^{k} pts(i), \quad (2)$$

with $k = 10$ in the experiments. The pts of one frame is defined as

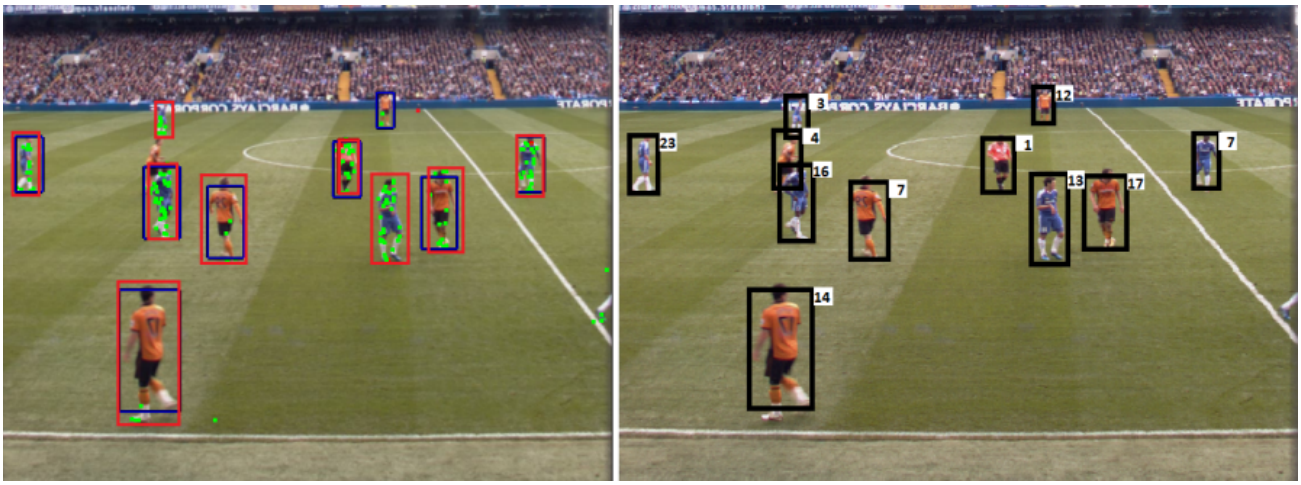

Figure 2. The left image shows the person detections from the HOG (blue Bounding boxes) and the Blob detector (red Bounding boxes) with the clustered feature points inside. The derived person tracking results are depicted (by Bounding boxes and the corresponding person IDs) on the right.

$$
\begin{aligned}
pts(i) = \quad & \alpha_c D_B(H^i, H^{i-1}) + \\
& \alpha_s \left(\tfrac{1}{w} |w^i_{BB} - w^{i-1}_{BB}| + \tfrac{1}{h} |h^i_{BB} - h^{i-1}_{BB}| \right) + \\
& \alpha_p \left(\frac{|P^i \cap BB^i|}{h^i_{BB}} - \frac{|P^{i-1} \cap BB^{i-1}|}{h^{i-1}_{BB}} \right) + \\
& \alpha_m + \tfrac{1}{|P^i|} \left(\sum_{j=1}^{|P^i|} ||P^i_j - P^{i-1}_j|| \right) + \\
& \alpha_d pds(i) + \tfrac{\alpha_a}{k} A(i),
\end{aligned}
\tag{3}
$$

where the superscript indices denote the frame number, D_B is the Bhattacharyya distance between the two color histograms H, w and h denote the width and height of the tile (with index BB of the bounding box), P is a set of tracked points and P_j a point coordinate vector. $pds(i)$ is the person detection score for the respective frame of the track, and $A(i)$ is the age of the track up to the current frame. α are the respective weights, set to equal values in our experiments.

The enhanced tracking strategy is described in the following. In the initial process of our approach, each newly detected feature point is related to the person region with minimum distance to its center. For each sufficiently large feature point cluster that is not among the previous tracking results a new person ID is assigned to the overlapping person region of the track. If a detected region contains feature points not yet linked to the same person ID or if at least two detected regions contain feature points with the same person ID, the assignment of the person ID to a person region is verified by the tracking score (with $k = 1$). On the basis of this verification the detected region and the feature points inside are (re-)linked to the person ID of the track having the best score. Due to the improved handling of ambiguous situations by assignment of person IDs to the appropriate person region the algorithm's performance is enhanced significantly by reducing the number of ID switches.

In order to handle cases where person detection fails inside a region already being tracked, propagation of previously detected person regions is introduced. The region's position is updated by the new overall location of the corresponding feature points on the basis of a color histogram comparison. In case of an insufficient number of feature points, the new position is calculated from the region's previous direction of movement and velocity depending on a verification of the color features and the age of the track. For occlusion handling the color histogram of each person region is updated over time while it is not occluded. In addition to the appearance model a motion model predicting motion from previous tracking is used.

4. Evaluation

For quantitative evaluation of the person detection and tracking results a soccer game data set has been used. Different test sequences (about 2,000 frames each) are used for evaluation. Compared to the evaluation in [6], we manually annotated ground truth tracks in addition to person bounding boxed in order to evaluate ID switches.

For person detection evaluation, we compared the bounding boxes of the ground truth data with the bounding boxes of person regions obtained from our approach. All visible soccer players are annotated in the center tiles of the sequences. Using a bounding box overlap threshold of 25% we obtain an average recall of 0.89 and average precision of 0.91 for all test sequences. Table 1 contains the results for the improved algorithm (a) and the previous version (b). The increase of the recall by 0.30 compared to earlier results is a consequent of two main enhancements introduced by the modified approach, namely improvements of the blob detection and propagation of previously detected

persons. To overcome missed detections the latter method uses prediction of motion from previous tracking results especially in cases where point tracking failed, e.g., due to motion blur. Both enhancements increase the number of person detections significantly.

For person tracking evaluation, we manually annotated ground truth tracks of the center tiles taken from two test sequences (see Figure 2). On the basis of different metrics for evaluation of continuous and stable tracks [4] we compared the ground truth tracks $T = (T_1, \ldots, T_n)$ with the tracks provided by the proposed algorithm $\tilde{T} = (\tilde{T}_1, \ldots, \tilde{T}_m)$.

For comparison of tracks the spatial and temporal overlap are taken into account. The measure for spatial overlap is defined by the person region representing T_i and the corresponding region of $\tilde{T}_j$ for one frame. The overlap is defined by the intersection divided by the union of both regions. The temporal overlap measure between T_i and $\tilde{T}_j$ is defined by the number of frames spatially overlapping between the two tracks, normalized by the number of frames in which at least one of them is visible. Using a spatial overlap threshold of 20%, the temporal coverage over annotated tracks is 0.92. The result is a consequence of the enhanced person detection under the presence of motion blur in combination with the propagation of the previously detected person regions. Consequently, the coverage approximately matches with the above mentioned precision and recall for person detection (with a spatial overlap threshold of 25%). To determine a correctly detected track [4] or true positive we use a threshold of 60% for the temporal overlap in addition to a spatial overlap threshold of 20%. As a result we obtain an average recall of 0.71 (each result track can be assigned to only one ground truth track). By omitting the temporal overlap threshold of 60% the average Track Completeness (TC) [4] can be estimated. For each T_i the $\tilde{T}_j$ with the maximum common overlap (temporal with a spatial overlap of at least 20%) is assigned. Processing the two test sequences we obtain an average TC of 0.73.

The ground truth contains 41 tracks with an average length of 353 frames, and the average overlap with detected tracks is 324 frames. An ID Change (IDC) [4] occurs if an algorithm track is assigned to more than one ground truth track. Furthermore, if two algorithm tracks change their corresponding ground truth tracks (e.g. due to failed occlusion handling) we are speaking of an ID swap (= 2 IDC). Assuming that each IDC is caused by one ID swap we estimated an ID change rate of approximately one ID swap per minute (1 every 1,500 frames) for the used test sequences. Thus, we obtain an average precision of 0.67 by handling all IDC as false positives. IDCs are mainly caused by situations where more than two persons overlap each other at the same time, usually under presence of fast motion (e.g., a group of players rushing to the ball). If a continuous ground truth track is covered by a set of different algorithm tracks, the set is not considered for calculation of the recall if each $\tilde{T}_j$ is assigned to one T_i only. This fragmentation of algorithm tracks is a consequence of the difficult discrimination between the upper bodies of the players at the far end of the soccer field because of the moving crowd and the moving advertising board behind. We can solve the disconnection by further trajectory analysis, e.g., on the basis of to the known number of maximum involved players.

By excluding different building blocks in the evaluation process, their respective influence has been analyzed (see Table 1 (c)-(e)). For instance, by omitting color verification the tracking recall decreases drastically (c). The track continuity is reduced due to the lack of support for occlusion handling, track propagation and assignment of person IDs. The latter influences the decrease of the tracking precision and thus increases number of ID changes as well. The influence of the blob detector is shown by the reduction of track coverage (d). Due to the higher number of missed detections the track length and thus the tracking recall are reduced. The results in the last column (e) underline the importance of the propagation of previous tracks in case of missed detection or due to occlusion.

In order to assess the influence of the scores proposed above we used a tracking score (with $k = 10$) threshold of 75% for evaluation. This increases average person detection precision significantly to 0.98 and decreases the ID changes along the tracks by 0.15, due to the selection of stable person detections and tracks. However, as a consequence of discarding tracks with low confidence value the average person detection recall decreases by 0.19 and the tracking recall by 0.10. Due to improvements of the use of the blob detector and the feature point tracker (using approximately 1,500 points has been found to be sufficient) each of these components has been sped up by about 25%. The total runtime for each image tile is about 85ms, consisting of 70ms for the improved fast HOG, about 10ms for the feature point tracker, 35ms for the blob detector. Analyzing a FullHD image sequence the runtime of these three components running in parallel is constant for different scenarios. The runtime for combining the results varies with the number of involved players. In our experiments we evaluated a maximum runtime of 15ms for tracking 17 athletes in one image tile, considered as close to the upper limit of involved players regarding a field sports scenario. Analyzing every second frame in the image sequence of a high resolution real-time broadcast camera is sufficient to provide reliable tracking in most scenarios. Therefore, the algorithm performance is suitable to support real-time broadcast systems.

5. Conclusion

We have proposed extensions and improvements to a person detection and tracking approach, adapting it better to the requirements of sports broadcasts. The proposed approach

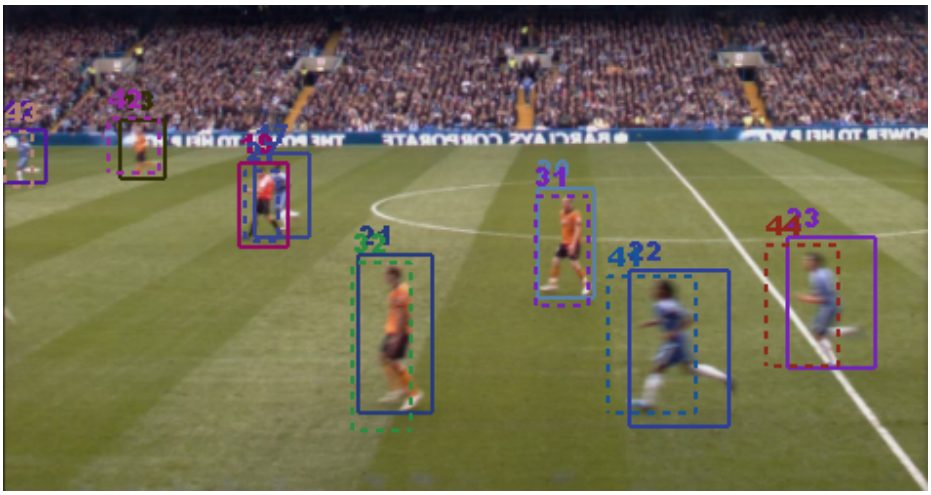

Figure 3. Annotated ground truth tracks and tracks detected by the proposed approach (represented by the dashed bounding boxes).

Measure	(a)	(b)	(c)	(d)	(e)
Recall (detection)	0.89	0.59	-	-	-
Precision (detection)	0.91	0.96	-	-	-
Recall (tracking)	0.71	-	0.41	0.51	0.10
Precision (tracking)	0.67	-	0.55	0.62	0.24
Track Completeness	0.73	-	-	-	-
Track coverage	0.92	-	0.84	0.78	0.70
IDC per min.	2.0	-	3.7	2.8	15.0

Table 1. Person detection and tracking evaluation (a) , and comparison with the results: from [6] (b), excluding color verification (c), excluding blob detection (d), excluding propagation of previous tracks (e) where applicable.

is able to perform real-time detection and tracking of persons in ultra-HD resolution panoramic video. We discussed improvements for cases in which gradients are temporally low (e.g., due to motion blur) and for handling occlusion issues. The approach provides sufficiently reliable results as input to a subsequent virtual director component, which selects streams for delivery to different target platforms.

Acknowledgement

This work has been partially supported by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248138 FascinatE (http://www.fascinate-project.eu).

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, 2005. 2

[2] D. Delannay, N. Danhier, and C. D. Vleeschouwer. Detection and recognition of sports(wo)men from multiple views. In *Third ACM/IEEE International Conference on Distributed Smart Cameras*, 2009. 2

[3] H. Fassold, J. Rosner, P. Schallauer, and W. Bailer. Real-time KLT Feature Point Tracking for High Definition Video. In V. Skala and D. Hildebrand, editors, *GraVisMa 2009 - Computer Graphics, Vision and Mathematics for Scientific Computing*, 2010. 2

[4] S. A. V. Fei Yin, D. Makris. Performance evaluation of object tracking algorithms. IEEE Intl. Workshop on Performance Evaluation of Tracking and Surveillance, 2007. 5

[5] M.-C. Hu, M.-H. Chang, J.-L. Wu, and L. Chi. Robust camera calibration and player tracking in broadcast basketball video. *IEEE Trans. Multimedia*, 13(2):266–279, 2011. 2

[6] R. Kaiser, M. Thaler, A. Kriechbaum, H. Fassold, W. Bailer, and J. Rosner. Real-time person tracking in high-resolution panoramic video for automated broadcast production. In *Conf. Visual Media Production*, 2011. 2, 3, 4, 6

[7] J. Kang, I. Cohen, and G. Medioni. Soccer player tracking across uncalibrated camera streams. In *IEEE Intl. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS) In Conjunction with ICCV*, pages 172–179, 2003. 2

[8] H. Li and M. Flierl. Sift-based multi-view cooperative tracking for soccer video. In *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, 2012. 2

[9] V. Prisacariu and I. Reid. FastHOG - a real-time GPU implementation of HOG. Technical report, Department of Engineering Science, Oxford University, 2009. 2

[10] R. Schäfer, P. Kauff, and C. Weissig. Ultra high resolution video production and display as basis of a format agnostic production system. In *Proceedings of International Broadcast Conference (IBC 2010)*, 2010. 1

[11] A. Yao, D. Uebersax, J. Gail, and L. V. Gool. Tracking people in brodcast sports. In *Proceedings of the 32nd DAGM conference on Pattern recognition*, 2010. 2

[12] G. Zhu, C. Xu, Q. Huang, Y. Rui, S. Jiang, W. Gao, and H. Yao. Event tactic analysis based on broadcast sports video. *IEEE Trans. Multimedia*, 11(1):49–67, 2009. 2