Electronic Communications of the EASST
Volume 62 (2013)

Specification, Transformation, Navigation
Special Issue dedicated to Bernd Krieg-Brückner on the Occasion of his 60th Birthday

(A) Vision for 2050 – Context-Based Image Understanding for a Human-Robot Soccer Match

Udo Frese, Tim Laue, Oliver Birbach, and Thomas Röfer

19 pages

Guest Editors: Till Mossakowski, Markus Roggenbach, Lutz Schröder
Managing Editors: Tiziana Margaria, Julia Padberg, Gabriele Taentzer
ECEASST Home Page: http://www.easst.org/eceasst/ ISSN 1863-2122

(A) Vision for 2050 – Context-Based Image Understanding for a Human-Robot Soccer Match

Udo Frese, Tim Laue, Oliver Birbach, and Thomas Röfer*

Udo.Frese@dfki.de
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH,
Cyber-Physical Systems, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany

Abstract: We believe it is possible to create the visual subsystem needed for the RoboCup 2050 challenge – a soccer match between humans and robots – within the next decade. In this position paper, we argue, that the basic techniques are available, but the main challenge will be to achieve the necessary robustness. We propose to address this challenge through the use of probabilistically modeled context, so for instance a visually indistinct circle is accepted as the ball, if it fits well with the ball's motion model and vice versa.

Our vision is accompanied by a sequence of (partially already conducted) experiments for its verification. In these experiments, a human soccer player carries a helmet with a camera and an inertial sensor and the vision system has to extract all information from that data, a humanoid robot would need to take the human's place.

Keywords: computer vision, context, RoboCup, robot soccer

1 Introduction

Soon after establishing the RoboCup competition in 1997, the RoboCup Federation proclaimed an ambitious long term goal (depicted in Figure 1).

> *"By mid-21st century, a team of fully autonomous humanoid robot soccer players shall win the soccer game, comply with the official rule of the FIFA, against the winner of the most recent World Cup."*
>
> *Hiroaki Kitano, Minoru Asada [KA98]*

Currently, RoboCup competitions take place every year. Within a defined set of different sub-competitions and leagues, incremental steps towards this big goal are made [Rob11]. Although a rapid and remarkable progress has been observed during the first decade of these robot competitions, it is not obvious, if and how the final goal will be reached. There exist rough roadmaps, e.g. by Burkhard et al. [BDF+02], but in many research areas, large gaps must be bridged within the next 40 years.

While this is obvious for several areas, e.g. actuator design and control, we claim that the situation is surprisingly positive for vision:

* This work has been supported by Deutsche Forschungsgemeinschaft grant FR2620/1-1 and E.C. project ECHORD (grant #231143), subproject GRASPY.

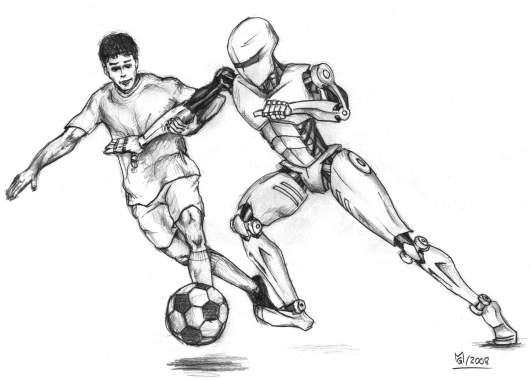


Figure 1: The RoboCup vision: An autonomous humanoid robot playing soccer against a human. Illustration courtesy of M. Görner.

> *Within the next decade, it will be possible to develop a vision system that is able to provide all environmental information necessary to play soccer on a human level.*

Annual RoboCup competitions are always bound to strict rule sets (defined for the state of the art of the competing robots) and demand competitive robot teams. Thus only incremental progress adapting to actual rule changes (which continuously raise the level of complexity) is fostered. By developing the aforementioned vision system independently of these competitions, we hope to set a new landmark which could guide the incremental development.

Because a *real* human level soccer robot will not be available for a long time, our vision is accompanied by a (partially already conducted) set of experiments that verify our claim without needing a robot.

This paper is organized as follows: Section 2 roughly identifies the challenges for playing robot soccer and compares them to the state of the art in robotics. In Section 3 we explain, why the basic techniques for the vision system are available. We argue, why the remaining challenge is robustness, for which we present our idea of a solution in Section 4. Finally, a sequence of experiments to verify our claim is described in Section 5.

2 Challenges for Playing Soccer

The global task of playing soccer consists of several different, interdependent challenges. We roughly categorize them according to the Sense-Think-Act cycle (see Figure 2). This should be considered as a possible architecture for illustration. In the following, the challenges are described in reverse order but with decreasing degree of difficulty.

2.1 Challenges for Actuation

The largest obvious gap may be observed in the field of actuation. Nowadays, the probably most advanced humanoid robot, Honda's ASIMO, is capable of running at a top speed of six kilometers per hour [Hon11]. This is an impressive result, but still more than five times slower than the

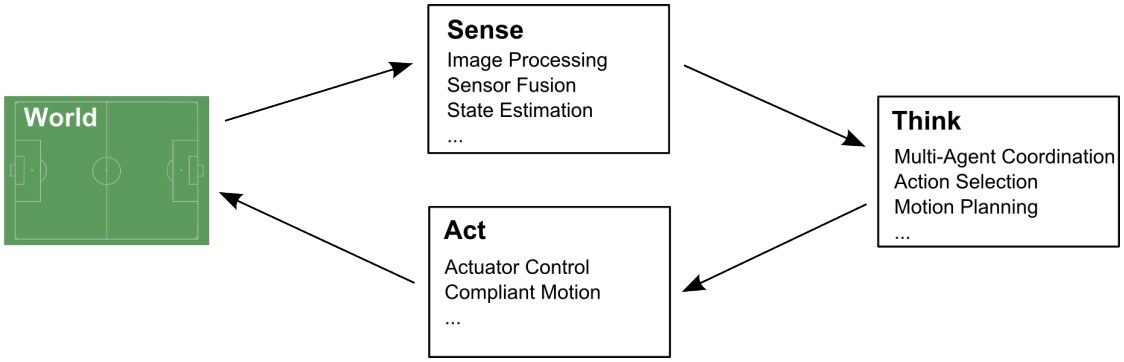


Figure 2: The Sense-Think-Act cycle roughly depicting major tasks for playing soccer with a humanoid robot.

top speed of a human soccer player. A similar gap regarding kicking velocity has been pointed out by [HLFH07, HLF+09]. They showed that a state-of-the-art robot arm (with a configuration comparable to a human leg) is six times slower than required to accelerate a standard soccer ball to an adequate velocity. It is still an open issue whether today's motor technology could be developed further on enough, or if more efficient actuators, e.g. artificial muscles, will be needed. Since soccer is a contact sport leading to physical human-robot interaction [HLFH07], not only direct position control but also approaches for compliant motion, such as impedance control, need to be taken into account. As pointed out by [HLF+09], joint elasticity appears to be a crucial aspect for velocity as well as for safety.

Additionally, the problems of energy efficiency and power supply need to be solved. The ASIMO robot for example is, according to [Hon11], capable of walking (with a speed of less than three kilometers per hour) for 40 minutes.

2.2 Challenges for Thinking

In this area, two different levels may be distinguished: motion planning and high-level multi-agent coordination. The latter is a research topic in the RoboCup Soccer Simulation League since a while and has reached a remarkable level. Dealing with the offside rule as well as playing one-two passes are standard behaviors, complex group tasks as playing keep-away soccer serve as a test bed for learning algorithms [SSK05]. This area could be considered to be already quite close to human capabilities.

On the other hand, when playing with real humanoid robots, sophisticated methods for motion planning are needed. The current research frontier on humanoid motion control is balancing and dynamic foot placement for walking robots. Algorithms for full-body motion planning exist [KKN+02], but are subject to restrictions that make them inapplicable to tasks as playing soccer.

Here is a big gap to human level soccer. As an example consider volley-kicking. The player has to hit the ball exactly at the right time, position, and velocity, with a motion compatible to the step pattern, allowing balancing, and considering opponents. Last but not least, all this must happen in real-time.

2.3 Challenges for Sensing

According to [KA98], it is evident that the robots' sensorial capabilities should resemble the human ones. The philosophical considerations of [SQH10], who evaluate the compatibility of the current FIFA rules and potential robot contenders, come to the same conclusion. Thus, we could assume to deal with data from up to two cameras and an inertial sensor, both mounted inside a robot's head, emulating the human eyes and vestibular system. The required information are estimates of the own position and the positions of the ball and of other players. In case of tackles or dribbling, the latter will be needed to be recognized in more detail (e.g. the positions of the feet and limbs).

Current solutions for these tasks and our idea how to bridge the remaining gap are presented in the following section.

3 The Vision System

Our main thesis is that the "sense" part of the RoboCup 2050 challenge can be realized within a decade starting from the current state of the art in computer vision. This is remarkable, since the "act" and "think" parts are apparently far away from reaching human level performance and for computer vision in general, this is also true. The reason, why we believe such a vision system can be realized, is that unlike a household robot for instance, a soccer robot faces a rather structured environment.

3.1 State of the Art

Relevant objects in a soccer match can be classified into two categories, static and dynamic ones. Line markings and goals are important static features and provide the field's boundaries. Detection of these in images is achieved through their known geometric appearance, using classical Hough transform [Bal81, YPIK89] or more elaborate methods [GS04].

Dynamical objects include players, referees and the ball. For the latter, its simple geometric appearance allows detection using well-known methods [Bal81, YPIK89]. As the ball might be accelerated up to $129 km/h$ [Wes02] from an instep kick (and even more when the ball has moved towards the player prior to the shot) its robust detection might be challenging for current sensing technology. This problem also holds when detecting other players. It is particularly difficult because we will need not only the general position but the detailed state of motion (articulated person tracking) for close range tackling and to infer the player's action for tactical purposes. For example, biomechanical characteristics of the ball kick [KK07, LAA+10] revealed that the knee might flex at a velocity of $860^\circ/s$ prior to the kick. Such rapid limb movement considerably increases the challenge detecting the full state of the players. Fortunately, people tracking is an important topic in computer vision with a large body of literature [FAI+05, MHK06]. Approaches which are capable of estimating the body pose as well as the one of the limbs in general [ARS10] and of sportsmen [RF03] (see Figure 3) have already been presented. Tracking of multiple persons while the observer is moving are studied in [ARS10, ELG07].

Furthermore, soccer scenes are well lit and lightly colored with green lawn and the players wearing colored clothes of high contrast. In the RoboCup competition, this idea is taken to an

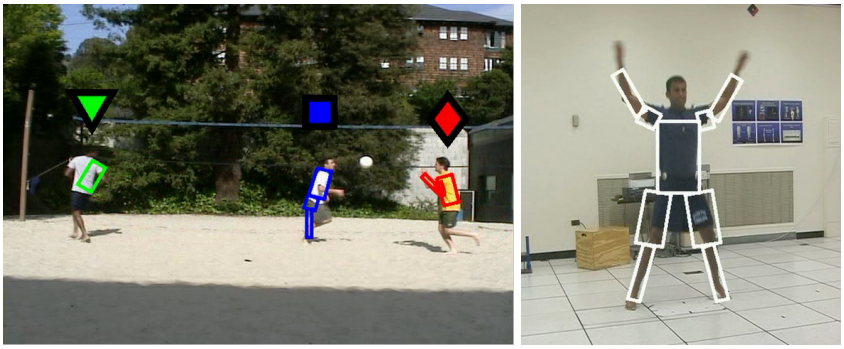


Figure 3: Tracking sportsmen and their limbs: Three persons which are running and passing a ball (left image), one person performing jumping jacks (right image). Both images are results of [RF03] and are used with permission of Deva Ramanan.

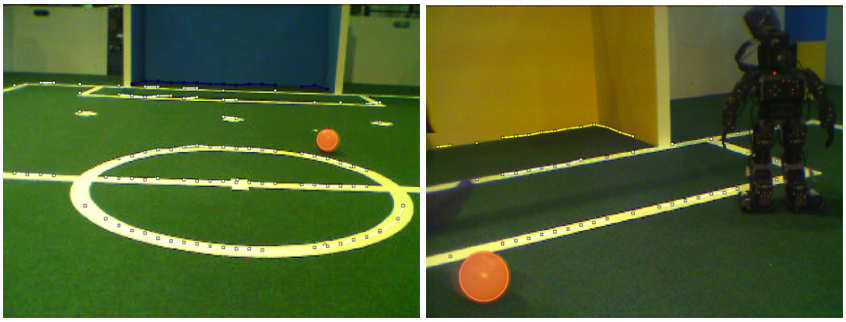


Figure 4: Two images taken and processed by a humanoid soccer robot on a standard RoboCup field [LR07]. The colored dots and lines indicate perceived parts of goals and field lines; the recognized ball is bordered by an orange circle.

extreme, where most teams rely on color segmentation on a pixel-per-pixel basis as their primary vision engine (see Figure 4). This will not be possible for real-world soccer, mainly due to changing lighting conditions. Still, color can provide a valuable additional cue, at least when looking below the horizon, where objects are in front of green lawn.

The background above the horizon, including the stadium and the audience is of course also visible and unfortunately rather undefined. However, if it is relevant for the soccer robot at all, then not for recognition, but only in the sense of a general landmark. For this purpose there are nowadays well working techniques, such as the Scale Invariant Feature Transform (SIFT) [Low04].

Overall, understanding a soccer scene from the player's perspective seems much easier than for instance understanding an arbitrary household, traffic, or outdoor scene. Indeed there are already half-automatic systems in the related area of TV soccer scene analysis, for example the ASPOGAMO system by [BHB+06, BGB+07], proofing that soccer scene understanding in general is on the edge of being functional.

3.2 Open Problems

So, is a vision system for the RoboCup 2050 challenge an easy task? We believe it is not. It is surprisingly a realistic task but well beyond the current state of the art. The first problem is that the camera is moving along with the head of the humanoid soccer robot. To predict a flying ball, the orientation of the camera must be known very precisely. It seems unrealistic that the necessary precision can be obtained from the robot's forward kinematic, since unlike an industrial robot, a humanoid robot is not fixed to the ground. So our solution is to integrate an inertial sensor with the camera and fuse the complementary measurements of both sensors in a probabilistic least-squares framework.

The second problem is the player's perspective. It is much more difficult than the overview perspective used in TV soccer scene analysis. In the TV perspective, the scale of an object in the image varies by a factor of about 3 [BHB+06, Figure 5] whereas in the player's perspective, it can vary by a factor of 250 assuming the distance to an object ranging from 0.5m to 125m. Hence, for instance the people detection algorithm must handle both extreme cases, a person only the size of a few pixels, where an arm or a leg may be thinner than a single pixel and a person much larger than the camera's field of view, only partially visible. Indeed, while for far away players it is probably sufficient to track their position as a point, precise articulated person tracking is needed for close range tackling and probably the most difficult challenge.

Furthermore, in an image from the player's perspective, other players will extend beyond the green lawn of the field into the general background. Hence it is not possible to search for non-green blobs as an easy first processing step. This can also happen for a flying ball, which is then particularly difficult to detect.

However, the third and not to be underestimated problem is that although computer vision systems that operate in natural environments (*i.e.* not requiring nicely setup scenes and lighting conditions) are making progress (see [ARS10, ELSG09] for challenging people detection), there is still plenty of work left to understand the broad range of natural scenes. So, to summarize, for the vision part of the RoboCup 2050 challenge, we do not need a new level of functionality as for many other grand challenges, but we need a new level of robustness.

4 Robustness Through Context

We propose to address the question of robustness by utilizing probabilistically modeled context information as a central idea. Conceptually, the overall scene understanding and prediction problem is formulated as a global likelihood optimization task. Using contextual associations is not entirely new [Ull95, SAS+07] and recent progress has shown that exploiting context leads to promisingly robust results. Recent work worth noting includes modeling of object to object relations as done by [RVG+07]. Here, the authors improved the object categorization accuracy by integrating a post-processing step that maximizes the object categorization agreement based on contextual relevance. More generally, [KH05] proposed a hierarchical framework allowing interaction between objects and regions within the image plane.

Leaving the image plane and therefore 2D models behind, [HEH06] propose a mechanism for estimating rough 3D scene geometry from a single image which is then used as an additional input to the object detection methods. Therefore, detections which fail to fit into the estimated scene

geometry at the detected position are rejected from the global scene interpretation (e. g. pedestrians who do not stand on the ground plane). This idea is taken one step further by [ELG07] and [LSCG08], where the relationships between objects and scene geometry are modeled, allowing suitable recognition of objects for a mobile robot application.

These promising results show that exploiting context within the image and the scene is particularly well suited to the task of creating a vision system for RoboCup 2050 and conversely that this task is well suited to study this methodology. In this paper, we will make this conceptual viewpoint concrete with the example of ball tracking. Similar thoughts, however in a more complex way, are of course also applicable to other vision tasks, in particular (articulated) people tracking.

4.1 Data-Driven Bottom-Up Processing

Most current vision systems use a data-driven bottom-up approach [FBH+01, R+05, BGB+07, as examples]. Usually, low level features are extracted from the image and then aggregated through several stages to high level information. Each stage may incorporate background knowledge at its own level but does not take information from higher levels into account. It simply takes some input from the previous lower level and passes the result of the computation to the next higher level.

As an example, a classical Hough transform starts by classifying pixels as edge or not by thresholding the result, for instance, of a Sobel filter. Similarly, the ASPOGAMO system [BGB+07] starts by classifying pixels as lawn or not on the basis of their color. This is a hard decision taken on the lowest level without any higher level knowledge, such as the fact that we are looking for a ball or the ball's motion model. Such a pixel-wise classification can be very ambiguous. Often we could, for instance, classify a borderline pixel correctly as belonging to the ball, although it looks rather greenish, if we considered the context of the ball or its motion model. However, in conventional vision systems, this knowledge does not exist on the low level and on the higher level, the fact that this pixel was borderline in the classification is lost due to committing to a hard decision on the lower level.

To make this idea more concrete, we will now describe a ball tracking system in the conventional data-driven design (Figure 5). We will later discuss how it would look like in the context-based design we propose in this paper.

In a data-driven approach, a *ball finder* algorithm is executed on every image yielding the 2D image circle (position, radius) corresponding to the ball (or several hypotheses thereof). These circles are passed to a probabilistic estimator, such as Gaussian-Maximum-Likelihood that finds the mostly likely states, i. e. ball positions and velocities. Such an estimator is based on model likelihoods, i. e. functions with uncertainty, that specify the following:

1. A *camera model* defines when the ball was at a certain position, where it would be in the image. This links 3D positions to 2D circles.

2. A *ball motion model* defines when the ball was at a certain position having a certain velocity, where it would be and which velocity it would have after one timestep. This links 3D positions and velocities over time.

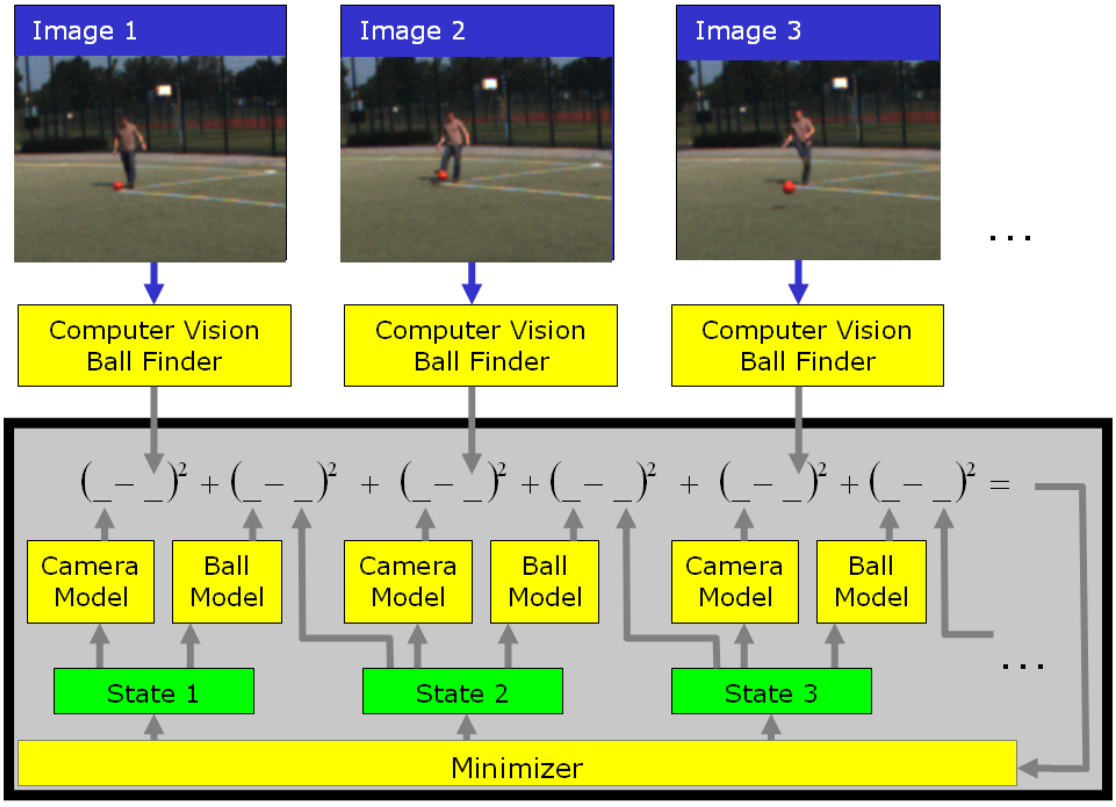


Figure 5: Information flow in a ball tracker based on the conventional data-driven paradigm (see text). The ball finder is not part of the optimization loop (gray box).

The estimator then finds the most likely states by minimizing a sum of squared differences, which indicates how likely an assumed sequence of states is. It consists of 1. the squared differences between where the ball has been observed by the ball finder and where the ball model would expect it to be observed if the assumed state sequence was true and 2. the squared difference between the states predicted by the ball model and the assumed corresponding states.

The models 1. and 2. are Gaussian likelihoods, i. e. formulas with uncertainty, so there is a wealth of algorithms for handling this kind of problems. In particular, figure 5 is a conceptual description. Most actual implementations do not minimize over the whole state sequence in every step but instead use an incremental estimator, such as the Extended or Unscented Kalman Filter [WB97, Bir08, Kur07]. However, this changes only the computation not the result. The key point of this approach is that the ball finder, more general the computer vision, is outside the optimization loop and there is no information flow from the estimator back to the ball finder. Hence the ball finder cannot take the high level context of the ball's motion into account when deciding on the low level, whether something is accepted as a ball or not.

4.2 Global Likelihood Optimization

A common experience in computer vision research is that many systems are not very robust regarding object appearance and illumination conditions. We believe that much of this brittleness originates from the data-driven bottom-up way of committing to hard decisions on an early level of processing. So our approach for increased robustness is to avoid early decisions and instead to conceptually take an overall likelihood optimization view on the problem.

The problem is to understand an image sequence, i. e. estimating over time. Indeed, successive images are linked by a motion model and this provides most of the context we want to build upon. However, we propose not to use incremental filters, such as EKF, but to look back into the raw images of the last few seconds at least. This approach has surprising advantages. Imagine the ball is kicked, but during the first 100ms there is too little contrast to the background so it is not detected. Now when it is detected, there is new information on where the ball has been before from the ball's motion model. The old images are still in memory and tracking the ball back in time is much less ambiguous than finding the ball without context and will probably succeed. Paradoxically, once the system has detected the ball, it has already observed it for 100ms. The first useful prediction is not delayed at all, because prior to that the ball must have been observed for some time anyway. In the example above, we would realize this (Figure 6) by replacing the ball finder by

3. a *ball appearance model* that indicates for a given circle in the image plane how much it looks like a ball. This links the 2D image circle to the actual image.

In the data-driven approach, the ball finder determines where the ball is and the camera model computes where the ball should be according to the state. Then both are compared with the difference indicating the likelihood of the state. In the global likelihood optimization approach, instead the camera model computes where the ball should be and then the ball appearance model computes how "ball-like" the image looks there. The difference is, that now the ball appearance model, i. e. the computer vision, is inside the optimization loop. Now the lower layer does not have to commit early to a single image circle, but instead it gradually assesses different circles in the image as requested by the minimization algorithm. In this approach, an indistinct ball would get a lower likelihood in 3. but this could be compensated by 1. and 2. if it fits well to the context of a flying ball. This mechanism allows the now implicit ball finder to utilize the context provided by the ball motion and implement the behavior described above.

At first sight, this approach resembles Multi Hypotheses Tracking (MHT) [CH96], however the MHT goes only half the way to the behavior in Figure 6. The MHT maintains multiple hypotheses on which observation (i. e. a circle) is caused by which target, including the hypothesis of a false alarm caused by no target. However, it requires the computer vision to commit to a discrete set of observations, which are then input to the MHT. So the discussed example works with the MHT only if the initial low-contrast ball is still distinct enough to become an observation.

Conceptually, a particle filter can implement the behavior in Figure 6 equivalently, where all past information is aggregated in the particle set and it is not necessary to store or operate on past images. In this case, the discussed example requires enough particles such that even the indistinct initial ball creates some particles which are then selected later when further images and the context support them. It will be a future question if this is more or less practical than an

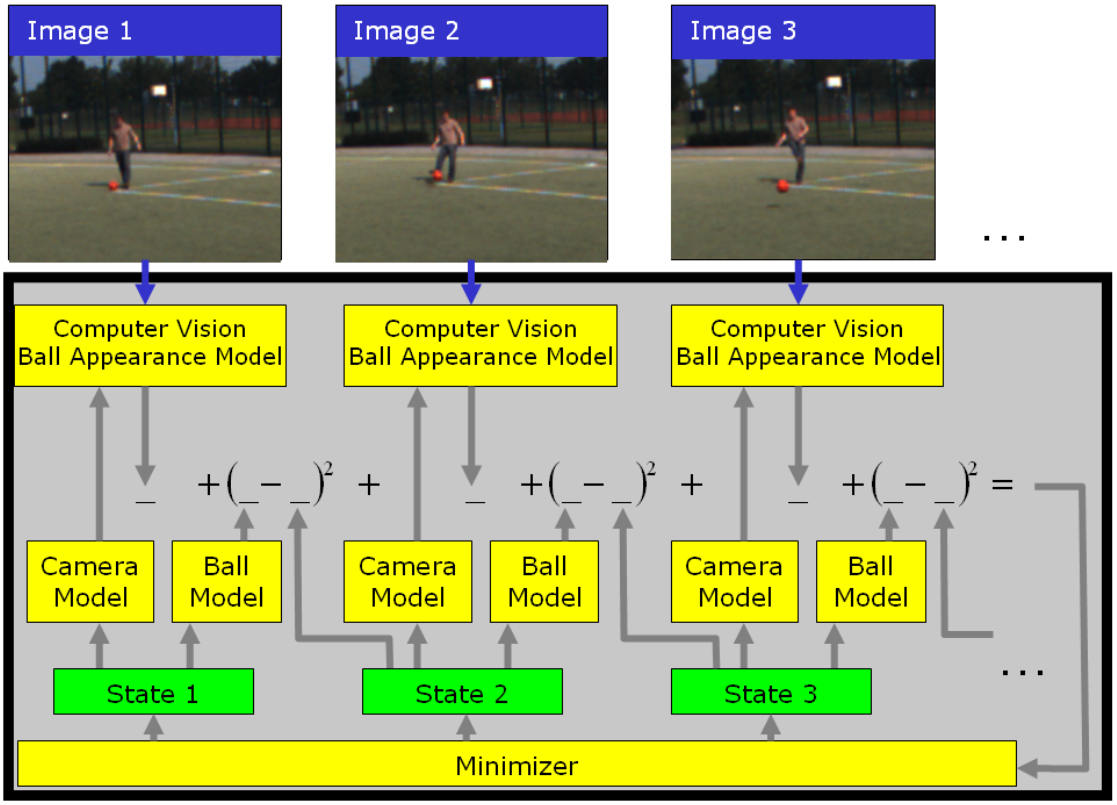


Figure 6: Information flow in a ball tracker based on the global likelihood optimization paradigm (see text). Here, computer vision, i.e. the ball appearance model, is part of the optimization loop (gray box).

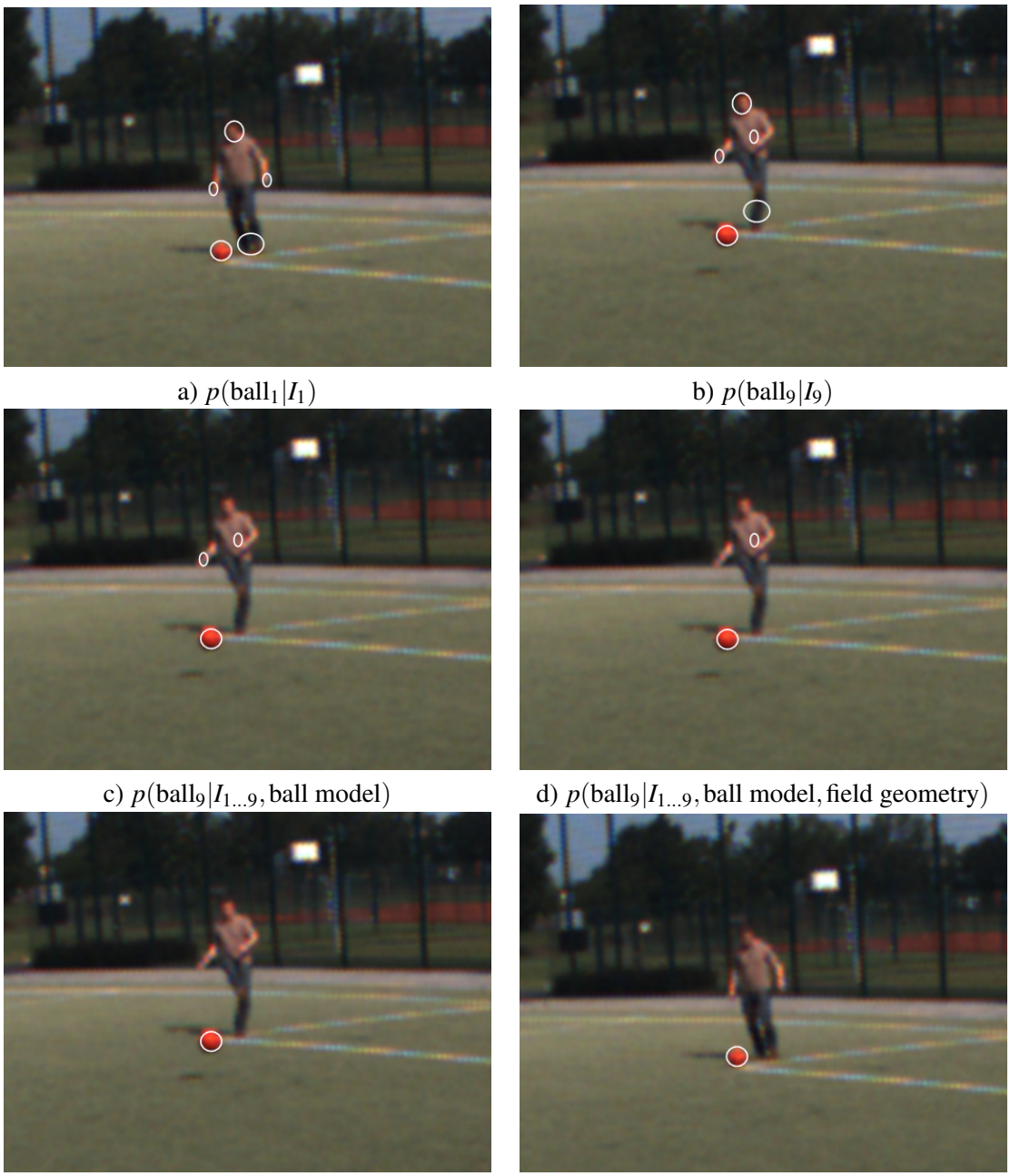
implementation operating on past images.

As a final remark, the provided functionality greatly resembles the priming mechanism known from psychology [KW03]. This makes us even more believe that the approach of a global likelihood optimization directly in the images is an elegant way to greatly increase robustness of computer vision algorithms.

4.3 Exploiting Semantic Context

Doubtless, if the goal is to track a flying ball, the ball's motion model provides the most valuable context. However, there is other, more semantic information that could be incorporated to increase robustness. Figure 7 illustrates how a broader context could help distinguishing the real ball from false responses of the computer vision. We have constructed this example to illustrate what we believe could happen. It is not based on real computation.

In a), only the first image I_1 is considered ($p(\text{ball}_1|I_1)$). The computer vision gives varying responses for different image circles. After applying the necessary threshold needed to give a final result, there are five possible ball locations in the example, namely the true ball and the



a) $p(\text{ball}_1|I_1)$

b) $p(\text{ball}_9|I_9)$

c) $p(\text{ball}_9|I_{1\ldots9}, \text{ball model})$

d) $p(\text{ball}_9|I_{1\ldots9}, \text{ball model}, \text{field geometry})$

e) $p(\text{ball}_9|I_{1\ldots9}, \text{ball model}, \text{field geometry}, \text{game semantics})$

f) $p(\text{ball}_1|I_{1\ldots9}, \text{ball model}, \text{field geometry}, \text{game semantics})$

Figure 7: An illustration of how different types of context information could help to distinguish the real ball from things that look like a ball to the computer vision (see text).

player's head, hands, and both feet together. The same happens in b), where in the example the 9th image I_9 is considered alone ($p(\text{ball}_9|I_9)$). Now in c) all images $I_{1...9}$ are considered together ($p(\text{ball}_9|I_{1...9}, \text{ball model})$). The link between the images is provided by the ball motion model as described in the previous section. In the example, the player's head and feet could be ruled out, because their motion does not correspond to the ball motion model. However, in this example, the two hands remain valid hypotheses, because their side-swinging motion would actual correspond to a distant flying ball. In d), the additional context of the field geometry is added ($p(\text{ball}_9|I_{1...9}, \text{ball model}, \text{field geometry})$). This context makes the player's right hand being a ball very unlikely, because from the image position and radius of the circle, the ball would be outside of the field. In e), additional information about game semantics is added $p(\text{ball}_9|I_{1...9}, \text{ball model}, \text{field geometry}, \text{game semantics})$. From the game situation, the system expects the player to kick the ball towards the camera. A ball flying sideward as the falsely detected left hand is implausible with this context and can be ruled out. Of course, as explained above, once the system has found a unique ball in image 9, it already knows backwards, via the ball motion model, which was the correct ball in image 1. This situation is shown in f) ($p(\text{ball}_1|I_{1...9}, \text{ball model}, \text{field geometry}, \text{game semantics})$). So the system has already observed the ball for 9 images and can immediately provide a prediction.

How could this example be realized? First of all, using game semantics requires to detect the players, lines and goals to have information about the game situation. But how could the semantics of soccer be modeled? The classical AI approach would be to use logic, for instance ontologies, to describe a soccer match. However, typical ontologies, such as OWL [BHH+04], express crisp knowledge in some sort of logic and are difficult to combine with the probabilistic and vague knowledge discussed in the previous section.

While in the long run modeling a rich and complex semantics, as ontologies provide, is desirable, probably the first step would be to start with a shallow semantics that is more easily integrated with probabilistic motion and perception models. A candidate would be to run a Hidden Markov Model (HMM) [HTWM04] for each player to model the different actions (kicking, running, tackling, etc.) a player may take, maybe subdivided into different phases. Then one could define how each phase would affect the player's motion and the motion of the ball. The advantage is that easy algorithms for inference in HMM exist and even more important that HMM can be well combined with continuous probabilistic estimation by the Interacting Multiple Model Filter (IMM) [BLK01, §11.6.6].

Overall, it is obvious that modeling the player's behavior is essential for understanding a soccer match as a whole and is helpful even for tracking the ball. However, it is still rather unclear how the required background knowledge can be modeled in a way that can effectively help a computer vision system.

5 Proposed Experiments

For a vision to become reality, realistic intermediate steps are necessary. It would not help, if we build a vision system now but then had to wait until a human level soccer robot is available. So we propose a sequence of experiments that, without a humanoid robot, ultimately allows to verify that the proposed system is appropriate for human level soccer (Figure 8).

Figure 8: Our proposed experiment: Mount a camera and an inertial sensor on the head of a human soccer player and use them to extract all the information, a humanoid soccer robot would need to take the human's place.

5.1 Helmet Camera with Inertial Sensor

The basic idea is to let a human soccer player wear a helmet with a camera and an inertial sensor and verify that the information extracted by the vision system from the sensor data would allow a humanoid robot to take the human's place.

As a first experiment, we propose to record data from a soccer match and run the vision system on that data offline. Since it is hard to obtain ground-truth data, we would use our expert's judgment to assess, whether the result would be enough for a humanoid robot to play soccer. It is very advantageous to work on recorded data allowing to reproduce results for debugging and analysis and to run the system even if it is still not real-time. Overall, it allows to first concentrate on functionality and robustness instead of computation time and integration.

We have already conducted a very first experiment [Kur07, Bir08], where the ball and the field lines are manually extracted from the recorded images (available on request). The ball's trajectory is predicted by least-square estimation using the likelihood functions 1. and 2., as well as corresponding equations for how the inertial sensor observes the free motion of the camera (Figure 9). The results indicate that if the ball can be detected in the image with about one pixel precision, the prediction would be precise enough. We believe that these kinds of studies which deliberately discard essential aspects, such as integration, real-time computation, or autonomy are undervalued by the RoboCup community who favors full system approaches. But even from a full system perspective, it is much more valuable to obtain an extensive result on a subsystem which then can guide the full system design than to do another increment on a full system.

5.2 Multi-Hypothesis Ball-Tracking

More recently, we presented a computer vision system for tracking and predicting flying balls in 3-D from a stereo-camera in real-time [BF09]. The system consists of two stages. First, a robust circle detector which avoids hard decisions and thresholds extracts the most ball-like circles

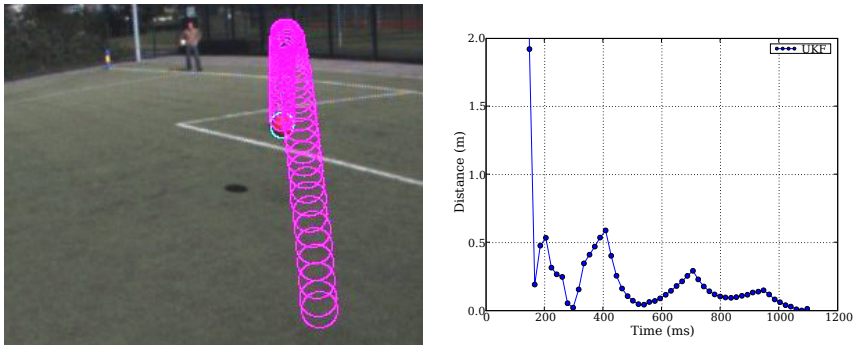


Figure 9: Predicting the trajectory of a flying ball from a moving camera-inertial system. As an initial study, the ball, the lines, and the goal corners have been manually extracted from the images. From this data, the trajectory of the ball is predicted (left). The right plot shows the error of the predicted touchdown point varying over time. It shows that, even though the camera is moving, the prediction is roughly precise enough for interception.

from the images. Then, measurements of both images are fused by a probabilistic multiple-target tracking algorithm and the most probable tracks are returned accounting for appearing and disappearing balls as well as false and missing measurements. Compared to the global likelihood optimization approach of Section 4.2, there is still the hard decision of whether a circle makes it into the measurement list passed to the MHT or not. However, unlike for other circle detectors, i. e. the Hough-transform, this is the only hard decision.

In the corresponding experiments, four people were throwing up to three volleyballs to each other (Figure 10). Although the circle detector detected balls with just a rate of 69.4% and the multiple-target tracker accidentally mis-associated measurements and tracks, the system successfully tracked all 32 visible trajectories.

To assess the accuracy of such a system, we conducted lab-experiments and their preliminary results will be presented here. We used the proposed setup from section 5.1 extended by a second camera for stereo (1616×1220 pixels resolution at $25Hz$). Several balls where thrown towards the cameras and tracked by the system described in the paragraph above. Additionally, an infrared tracking system provided ground-truth by tracking the ball positions independently. The balls were covered with retro-reflective foil for that purpose. Comparing the predicted trajectories after each integrated image pair with the ground-truth position of the ball when it passes the cameras allows us to evaluate the prediction performance. From our results it can be seen that an accuracy of about $25cm$ is reached $240ms$ (six frames) after the ball left the thrower's hand. Furthermore, an accuracy below $2cm$ is achieved when the ball leaves the camera image. Such a precision is enough to use the proposed system to let a robot catch thrown balls, as we showed in ongoing research [BSW+11, BFB11].

The above-mentioned system follows the data-driven bottom up approach as outlined in section 4.1. Especially the performance in the beginning can be improved by integrating the ball's motion as context. Because of imprecise detections in the beginning due to small balls in the image, tracking results are poor. By using the ball's motion as context these tracking results could be improved. We already use prediction to guide the search for circles in the image. This greatly

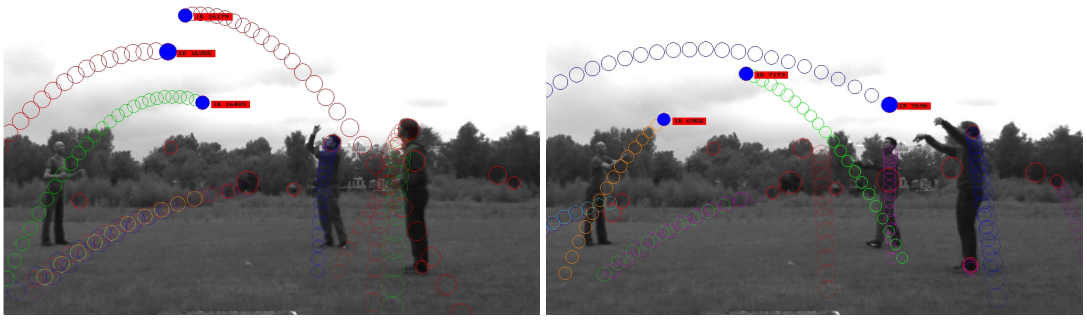


Figure 10: Camera view of the tracking scenario and overlaid detection results (red circles) and predicted ball trajectories (coloured circles).

increases tracking robustness but works only for *future* images. Within the global maximization approach one could also look for better suiting circles in *past* images as proposed in section 4.2. After such a global optimization we expect the tracking performance to be improved, since no early commitment to possibly wrong measurements is made.

5.3 Motion Capture Suit

Departing from the experiment above, one might ask whether more sensors are needed than just camera and inertial. Both human and humanoid robot can derive their motion from the joint angles. This provides the horizontal motion (odometry) and the height over ground. The horizontal motion facilitates localization and the height derived from vision is much less precise. Indeed, we experienced that the uncertain height is a major part of the error in Figure 9.

An intriguing idea to solve this gap is to equip the human player with a *tracker-less* motion capture suit [Xse07] measuring joint angles. Apart from kinematic information, the inertial sensors in such a suite could also detect collisions between the foot and the ball or between the foot and the ground thereby providing most of the information that humans and robots obtain from the haptic senses.

5.4 Virtual Reality Display

The experiments above have the drawback that they are evaluated by an expert looking at the vision system's output. The most direct proof that this is all you need for playing soccer would be to give a human just that output via a head mounted display and see whether s/he can play.

The approach is of course fascinating and direct, but we have some concerns regarding safety. Anyway, this experiment becomes relevant only after we are convinced in principle, that the system is feasible. So this is something to worry about later.

6 Conclusion

In this position paper, we have outlined the road to a vision system for a human-robot soccer match. We claim that, since soccer is a structured environment, the basic techniques are available

and the goal could be reached within a decade. The main challenge will be robustness, which we propose to address by optimizing a global likelihood function working on a history of raw images. We have outlined a sequence of experiments to evaluate such a vision system with data from a camera-inertial system mounted on the head of a human soccer player.

The reason for being confident such a system can be realized within a decade is the insight that it does not need general common-sense-reasoning AI. This is good news for the RoboCup 2050 challenge. But it suggests that, even when we meet that challenge, it does not imply we have realized the dream of a thinking machine, the whole challenge had started with.

That would not be the first time.

Bibliography

[ARS10] M. Andriluka, S. Roth, B. Schiele. Monocular 3D Pose Estimation and Tracking by Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2010.

[Bal81] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition* 13(2):111 – 122, 1981.

[BDF+02] H.-D. Burkhard, D. Duhaut, M. Fujita, P. Lima, R. Murphy, R. Rojas. The Road to RoboCup 2050. *IEEE Robotics and Automation Magazine* 9(2):31–38, 2002.

[BF09] O. Birbach, U. Frese. A Multiple Hypothesis Approach for a Ball Tracking System. In *Computer Vision Systems*. LNCS 5815, pp. 435–444. 2009.

[BFB11] O. Birbach, U. Frese, B. Bäuml. Realtime Perception for Catching a Flying Ball with a Mobile Humanoid. In *Proceedings of the 2011 International Conference on Robotics and Automation (ICRA)*. Shanghai, China, 2011.

[BGB+07] M. Beetz, S. Gedikli, J. Bandouch, B. Kirchlechner, N. v. Hoyningen-Huene, A. Perzylo. Visually Tracking Football Games Based on TV Broadcasts. *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India*, 2007.

[BHB+06] M. Beetz, N. v. Hoyningen-Huene, J. Bandouch, B. Kirchlechner, S. Gedikli, A. Maldonado. Camerabased Observation of Football Games for Analyzing Multi-agent Activities. In *International Conference on Autonomous Agents*. 2006.

[BHH+04] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, L. Stein. OWL Web Ontology Language Reference. Technical report REC-owl-ref-20040210/, W3C, 2004.

[Bir08] O. Birbach. Accuracy Analysis of Camera-inertial Sensor Based Ball Trajectory Prediction. Master's thesis, Universität Bremen, 2008.

[BLK01] Y. Bar-Shalom, X. Li, T. Kirubarajan. *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., 2001.

[BSW+11] B. Bäuml, F. Schmidt, T. Wimböck, O. Birbach, A. Dietrich, M. Fuchs, W. Friedl, O. Eiberger, M. Grebenstein, C. Borst, U. Frese, G. Hirzinger. Catching Flying Balls and Preparing Coffee: Humanoid Rolling Justin Performs Dynamic and Sensitive Tasks. In *Proc. of the 2011 IEEE Int. Conf. on Robotics and Automation (ICRA 2011)*. 2011. (Video).

[CH96] I. J. Cox, S. L. Hingorani. An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the Purpose of Visual Tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 18(2), 1996.

[ELG07] A. Ess, B. Leibe, L. V. Gool. Depth and Appearance for Mobile Scene Analysis. In *Proceedings of the IEEE International Conference on Computer Vision*. 2007.

[ELSG09] A. Ess, B. Leibe, K. Schindler, L. van Gool. Robust Multiperson Tracking from a Mobile Platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(10):1831–1846, 2009.

[FAI+05] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, D. Ramanan. Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis. *Foundations and Trends in Computer Graphics and Vision* 1(2/3):77–254, 2005.

[FBH+01] U. Frese, B. Bäuml, S. Haidacher, G. Schreiber, I. Schaefer, M. Hähnle, G. Hirzinger. Off-the-Shelf Vision for a Robotic Ball Catcher. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Maui*. Pp. 1623 – 1629. 2001.

[GS04] D. Guru, B. Shekar. A simple and robust line detection algorithm based on small eigenvalue analysis. *Pattern Recognition Letters* 25(1):1–13, 2004.

[HEH06] D. Hoiem, A. A. Efros, M. Hebert. Putting Objects in Perspective. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2006.

[HLF+09] S. Haddadin, T. Laue, U. Frese, S.Wolf, A. Albu-Schäffer, G. Hirzinger. Kick it with Elasticity: Safety and Performance in Human-Robot Soccer. *Robotics and Autonomous System - Special Issue on Humanoid Soccer Robots* 57(8):761–775, 2009.

[HLFH07] S. Haddadin, T. Laue, U. Frese, G. Hirzinger. Foul 2050: Thoughts on Physical Interaction in Human-Robot Soccer. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2007.

[Hon11] Honda Worldwide Site. Honda World Wide — Asimo. 2011. http://world.honda.com/ASIMO/

[HTWM04] W. Hu, T. Tan, L. Wang, S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics, Part C* 34(3):334–352, 2004.

[KA98] H. Kitano, M. Asada. RoboCup Humanoid Challenge: That's One Small Step for A Robot, One Giant Leap for Mankind. In *International Conference on Intelligent Robots and Systems, Victoria*. Pp. 419–424. 1998.

[KH05] S. Kumar, M. Hebert. A Hierarchical Field Framework for Unified Context-Based Classification. In *Proceedings of the IEEE International Conference on Computer Vision*. 2005.

[KK07] E. Kellis, A. Katis. Biomechanical characteristics and determinants of instep soccer kick. *Journal of Sports Science and Medicine* 6:154–165, 2007.

[KKN+02] J. J. Kuffner, S. Kagami, K. Nishiwaki, M. Inaba, H. Inoue. Dynamically-stable Motion Planning for Humanoid Robots. *Auton. Robots* 12(1):105–118, 2002.

[Kur07] J. Kurlbaum. Verfolgung von Ballflugbahnen mit einem frei beweglichen Kamera-Inertialsensor. Master's thesis, Universität Bremen, 2007.

[KW03] B. Kolb, I. Wilshaw. *Fundamentals of Human Neuropsychology*. Worth Publishers, 2003. pp. 453-454, 457.

[LAA+10] A. Lees, T. Asai, T. B. Andersen, H. Nunome, T. Sterzing. The biomechanics of kicking in soccer: A review. *Journal of Sports Sciences* 28:805–817, 2010.

[Low04] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2):91 – 110, 2004.

[LR07] T. Laue, T. Röfer. Particle Filter-based State Estimation in a Competitive and Uncertain Environment. In *Proceedings of the 6th International Workshop on Embedded Systems*. VAMK, University of Applied Sciences; Vaasa, Finland, 2007.

[LSCG08] B. Leibe, K. Schindler, N. Cornelis, L. V. Gool. Coupled Object Detection and Tracking from Static Cameras and Moving Vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(10):1683–1698, 2008.

[MHK06] T. B. Moeslund, A. Hilton, V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104(2):90–126, 2006.

[R+05] T. Röfer et al. GermanTeam RoboCup 2005. 2005. http://www.germanteam.org/GT2005.pdf.

[RF03] D. Ramanan, D. Forsyth. Finding and tracking people from the bottom up. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2003.

[Rob11] RoboCup Federation. RoboCup Official Site. 2011. http://www.robocup.org

[RVG+07] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie. Objects in Context. In *Proceedings of the IEEE International Conference on Computer Vision*. 2007.

[SAS+07] G. Schmidt, M. Athelogou, R. Schönmeyer, R. Korn, G. Binnig. Cognition Network Technology for a Fully Automated 3D Segmentation of Liver. In *Proceedings of the MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge*. Pp. 125–133. Brisbane, Australia, 2007.

[SQH10] P. Stone, M. Quinlan, T. Hester. Can Robots Play Soccer? In Richards (ed.), *Soccer and Philosophy: Beautiful Thoughts on the Beautiful Game*. Popular Culture and Philosophy 51, pp. 75–88. Open Court Publishing Company, 2010.

[SSK05] P. Stone, R. S. Sutton, G. Kuhlmann. Reinforcement Learning for RoboCup-Soccer Keepaway. *Adaptive Behavior* 13(3):165–188, 2005.

[Ull95] S. Ullman. Sequence Seeking and Counter Streams: A Computational Model for Bidirectional Information Flow in the Visual Cortex. *Cerebral Cortex* 5(1), 1995.

[WB97] G. Welch, G. Bishop. An Introduction to the Kalman Filter. Technical report TR 95-041, University of North Carolina, 1997.

[Wes02] J. Wesson. *The Science of Soccer*. IOP Publishing Ltd., Dirac House, Temple Bacl, Bristol, 2002.

[Xse07] Xsens Technologies B.V. Moven, Inertial Motion Capture, Product Leaflet. 2007.

[YPIK89] H. Yuen, J. Princen, J. Illingworth, J. Kittler. A Comparative Study of Hough Transform Methods for Circle Finding. In *Alvey Vision Conference*. Pp. 169–174. 1989.