Evaluation of Probabilistic Occupancy Map for Player Tracking in Team Sports

Matej ŠMÍD, Jan POKORNÝ

Dept. of Cybernetics, Czech Technical University, Technická 2, 166 27 Praha, Czech Republic

smidm@cmp.felk.cvut.cz, pokorj@gmail.com

Abstract. *In this paper we focus on tracking of players in team sports with multiple cameras. We evaluate a state of the art multi-camera multi-target tracking algorithm on a novel floorball dataset. We chose recent Probabilistic Occupancy Map and K-Shortest Path algorithms with a publicly available implementation that enables easy deployment. Both algorithms are mathematically well funded and capable of near real-time performance. We show quantitative evaluation of the algorithms on a real-world problem, and demonstrate influence of different background subtraction methods that are a necessary preprocessing step.*

Keywords

Visual tracking, multi-camera, multi-target, team sports, evaluation.

1. Introduction

Visual tracking in sports is a promising field for applied computer vision. The possible uses range from real-time game analysis (such as referee assistance), video augmentation for presentation purposes, automatic video summarization, to game analysis for players training.

In this paper we focus on tracking players positions in indoor team sports using multiple synchronized cameras.

We evaluate the state of the art Probabilistic Occupancy Map (POM) [5][2] and K-Shortest Paths optimization (KSP) [1] algorithms. The both algorithms are frequently cited, published recently, and the authors are actively researching in the area. The authors have also published an implementation of the both algorithms.

In Section 2 we explain briefly basic principles of the POM and the KSP algorithms. Also we include description of the used background subtraction algorithms whose results are needed for the POM. Section 3 introduces the used floorball dataset and the associated ground truth for the evaluation. In Section 4 we evaluated the results of POM and KSP using standard CLEAR metrics. We discuss the influence of the background subtraction algorithms and several POM and KSP parameters on the overall performance.

2. Algorithms

A set of algorithms to estimate tracks of players in a pitch observed by multiple cameras is presented in this section. The POM algorithm estimates probabilities of players occupying particular positions on the ground plane in a specified time frame. On the input of the POM are binary images representing moving objects that were generated by a background subtraction algorithm. The K-Shortest Paths finds the most probable tracks of individuals in a sequence of occupancy maps.

2.1. Probabilistic Occupancy Map

We define a grid on the ground plane spanning all admissible player locations G, see Figure 1. The goal is to estimate probability $P(\mathbf{X}|\mathbf{B})$, where $\mathbf{X} = (X^1, \ldots, X^G)$ is a vector of boolean random variables standing for the occupancy of all locations on the ground plane and $\mathbf{B} = (B^1, \ldots, B^C)$ are binary images representing foreground objects in C cameras, see Figure 5.

We have to state two independence assumptions. First is that individuals do not take into account other individuals when moving around

$$P(X^1, \ldots, X^G) = \prod_k P(X^k). \tag{1}$$

Second independence assumption is that all statistical dependencies between views are caused by moving individuals. This can be stated as

$$P(B^1, \ldots, B^C|\mathbf{X}) = \prod_c P(B^c|\mathbf{X}). \tag{2}$$

For following steps we need to define a generative image model $P(\mathbf{B}|\mathbf{X})$ first. We introduce a synthetic image $A^c(\mathbf{X})$ that is generated by placing filled rectangles of the human height on locations where $X^k = 1$. This can be done for

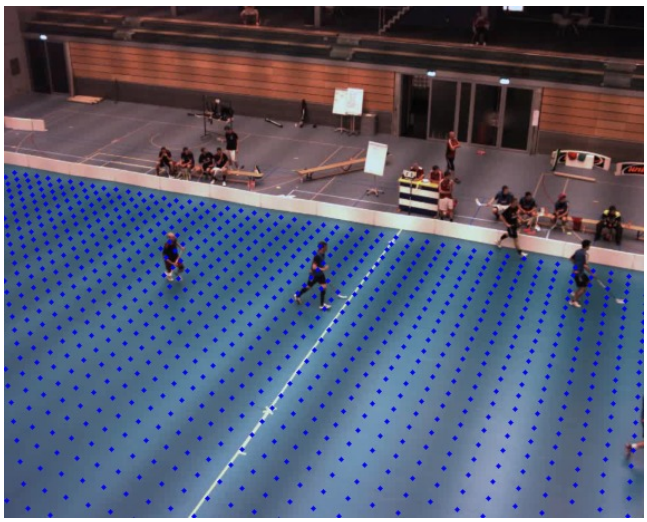

Fig. 1. The ground plane with the grid of possible players locations. Observe that the grid covers only the pitch region. A grid location is represented as a number $1, \ldots, G$.

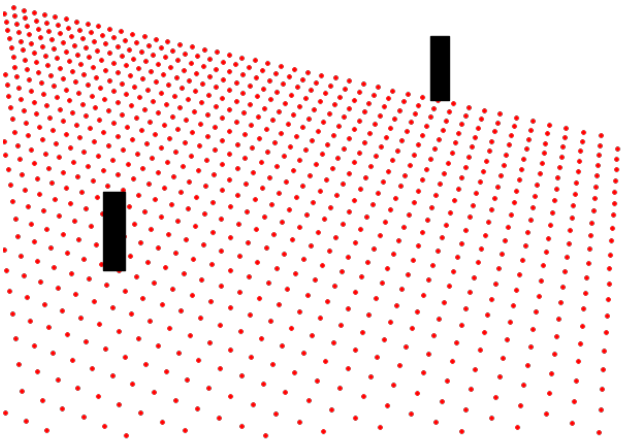

Fig. 2. Synthetic image A_c with $X_{10} = 1$ and $X_{750} = 1$. The red grid serves only for visualization purposes, it is not part of the synthetic image.

all camera views c. An example is in Figure 2. We also introduce a distance function Ψ that quantifies a difference between two grayscale images, for details see [5]. We can state the generative image model as

$$\begin{aligned} P(\mathbf{B}|\mathbf{X}) &= \prod_c P(B^c|\mathbf{X}), \\ &= \prod_c P(B^c|X^c), \\ &= \frac{1}{Z}\prod_c e^{-\Psi(B^c,A^c)}. \end{aligned} \tag{3}$$

Now we can proceed to the final step of the occupancy map derivation. We will approximate conditional occupancy probability $P(\mathbf{X}|\mathbf{B})$ with a product law $Q(\mathbf{X}) = \prod_k Q(X^k)$. Let the marginal probabilities of Q be $q_1, \ldots, q_G$. These directly approximate the marginal occupancy probabilities $P(X^1 = 1|\mathbf{B}), \ldots, P(X^G = 1|\mathbf{B})$, in other words the occupancy map itself. We want to minimize the Kullback-Leibler divergence between Q and the "true" posterior $P(\cdot|\mathbf{B})$

$$\mathrm{KL}(Q, P(\cdot|\mathbf{B})). \tag{4}$$

By plugging the generative model in, minimizing the above divergence and further taking an approximation we get an update rule

$$\begin{aligned} q_k = 1/(1 + \exp(\lambda_k + \sum_c \Psi(B_c, E_Q(A_c|X_k = 1)) \\ - \Psi(B_c, E_Q(A_c|X_k = 0)))), \end{aligned} \tag{5}$$

where λ_k is a constant, Ψ is a normalized pseudodistance between binary foreground image and augmented synthetic image, $E_Q(A_c|X_k = 1)$ and $E_Q(A_c|X_k = 0)$ are expectations of synthetic image over the probability distribution Q with the rectangle corresponding to X_k forced to 1 or 0. Again, for details consult [5]. The expression under the sum represents comparison of the given binary foreground image B_c with a synthetic image A_c augmented to include (first term) and exclude (second term) player on the position k. If the player is actually present on the location k, the value q_k increases.

We find the Q closest to the "true" $P(\cdot|\mathbf{B})$ by iteratively evaluating update rule q_k in Eq. 5. See Figure 3.

2.2. K-Shortest Paths

We present here an algorithm that has a sequence of probabilistic occupancy maps on the input and produces trajectories of players over the sequence [1]. First we construct a graph with directed edges and KT vertices, where K is a number of locations on the ground grid and T is a number of time steps. Vertices represent locations in time and edges represent all admissible player motions. Edges between time steps connect only locations that a player is able to reach in one time step. We introduce virtual vertices v_{source} and v_{sink} that link to locations where a player can enter or leave the scene, e.g. doors, area borders. v_{source} and v_{sink} also link to all nodes in the first and the last time step. This allows to handle players that enter or leave the scene in the same manner as players that are present in the whole sequence. A sample graph can be seen in Figure 4.

To every vertex we assign estimate of the marginal posterior probability of a player presence

$$\rho_i^t = \hat{P}(M_i^t = 1|\mathbf{B}^t), \tag{6}$$

where $\mathbf{B}^t$ is set of images of foreground objects in time t and M_i^t is a random variable standing for presence of a player on a location i in time t. This estimate is directly the output of

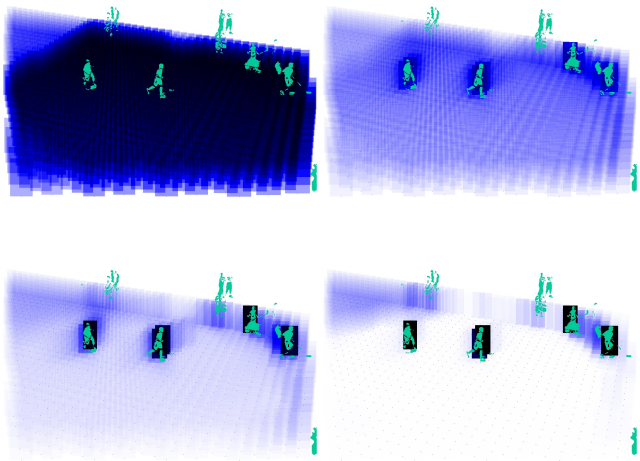

Fig. 3. Iteratively estimating q_ks. Every q_k corresponds to a rectangle on a location k. From left to right, from top to bottom: 1st, 11th, 24th, 100th iteration. Starting from an uniform prior, probability quickly peaks in the true locations.

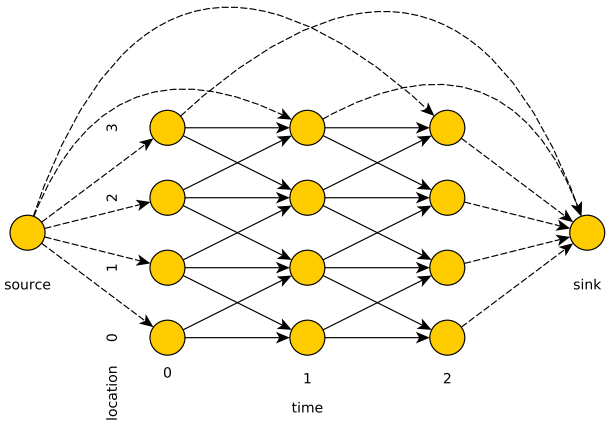

Fig. 4. Simple example of a graph representation for the K-Shortest Paths algorithm. Full edges represent admissible player motions. Dashed edges represent entering or leaving the scene and initial and final player positions.

the POM. Further we assign a cost to all edges not connecting v_{source} or v_{sink}

$$c(e^t_{i,j}) = -log\left(\frac{\rho^t_i}{1-\rho^t_i}\right). \tag{7}$$

The edges connecting source or sink have the cost set to zero. The flow in the graph $f^t_{i,j}$ represents estimated number of objects moving from location i to location j in time t. We want to find

$$\mathbf{f}^* = \arg\min_{f\in\mathfrak{H}} \sum_{t,i} c(e^t_{i,j}) \sum_{j\in\mathcal{N}(i)} f^t_{i,j}, \tag{8}$$

where $\mathfrak{H}$ is a set of feasible flows in the graph. This minimization problem can be solved in near linear time by the k-shortest disjoint paths algorithm [8]. With flows $\mathbf{f}^*$ we can reconstruct the players tracks by backtracking from the sink vertex.

This version of KSP algorithm completely ignores appearance or previous motion of the players and thus performs weak at preserving player identities over the sequence.

2.3. Background Subtraction

On the input of the POM are binary images representing moving objects in the foreground. We experimented with three different background subtraction algorithms: simple thresholded background subtraction, moving median and adaptive background mixture model.

First we prepared background images without foreground objects by taking median of frames in parts of the sequence when all foreground objects are moving.

The simple thresholded background subtraction works as follows: we convert an image and a reference background to the HSV colour space, take the absolute value of the hue channels difference and threshold the difference.

The moving median algorithm is similar. We take the moving median as the background model, convert the background and the current frame to greyscale, subtract the greyscale background from the current frame and threshold the absolute value of the result.

Under the Improved Adaptive Background Mixture Model [7][4] is each pixel modeled as a mixture of K normal distributions. Every distribution represents a colour. A subset of B out of K distributions model the background. An evaluated pixel is considered as foreground when its value is more than 2.5 standard deviations away from any of the B distributions. The pixels not belonging to the foreground update the background model using the online EM algorithm.

Results of the background subtraction algorithms are demonstrated in Figure 5.

Fig. 5. The foreground object masks created by the three background subtraction algorithms. From top to bottom: simple thresholded difference in the hue channel, thresholded difference from a moving median, adaptive background model proposed by KaewTraKulPong.

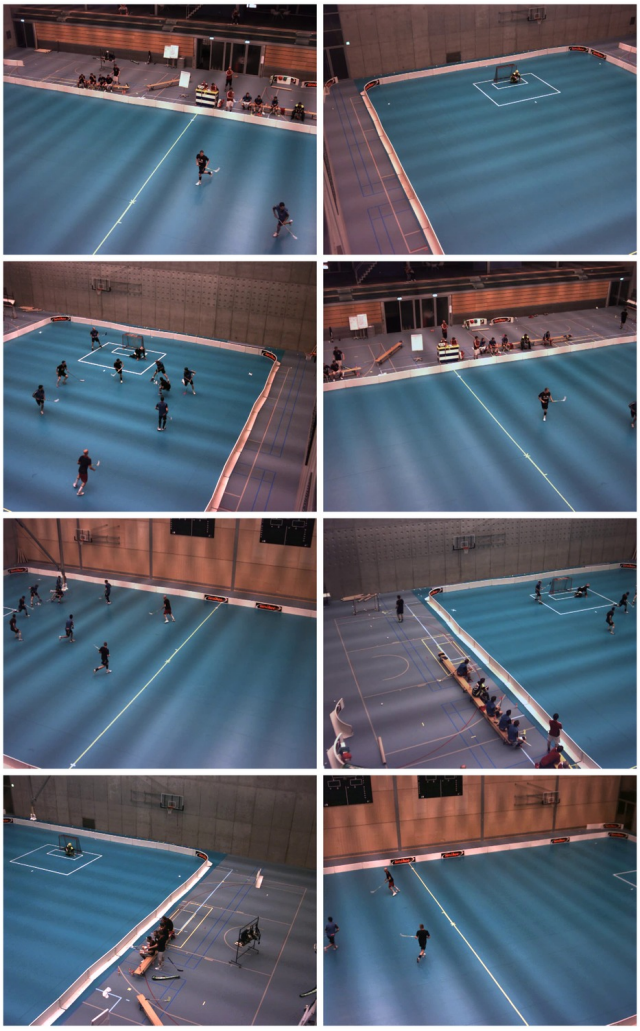

Fig. 6. Snapshot from the floorball dataset. The sequence is 7 minutes long, taken by 8 synchronized cameras.

3. Dataset

We evaluated the tracking algorithms on an indoor floorball video sequence, see Figure 6. It was acquired in a sports hall using 8 synchronized cameras positioned under the ceiling. Most positions on the field are visible in 2 to 4 views. There are 12 players of 2 teams including 2 goal keepers present on the field. The frame rate is 20 frames per second, resolution is 960×768 pixels. For this evaluation we used the first 21 seconds until the first players switch, this duration corresponds to 420 frames.

The camera calibration was done using the Tsai's method [9]. The radial distortion was not compensated, but the error is negligible.

The players of the blue team wear dark blue jersey, the players of the red team wear black jersey with red shoulder part. The jerseys are without numbers. The players wear shorts of various colours.

We marked all players in all camera views every 2 seconds or 40th frame until the first player switch in the 21st

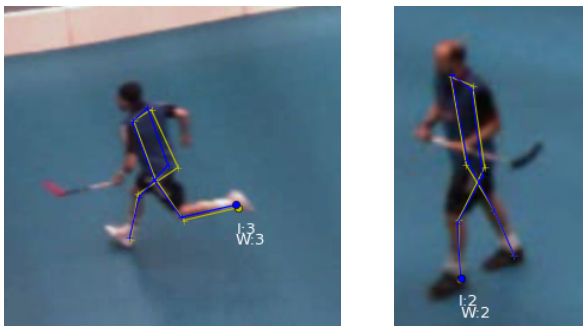

Fig. 7. The evaluation ground truth: a player is described by 8 points. The manually defined 2D ground truth is displayed in yellow colour. The 3D ground truth was reconstructed by the linear triangulation from multiple views. The reprojected 3D ground truth is displayed in blue colour.

second. We annotated the players with more points than necessary for tracking evaluation because of our future plans that include articulated player tracking. A player is marked with 8 points: feet, knees, both hip sides and shoulders. We reconstructed player 3D positions with the linear triangulation method [6, p. 312] and verified correctness by reprojecting 3D positions back to the camera views. The reprojection error is measured as a distance between a manually defined 2D position and the reprojected 2D position.

4. Experimental Results

We performed several experiments with the POM and KSP algorithms on the floorball dataset. We have tested three different background subtraction algorithms and measured the influence of several parameters on the performance. For the quantitative evaluation were used the CLEAR metrics, widely used benchmarks for multi-target tracking [3].

The floorball pitch has dimensions of $40 \times 20\,\mathrm{m}$. We laid on the pitch a 40×79 grid with cell centers spaced in $0.5\,\mathrm{m}$ intervals, see Figure 1. In total the pitch has 3160 locations. With the help of the camera calibration we projected human sized bounding boxes on all positions on the grid in all camera views.

The ground truth for the floorball dataset defines multiple points per player. For this evaluation we used as a player reference point the mean coordinate of the ground truth points shifted to the ground. CLEAR evaluation has matching distance threshold, which we fixed to 1 meter. The detections with a distance to the nearest ground truth greater than the threshold are considered false positives. CLEAR defines the multiple object tracking precision

$$\mathrm{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}, \tag{9}$$

where c_t is the number of matches between detections and the ground truth in time t and d_t^i is distance between detection and ground truth. MOTP is the average matching error made. Second metric is the multiple object tracking accuracy

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN_t} + \mathrm{FP_t} + \mathrm{mme_t})}{\sum_t g_t}, \tag{10}$$

where $\mathrm{FN_t}$ is the number of false negatives in time t, $\mathrm{FP_t}$ is the number of false positives, $\mathrm{mme_t}$ is the number of mismatches and g_t is the real number of objects present at the time t. Mismatch occurs when a player identity is changed. We introduced our own third metric MOTA2 that omits the $\mathrm{mme_t}$ term from MOTA.

We experimented with the following factors. The three used background subtraction algorithms were already discussed in Section 2.3. To compensate the larger player silhouettes caused by a floorball stick often held further from a body we varied the human dimensions parameter. We also experimented with the maximum number of allowed trajectories. The results are shown in Table 1 and a sample evaluated frame is presented in Figure 8.

Out of the three background subtraction algorithms was the best performing the moving median algorithm. The simple subtraction performed poorly due to the large amount of noise from the background that was included into the foreground, see Figure 5. Increasing the human dimensions parameter was not helpful and resulted into more false negatives. From the more detailed inspection of the detected tracks in the sequence we can see that the most challenging players are the goal keepers because of their slow and restricted movement. The all background subtraction algorithms partially fail on them and add them into the background. We have evaluated the experiments on the ground truth with the goal keepers included and also without them. Examples of a missed and a detected goal keeper can be seen in Figure 8. Note the relative high number of mismatches in all meaningful results. As was mentioned at the end of Section 2.2, the algorithm ignores appearance and thus is not able to distinguish between different identities when more players meet closely. Preserving identity of players was not expected. For the evaluation without players identity, we include alternative MOTA2 benchmark that ignores the mismatch errors.

5. Conclusion

We evaluated the two state of the art algorithms with publicly available implementations on the real world dataset. We have created the extensive ground truth for the floorball dataset and performed the quantitative evaluation of the algorithms. The results demonstrate that the evaluated algorithms are sufficiently general to deliver good results on a typical indoor team sport dataset. The weak points were caused by the background subtraction in the special case of minimally moving goal keepers. Also we cannot expect working identity preservation without an appearance model on sequences where people come close to each other.

background subtraction	human dimensions (m)	max. trajectories	goal-keepers in GT	MOTA	MOTP (m)	MOTA2	FN	FP	mismatches
median	0.9 x 1.85	10	no	0.66	0.29	0.82	10	10	17
median	0.9 x 1.85	30	no	0.66	0.29	0.76	2	24	11
median	1.3 x 1.85	10	no	0.60	0.35	0.73	15	15	14
median	1.3 x 1.85	30	no	0.60	0.35	0.73	15	15	14
adaptive	0.9 x 1.85	10	no	-0.34	0.33	-0.33	95	51	1
adaptive	0.9 x 1.85	30	no	-0.34	0.33	-0.33	95	51	1
simple	0.9 x 1.85	10	no	-0.41	0.22	-0.41	105	50	0
simple	0.9 x 1.85	30	no	-0.41	0.22	-0.41	105	50	0
median	0.9 x 1.85	12	yes	0.73	0.29	0.82	12	12	12
median	0.9 x 1.85	30	yes	0.73	0.29	0.82	12	12	12
median	1.3 x 1.85	12	yes	-0.50	0.36	-0.48	109	87	2
median	1.3 x 1.85	30	yes	-0.50	0.36	-0.48	109	87	2
adaptive	0.9 x 1.85	12	yes	-0.17	0.29	-0.17	110	44	0
adaptive	0.9 x 1.85	30	yes	-0.17	0.29	-0.17	110	44	0
simple	0.9 x 1.85	12	yes	-0.33	0.33	-0.33	126	49	0
simple	0.9 x 1.85	30	yes	-0.33	0.33	-0.33	126	49	0

Tab. 1. The results of the experiments. In the left part are the experiment parameters, in the right part are the results. MOTA measures amount of false positives, false negatives and mismatches: bigger value is better. MOTP is mean error in meters. MOTA2 is the same measure as MOTA without counting mismatches in. FN are false negatives or missed detections. FP are false positives. Red values are the best in column.

In future work we would like to include a robust single player appearance model suitable for team sports.

Acknowledgments

Matšj Šmíd would like to thank his supervisor prof. Ing. Jiří Matas, Ph.D.

The authors would like to thank Software Competence Center GmbH and it's partners for floorball sequence acquisition, calibration and for creating 2D ground truth.

Matěj Šmíd was supported in part by the V3C – Visual Computing Competence Center under Technology Agency of the Czech Republic and in part by CTU Student Grant Competition, grant no. SGS13/142/OHK3/2T/13.

References

[1] BERCLAZ, J., FLEURET, F., TURETKEN, E., AND FUA, P. Multiple Object Tracking using K-Shortest Paths Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 9 (2011), 1806–1819.

[2] BERCLAZ, J., SHAHROKNI, A., FLEURET, F., FERRYMAN, J., AND FUA, P. Evaluation of probabilistic occupancy map people detection for surveillance systems. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance* (2009).

[3] BERNARDIN, K., AND STIEFELHAGEN, R. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing 2008* (2008), 1–10.

[4] BRADSKI, G. The opencv library. *Doctor Dobbs Journal* (2000).

[5] FLEURET, F., BERCLAZ, J., LENGAGNE, R., AND FUA, P. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 2 (Feb. 2008), 267–82.

[6] HARTLEY, R., AND ZISSERMAN, A. *Multiple view geometry in computer vision*, second ed. Cambridge University Press, Cambridge, 2003.

[7] KAEWTRAKULPONG, P., AND BOWDEN, R. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-Based Surveillance Systems*. 2002.

[8] SUURBALLE, J. Disjoint paths in a network. *Networks* (1974).

[9] TSAI, R. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *Robotics and Automation, IEEE Journal of* (1987).

Authors

Matěj ŠMÍD – currently PhD student at the Center for Machine Perception, Department of Cybernetics, FEE, CTU interested in sports tracking and articulated objects tracking. After finishing Master studies at CTU, Matěj spent more than 5 years working in the field of computer vision at Software Compentence Center Hagenberg GmbH. He also taught computer vision at University of Applied Sciences Upper Austria.

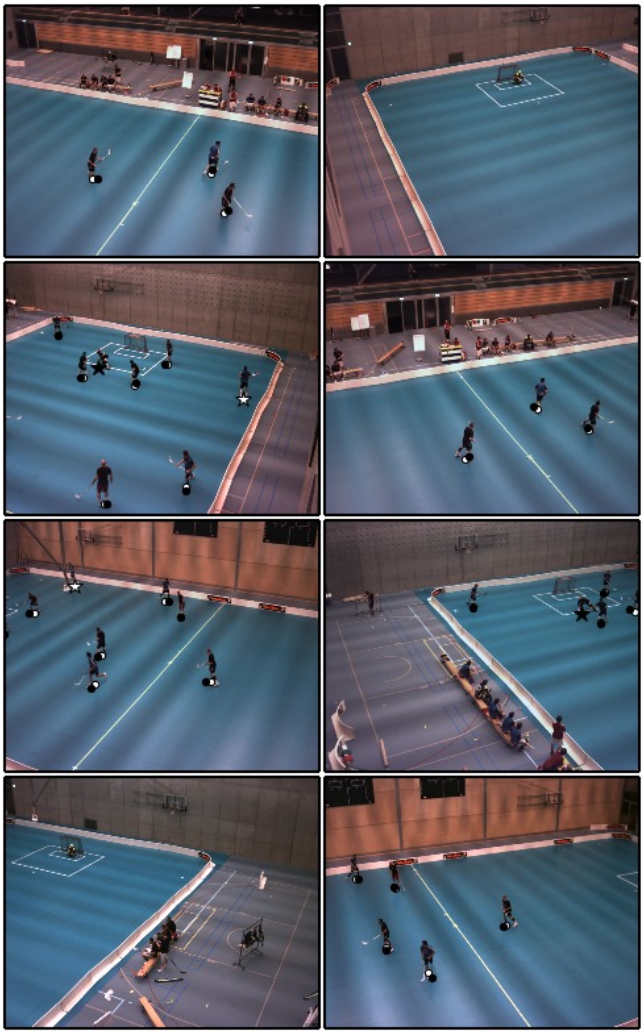

Fig. 8. Frame 160 from the floorball sequence with detected positions evaluated against ground truth. Black circles are detected positions, black stars are detected false positives, white circles are ground truth positions, white stars are false negatives.

Jan POKORNÝ – master student at the FEE, CTU – specialization in software engineering and computer vision. Jan's interest in computer vision began during his work on the bachelor thesis, after which he was offered an internship at Software Compentence Center Hagenberg GmbH. Jan is recently on the internship in Toyota Motor Europe, where he participates on the development of algorithms for 2D and 3D computer vision.