(A) VISION FOR 2050
The Road Towards Image Understanding for a Human–Robot Soccer Match

Udo Frese
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), 28359 Bremen, Germany
Udo.Frese@dfki.de

Tim Laue
SFB/TR 8 Spatial Cognition, Faculty 3 - Mathematics and Computer Science, Universität Bremen, 28359 Bremen, Germany
timlaue@informatik.uni-bremen.de

Keywords: Vision; Recognition; Image Processing; State Estimation; Sensor Fusion; Robot Soccer; Sport Robotics

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.

1 INTRODUCTION

Soon after establishing the RoboCup competition in 1997, the RoboCup Federation proclaimed an ambitious long term goal.

> "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."
>
> *Kitano and Asada (1998)*

Currently, RoboCup competitions take place every year. Within a defined set of different subcompetitions and leagues, incremental steps towards this big goal are made (RoboCup Federation, 2008). 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. (2002), but in many research areas, huge 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:

> 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 rise 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 Sect. 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 Sect. 4. Finally, a sequence of experiments to verify our claim is described in Sect. 5.

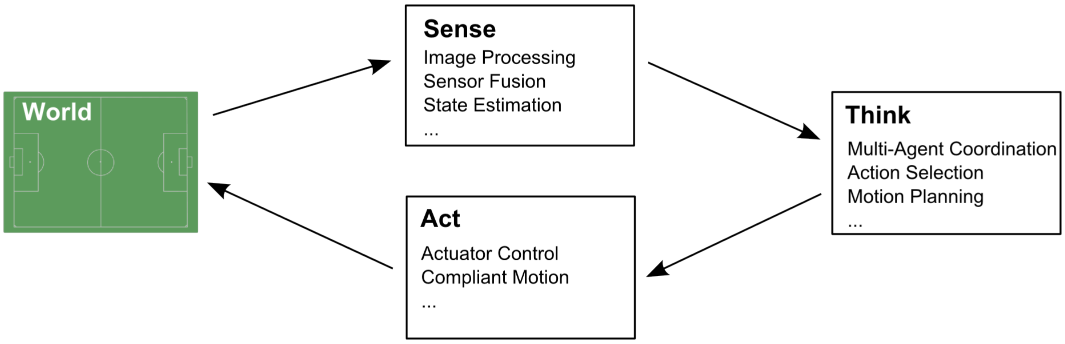

Figure 1: The Sense-Think-Act cycle roughly depicting major tasks for playing soccer with a humanoid robot.

2 CHALLENGES FOR PLAYING SOCCER

The global task of playing soccer consists of several different, interdepending challenges. We roughly categorize them according to the Sense-Think-Act cycle (see Fig.1). This should be considered as a possible architecture for illustration. In the following, the challenges are described in reverted order but with decreasing degree of difficulty.

2.1 Challenges for Actuation

The hugest 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 (Honda Worldwide Site, 2007). This is an impressive result, but still more than five times slower than the top speed of a human soccer player. A similar gap regarding kicking velocity has been pointed out by Haddadin et al. (2007). 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 (Haddadin et al., 2007), not only direct position control but also approaches for compliant motion, such as impedence control, need to be taken into account.

Additionally, the problems of energy efficency and power supply need to be solved. The ASIMO robot for example is, according to Honda Worldwide Site (2007), 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 aspects 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 keepaway soccer serve as a testbed for learning algorithms (Stone et al., 2005). 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 (Kuffner et al., 2002), 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 Kitano and Asada (1998), it is evident that the robots' sensorial capabilities should resemble the human ones. Thus, we could assume to deal with data from cameras and inertial sensors 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 lightyears 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

The objects relevant in a soccer match are the ball, the goals, the line markings and of course the players. Ball, goal and line markings are geometrical features, i.e. circles and lines. There is a large number of algorithms for detecting them in images, from the classical Hough transform (Davies, 2004) up to a range of more elaborate methods (Guru and Shekar, 2004).

Recognizing other players is more challenging. It is particularly difficult because we will probably need not only the general position but the detailed state of motion for close range tackling and to infer the player's action for tactical purposes. Fortunately, people tracking is an important topic in computer vision with a large body of literature (Price, 2008; Ramanan and Forsyth, 2003).

Furthermore, soccer scenes are lightly colored with green lawn and the players wearing colored clothes of high contrast. In the RoboCup competition, this idea is taken to an extreme, where most teams rely on color segmentation on a pixel-per-pixel basis as their primary vision engine. 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) (Lowe, 2004).

Overall, understanding a soccer scene from the player's perspective seems much easier then 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 Beetz et al. (2006, 2007) 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-square 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 (Beetz et al., 2006, Fig. 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 maybe thinner than a single pixel and a person much larger than the camera's field of view, only partially visible. 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 most severe problem is, that from our experience, most of the academic computer vision systems perform on the level of lab demonstrators requiring nicely setup scenes and lighting conditions and usually considerable parameter tweaking. 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, formulating the overall scene understanding and prediction problem as a global likelihood optimization task. This idea in general is not entirely new (Ullman, 1995; Binnig, 2004; Leibe et al., 2007), but we believe it is particularly well suited to this task and also the task is well suited to study this methodology.

4.1 Data-Driven Bottom-Up Processing

Most current vision systems use a data-driven bottom-up approach (Frese et al., 2001; Röfer et al., 2005; Beetz et al., 2007, 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 some background knowledge at its particular 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. Similar the system by Beetz et al. 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, on the low level this knowledge does not exist 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.

4.2 Global Likelihood Optimization

We believe, that much of the brittleness of current vision systems originates from this phenomenon. So our approach for increased robustness is an overall likelihood optimization. In the example above, the variables to be optimized would be the 2D image circle (center, radius) and the 3D position and velocity of the ball over time. The likelihood would be the product of the following likelihoods for all images:

1. a motion model likelihood binding the 3D positions and velocities over time;

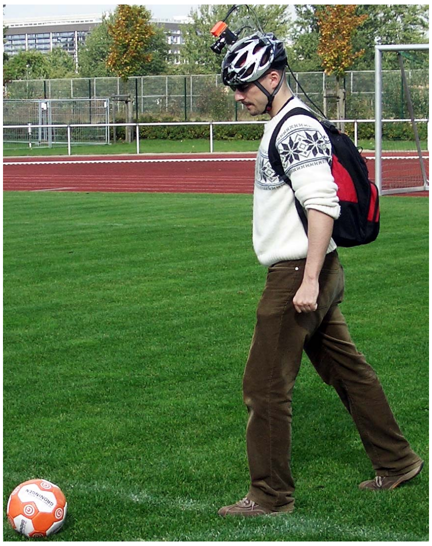

Figure 2: Our proposed experiment: Mount a camera and an inertial sensor on the head of a human soccer player and use them to extract all the information, a humanoid soccer robot would need to take the human's place.

2. a camera model likelihood binding 2D circles to 3D positions;
3. a circle edge likelihood, indicating how much contrast there is in the image along the outline of the hypothesized 2D circle;
4. a circle color likelihood, indicating how well the color inside the 2D circle corresponds to the ball.

The first two factors are Gaussians expressing the models as formulas with uncertainty (Birbach, 2008). The last two look directly into the image for a specific circle, returning a gradual result. In this approach, an indistinct ball would get a lower likelihood in 3. and 4. but this could be compensated by 1. and 2. if it fits well to the context of a flying ball.

The problem is understanding 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 or a particle filter, 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

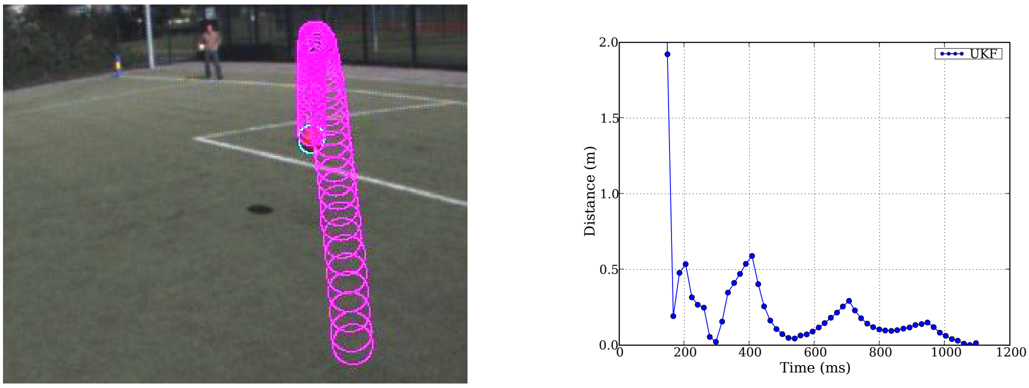

Figure 3: Predicting the trajectory of a flying ball from a moving camera-inertial system. As an initial study, the ball, the lines and the goal corners have been manually extracted from the images. From this data, the trajectory of the ball is predicted (left). The right plot shows the error of the predicted touch down point varying over time. It shows, even though the camera is moving, the prediction is roughly precise enough for interception. See `http://www.sport-robotics.com/icinco/`.

already observed it for 100ms. The first prediction is not delayed at all, because prior to that the ball must have been observed for some time anyway.

Overall, we believe that the approach of a global likelihood optimization directly in the images is an elegant way to greatly increase robustness.

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 (Fig.2).

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 asses, 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 its still not real-time. Overall, it allows to first concentrate on functionality and robustness instead of computational efficiency and technical integration.

We have already conducted a very first experiment (Kurlbaum, 2007; Birbach, 2008), 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 (Fig.3). 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 this kind of studies which deliberately discard essential aspects, such as integration, real-time computation, or autonomy are undervalued by the 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 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 own 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 Fig. 3.

An intriguing idea is to equip the human player with a *tracker-less* motion capture suit (Xsens Technologies B.V., 2007) measuring joint angles. Apart from providing the kinematic information discussed above, it also provides the trajectory of both feet. If the player hits the ball, one can compare the predicted ball trajectory with the real foot trajectory and evaluate the precision. This is important since ground truth is not available.

5.3 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 rather 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, we are 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.

REFERENCES

Beetz, M., Gedikli, S., Bandouch, J., Kirchlechner, B., v. Hoyningen-Huene, N., and Perzylo, A. (2007). Visually tracking football games based on tv broadcasts. *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India.*

Beetz, M., v. Hoyningen-Huene, N., Bandouch, J., Kirchlechner, B., Gedikli, S., and Maldonado, A. (2006). Camerabased observation of football games for analyzing multiagent activities. In *International Conference on Autonomous Agents.*

Binnig, G. (2004). Cellenger automated high content analysis of biomedical imagery.

Birbach, O. (2008). Accuracy analysis of camera-inertial sensor based ball trajectory prediction. Master's thesis, Universität Bremen, Mathematik und Informatik.

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

Davies, E. R. (2004). *Machine Vision. Theory, Algorithms, Practicalities.* Morgan Kauffmann.

Frese, U., Bäuml, B., Haidacher, S., Schreiber, G., Schaefer, I., Hähnle, M., and Hirzinger, G. (2001). Off-the-shelf vision for a robotic ball catcher. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Maui*, pages 1623 – 1629.

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

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

Honda Worldwide Site (2007). Honda World Wide — Asimo. `http://world.honda.com/ASIMO/`.

Kitano, H. and Asada, M. (1998). 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*, pages 419–424.

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

Kurlbaum, J. (2007). Verfolgung von ballflugbahnen mit einem frei beweglichen kamera-inertialsensor. Master's thesis, Universität Bremen, Mathematik und Informatik.

Leibe, B., Cornelis, N., Cornelis, K., and Gool, L. V. (2007). Dynamic 3D Scene Analysis from a Moving Vehicle. In *IEEE Conference on Computer Vision and Pattern Recognition.*

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91 – 110.

Price, K. (2008). The annotated computer vision bibliography. `http://www.visionbib.com/`.

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

RoboCup Federation (2008). RoboCup Official Site. `http://www.robocup.org`.

Röfer, T. et al. (2005). GermanTeam RoboCup 2005. `http://www.germanteam.org/GT2005.pdf`.

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

Ullman, S. (1995). Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, 5(1):1–11.

Xsens Technologies B.V. (2007). *Moven, Inertial Motion Capture, Product Leaflet.* XSens Technologies.