A Pragmatic Global Vision System for Educational Robotics

John Anderson and **Jacky Baltes**
Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, R3T 2N2 Canada
Email: andersj,jacky@cs.umanitoba.ca

Abstract

This paper advocates the use of global vision as a tool for increasing the effectiveness of robotics education, and describes the design and functionality of advanced global vision systems used in our own programs. Our experiences with using global vision as a basis for teaching robotics and AI have led us to use this as a standard method for teaching undergraduates. Our recent vision systems (DORAEMON and ERGO) have consistently been improved to perform accurately and robustly over a wide range of applications. DORAEMON uses a sophisticated camera calibration method and colour model to remove the need for an overhead view of the world. ERGO minimized the use of colour information to provide more robust object recognition under varying lighting scenarios. Most recently, these video servers have been used by undergraduates to develop autonomous robots for a mixed virtual/physical world.

Introduction: Global Vision in an Educational Environment

Perception is the most difficult element to present realistically when educating students in hands-on robotics. While elements of mechanics and control can be nicely scaled down by using simplified robotic software, and simple IR based platforms such as remote controlled toys can easily be adapted for students, limiting perception severely limits the applications that can be developed. It is certainly possible to develop interesting robotics projects using simple perceptual devices: a single sonar, for example, can be used to avoid obstacles directly in front of the robot, while a light sensor can be used to give a basic goal for a robot. However, the sophistication of resulting applications will always be limited without vision. The problem with vision is that it generates an enormous amount of data, and requires sophisticated algorithms to deal with even relatively simple issues such recognizing basic shapes, let alone judging distances, dealing with noise, and tracking objects over time. This presents two significant challenges to those wishing to use it as a basis for undergraduate or high-school robotics education. First, handling voluminous information and sophisticated processing on inexpensive robots, and second, ensuring that vision can be employed by students without overwhelming them in complexity. Another set of challenges also arises from the standpoint of managing such an educational program, namely setting up and calibrating a vision system, together with ongoing maintenance and the effort required to adapt it to different problem-solving environments.

This paper presents our approach to employing vision in educational robotics, together with recent work on a system that can be used for undergraduate and high-school robotics courses in addition to advanced research. Our approach begins by accepting that sophisticated visual processing is beyond the capabilities of lower-level students using inexpensive robots. Currently available consumer platforms (e.g. PDAs) achieve only a very low frame rate when visual processing is run locally, so many popular applications such as robotic soccer would be out of the question.

While we also work with local vision robots in our research, the problem from a robotics education standpoint is that many basic AI topics (e.g. planning, multi-agent cooperation) require good localization and mapping for successful student implementations. Even though a lot of progress has been made in recent years in localization and mapping (Thrun *et al.* 2001), these areas are still open problems when using visual feedback from poor quality cameras and robots with noisy ego-motion sensors. Because of this, robots relying only on local vision for sensing are often limited to more reactive behaviours such as chasing a ball.

Common student lab equipment such as Lego MindStorms are also too weak to do any local visual processing. Even if inexpensive, high-performance mobile platforms become available in the near term, from a student standpoint, the sophisticated algorithms necessary to support local vision are out of the question. Frameworks for local vision are also not simple to adapt to new environments.

Like many robotic soccer leagues (e.g. the RoboCup F-180 and the previous E-League), we advocate the simplicity of using a global vision approach, where a single third-party view is provided to all members of a robot team, analogous to the view of a commentator in a soccer game. While most robotics leagues take this route to limit local processing requirements, we find the real advantage in employing global vision is that it allows the introduction of basic computer vision concepts at a far less overwhelming level. Students thus get to see some of the issues involved in employing such systems in the real world, while drastically lowering the information load that would be required by local vision. The end result is that we can have systems that employ vision, *using tools that are simple enough that students can eventually calibrate them and modify them for new domains themselves.* Students can learn the rudiments of computer vision and benefit from having more interesting robotic domains to work in.

Global vision shares many of the problems associated with local vision. Objects of interest must be identified and tracked, which requires dealing with changes in appearance due to lighting variation and perspective. Since objects may not be identifiable in every frame, tracking objects across different frames is often necessary even if the objects are not mobile. The problem of interpreting objects that are juxtaposed as one large object rather than several distinct objects, and other problems related to the placement and motion of objects in the environment, are also common.

In domains such as robotic soccer, where pragmatic real-time global vision is large part of the application, many of the more difficult problems associated with global vision have been dealt with through the introduction of assumptions that greatly simplify the situation. The cost of such assumptions is that of generality: systems can only operate where the assumptions they rely upon hold. For example, global vision systems for robotic soccer (e.g. (Bruce & Veloso 2003; Browning *et al.* 2002; Simon, Behnke, & Rojas 2001; Ball, Wyeth, & Nuske 2004)) generally require a camera to be mounted perfectly overhead in order to provide a simple geometric perspective (and thus ensure that any object is the same size in the image no matter where in the field of view it appears), simplify tracking, and eliminate complex problems such as occlusion between agents. If a camera cannot be placed perfectly overhead, these systems cannot be used. Such systems also typically recognize individuals by arrangements of coloured patches, where the colours (for the patches and other items such as the ball) must be pre-defined, necessitating constant camera recalibration as lighting changes. Such systems can thus only operate in environments where lighting remains relatively consistent.

These systems bear little resemblance to human vision: children playing with remote-controlled toys, for example, do not have to climb to the ceiling and look down from overhead to understand their motions. Human vision does not require significant restrictions on lighting consistency, nor any specialized markings on objects to be tracked. In order to advance the state of the art in robotics and artificial intelligence, we must begin to make vision systems more generally intelligent. The most obvious first steps in this direction are considering the assumptions necessary to make current a global vision system fuctional, and then to find ways of removing these while still keeping the system accessible to the understanding of students.

Our approach to real time computer vision arises from a desire to produce a more generally intelligent approach to global vision for teams of robots, not only for technological advancement, but to make the system simpler to maintain and use and therefore more applicable to an educational environment. For example, a system that does not assume that a camera has a perfect overhead mount is not only more generally useful, but requires less set-up time in that a perfect overhead mount does not need to be made. Similarly, an approach that can function in a wide range of lighting conditions saves the time and expense of providing specialized lighting for a robotic domain or recalibrating the system each time the lighting changes. In conjunction with teaching robotics to graduate and undergraduate students over the past six years, we have developed a series of real-time global vision systems. While designed for the robotic soccer domain, these are also generally useful anywhere global vision can be used. These systems have been used in RoboCup and FIRA robotic soccer competitions by ourselves and other teams, and have also been employed in such applications as robotic education and imitation learning. More importantly from the standpoint of this paper, they are used routinely by students in introductory robotics classes. All are open source, and can be easily obtained for educational use or as a basis for further research work (Baltes & Anderson 2006).

The remainder of this paper deals with the specific design points made in our global vision systems to make them more generally applicable as well as practical for student use. We begin by examining the steps necessary to deal with a more general camera position, tracking objects when the camera is not perfectly overhead. We then turn to dealing with simplifying the objects being tracked by removing assumptions about color recognition, in order to assumptions about the objects being tracked, in in order to produce a vision system that does not rely on colour calibration.

Doraemon: Real-Time Object Tracking without an Overhead Camera

DORAEMON (Anderson & Baltes 2002; Baltes 2002) was the first of our systems to allow objects to be tracked from an oblique camera angle. The system acts as a server, taking frames from a camera, and producing a description of the objects tracked in frames at regular intervals, sending these over a network to clients (agents controlling robots, for example) subscribing to this information stream. Figure 1 is a sample visual frame used as input to DORAEMON to illustrate the problems involved in interpreting visual images without using a perfect overhead viewpoint. The image is disproportionate in height because it is one raw field from the interlaced video stream provided by the camera. It is easy to see that features are hard to extract, in part because the shape of coloured patches are elongated by the visual perspective, and in part because colour is not consistent across the entire image.

In order to be able to track images from an oblique angle, a calibration must be provided that allows an appropriate translation from a particular pixel in a visual frame to a coordinate system in the real world. The calibration process used by DORAEMON, described in detail in (Anderson & Baltes 2002), utilizes the well-established Tsai camera calibration (Tsai 1986), which can compute a camera calibration from a single image. We have used the Tsai calibration since 1998 with good results on mono-plane calibrations. This

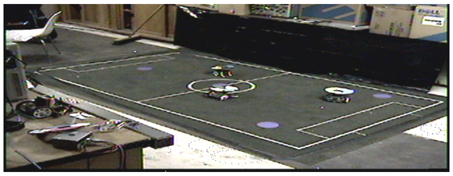

Figure 1: A sample visual frame taken from an oblique angle.

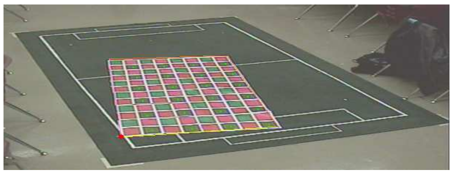

Figure 2: Tsai Camera Calibration used in Doraemon

method computes six external parameters (based on camera position) and six internal parameters using a set of calibration points from an image with known world coordinates. This requires a set of coordinates to be imposed on the world via a sample visual image (a calibration carpet with a repetitive grid). Even using an oblique view of the playing field, the calibration results in object errors of less than 1 cm.

Objects in DORAEMON are identified by the size and arrangement of coloured patches. The simplest objects may be simply a single coloured area of a given size - e.g. a ball might be described as an orange item 5cm in diameter. More sophisticated items (e.g. individual robots) are identified using unique arrangement of coloured patches on the top surface, as shown in Figure 1 (e.g. a blue patch for the front of all robots on one team, with an arrangement of other colours uniquely identifying each team member). The system is thus heavily dependent on accurate colour models. DORAEMON uses a sophisticated 12 parameter color model that is based on red (R), green (G), and blue (B) channels as well as the difference channels red-green (R-G), red-blue (R-B), and green-blue (G-B). The channel differences are less sensitive to lighting variations than the raw channels, and allow more robust colour recognition than the raw channels alone. While there are other models that are less sensitive to brightness, (for example, HSI), this approach attempts to balance sensitivity with computational resources. The channel differences are similar to the hue values used in HSI, for example, while this model is less computationally expensive.

Each frame is colour thresholded and the recognized patches are matched against the size and configuration information provided. Not every object will be recognized in every frame (e.g., because of fluctuations in lighting). To compensate for this, the locations of recognized objects in previous frames are used both to infer likely positions in future frames and to calculate the speed and orientation of motion of tracked objects.

Occlusion in robotic soccer is normally not an issue for tracking robots, even with an oblique camera, since the markers are on top of the robots and are thus the highest points on the field. Occlusion certainly happens when tracking the ball, however, and is also possible in any tracking scenario where obstacles on the field could be taller than robots. There is also the possibility that robots may abut one another, presenting a display of coloured patches that is similar to a different robot altogether, or presented in such a way that no one robot is easily recognizable. These situations are dealt with by tracking objects over time as well - an object may be lost temporarily as it passes behind an obstacle, or may be more momentarily unrecognized due to abutting other tracked objects - because objects are intended to be in motion, such losses will be momentary as new information allows them to be disambiguated.

```
7 6188 0.000290976 ; #defined objects, frame#, time diff
1605.82 -708.394 1321.44 ; x, y, z coordinates of camera
2 spot1 Found 1232.5 416.374 0 0 0 0 ; object information
2 spot2 Found 1559.22 417.359 0 0 0 0
2 spot3 Found 1260.55 812.189 0 0 0 0
2 spot4 Found 902.726 1002.43 0 0 0 0
2 spot5 Found 746.045 735.631 0 0 0 0
1 ball1 Found 1677.99 1205.55 50 0 -2.75769 1.19908
0 car54 Found 1783.53 873.531 100 2.63944 1.47684 -6.49056
```

Figure 3: A sample message from Doraemon

DORAEMON transmits information about tracked objects (position, orientation, velocity) in ASCII over ethernet to any client interested in receiving it. A sample message is shown in Figure 3.

The first line of each message contains the number of objects that video server is configured to track, followed by the video frame number and time difference in seconds between this message and the previous one. The next line contains the x, y, and z coordinates of the camera, and following this is a line for each object being tracked. Each of those lines consists of a numeric object class (e.g. a ball, robot, etc.), the unique defined identifier for the object, whether the object was located in the current frame or not, the x, y, and z coordinates of the object, the orientation of the object in radians, and the velocity of the object in mm/second in the x and y dimensions.

Doraemon takes several steps beyond global vision systems that maintain a fixed overhead camera in terms of being able to deal with the real world. It is quick to calibrate and simple to recalibrate when this is necessary (e.g. due to camera shift or changing lighting during use). However, DORAEMON is heavily dependent on good colour models, something that is not easy to maintain consistently over time in real-world domains without recalibration, and relies on a fairly naive model for dealing with occlusion. These dependencies led to the development outlined in the next section.

Ergo: Removing Dependence on Predefined Colours

The reliance on colour thresholding by both DORAEMON and related systems places some severe restrictions on the possible applications of a global vision system. Not only are lighting variations a problem, but the colours themselves must be chosen so that there is enough separation between them to allow them to be distinguished across the entire field of play, and the quality of the camera used is also a major issue. In practice, even with the extra colour channels employed by DORAEMON tracking is practically limited to around 6 different colours by these restrictions.

To increase the applicability of global vision to a broader array of real-world tasks, as well as to increase the robustness of the system in robotic soccer, we focussed on two major changes in approach: the use of motion detection to focus on areas of interest in the field, and different methods of marking objects that deemphasize the use of colour. These and other extensions resulted in the next generation of our global vision system, known as ERGO (Furgale, Anderson, & Baltes 2005).

One additional pragmatic step was also necessary in ERGO in order to attain a comparable frame rate as that employed in the original DORAEMON: the resolution of interpolated images was decreased, in order that interpolation did not inordinately slow down visual analysis. The result of this introduced an additional challenge, in that a typical 5cm soccer ball would now occupy only a 1-4 pixel range in the reduced resolution, allowing a ball to easily be interpreted as noise (Figure 4).

Figure 4: Captured field and corresponding low-resolution interpolated image in Ergo. Note that the ball is easily visible in the former image, but blends with noise on the field lines in the latter.

Rather than performing direct colour thresholding of camera images, ERGO thresholds for motion across pixels in each frame compared to a background image. An adaptation of $\Sigma\Delta$ background estimation (Manzanera & Richefeu 2004) is used, which provides a computationally inexpensive means of recursively estimating the average color and variance of each pixel in a camera image.

Detecting motion involves setting a threshold above which variation across pixels will be considered to be motion. In experimenting with this, it was found that increasing a global threshold enough that all noise would be eliminated also had the effect of eliminating any object of the size of a typical robotic soccer ball, since the size of such an object in the image (¡=4 pixels) is easily interpreted as noise. To deal with this, a means was required to consider variation more locally and eliminate noise, while still being able to pick up the motion of small objects, and so a combination of local and global thresholding was employed. A threshold is set for each pixel by examining the variance for each pixel in the background image, then apply a convolution in order to consider a pixel's variance across its 9-pixel neighbourhood. This local threshold is then scaled by a global threshold. To detect motion, each incoming image has its sum-squared error calculated across all pixels against the background image, the same convolution is applied to the result, and each value is compared to its corresponding pre-computed threshold. The use of the convolution has the effect of blending motion in small areas to eliminate noise, while making the movement of small objects such as the ball more obvious by also considering small changes in neighbouring pixels. The individual motion pixels are then merged together into regions.

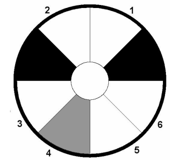

Figure 5: A new approach to labeling objects for tracking (Furgale, Anderson, & Baltes 2005): fixed black areas allow orientation to be recognized, while white and non-white values in locations 1-6 represent identity

ERGOalso introduced a new pattern representation. The two basic requirements of a representation are the determination of identity and orientation (since the remaining item of interest, velocity, can be obtained through knowing these over time). Previous research (Bruce & Veloso 2003) has shown that asymmetrical patterns can be used to allow a range of objects to be identified with fewer colours, and these ideas were extended to develop a representation and associated matching mechanism for tracking objects while minimizing the need for predefined colours.

The marking approach designed for Ergo divides the marker for a robot (or similar moving object) into a circular series of wedges (Figure 5). Two black wedges are the same on all robots, allowing a tracking algorithm to determine the labeled object's orientation. The remaining six wedges are marked with white and non-white (i.e. *any* colour other than white or black) to allow the determination of identity. Marking only two of these segments would allow up to twenty-one individuals to be identified uniquely (the centre is left open for a possible team identifier if desired).

An associated algorithm for identifying objects assumes that such a marking system is in use, and begins with a set of hypotheses of objects of interest, based on the regions of the camera image that have been flagged as motion. The original image is reinterpolated with a higher resolution in (only) three concentric circular strips of pixels (each 64 pixels long) around the centre of each region of motion. The mean is taken across these, resulting in a single array of 64 elements, providing an encoding for that region of motion that can be matched against the labeled pattern described above. To be able to match the pattern in this strip, two boundaries must be determined in this strip: the boundary between black and the marker that is neither black nor white, and the boundary between that and white. These boundaries are determined using a histogram of intensity values produced as part of the reinterpolation.

Once these thresholds are available, the identification algorithm begins by looking for the two black regions, and the average of the centre between these is the orientation. These wedges also provide the plane on which the pattern, and based on that plane the recorded centre of the object is refined. The remaining parts of the interpolated strip are then partitioned relative to the black wedges and the identification pattern can then be determined by counting the number of white wedges and the number of wedges that are neither white nor black.

This identification algorithm is very effective and computationally minimal, but is complicated in application by two factors. First, the list of regions of motion may be significantly larger than the number of objects to be tracked (due to extraneous movement by other objects, for example): large enough that this algorithm cannot process them all in real time in the data directed manner that would be ideal. Second, successful identification of an object relies on an accurate centre point. If two or more moving objects appear in close proximity to one another (or even partly occlude one another), motion analysis will view this as one large region of motion, with a centre that will not be helpful in identifying anything. This algorithm thus needs to be applied in a more goal-directed manner, and have some means of dealing with clumps of objects.

ERGO deals with these problems by tracking objects across images and predicting their next position, which provides for a goal directed application of this algorithm. Some objects may thus be found very quickly, since their centre point will be predicted and can easily be confirmed using the identification algorithm. The area in the image occupied by object recognized during this phase is masked during motion analysis. This masking serves two purposes: it produces no hypothesis, since the object has already been dealt with, but it also may serve to remove one of a group of objects that may appear together in a moving region. Masking the area will then leave a smaller region and a smaller number of grouped objects (possibly only one, which can then be handled as any other object would).

There are realistically two possibilities for the remaining objects: a region of motion is outside the predicted area for the object, or it is part of a clump of objects occupying a larger region. To deal with the former, ERGO examines the sizes of all unexplained regions of motion, and if it is a size that could suitably match an object of interest, it is passed to the identification algorithm. In the case of multiple objects occupying the same space, the regions of interest will be those that are too large for any one object. If any of these regions were to contain more than one object, at least one recognizable object will be touching the edge of the region, and so the edge is where recognition efforts are focussed.

Not every object is large enough to be labeled using the scheme shown in Figure 5, nor do all objects need an encoding to uniquely identify them. In robotic soccer, for example, the ball is physically unique, and its nature does not require a pattern for identification. The use of motion tracking to distinguish an element as small as the ball has already been described. In frames where this motion tracking does not allow the ball to be found, the ball's location is predicted from the previous frame, and an area eight times the ball's size is scanned for regions of the correct size and dimension after colour thresholding. Colour thresholding here is simply used to distinguish regions at all given that motion detection has failed, and no predefined colours are employed.

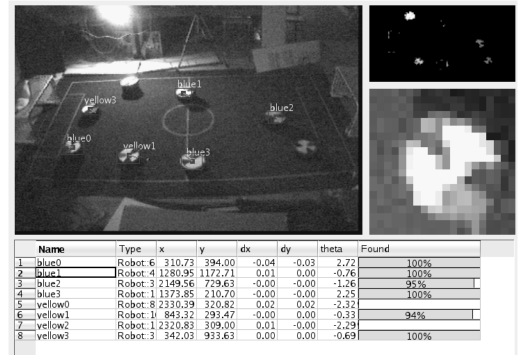

Figure 6: Using Ergo under very poor lighting conditions (Furgale, Anderson, & Baltes 2005)

These techniques allow ERGO to perform well under very challenging conditions. Figure 6 illustrates an extreme example, with lighting positioned across the viewing area, causing a wide disparity in brightness, and significant shadowing. Motion tracking is shown in the upper right, and the system output in the bottom of the image. All robots are identified except for one completely hidden in shadow, and the other in complete glare from the lighting source.

ERGO has gone a long way in making a global vision system more easily employed by and understood by students, in that it has both removed the need for a fixed overhead camera as well as any predefined colours, and thus can operate across a much broader range of condition s than previous systems. There are still assumptions it operates under, the largest being that a pattern can be used to consistently identify objects that need to be tracked.

Discussion

This paper has reviewed some of the issues involved in creating pragmatic global vision systems that are useable by students. We have discussed the assumptions on which traditional systems are based, pointed out how these differ with the observed abilities of human vision, and described how these assumptions limit the applicability and generality of existing systems. We then described techniques that allow some of these assumptions to be discarded, and the embodiment of these techniques in our production global vision systems, DORAEMONand ERGO.

In our undergraduate robotics course, students begin by using these vision systems on the first day of classes. The students are shown the principles behind the software they are using, and know enough in a single laboratory to understand how to calibrate the system and maintain it under varying lighting conditions as they do their work. As illustrated in this paper, the systems themselves go a long way toward making this task simple. Unlike using local vision, the students do not need to understand visual principles to get robots to operate with this system: at first they need to understand only enough to calibrate the system and to understand the output provided so that visual information can be used in their programs. As they work with the system, however, they understand more and more about issues such as occlusion and tracking objects over time by seeing the situations in which the system performs better in those in which specific problems occur. This allows the students to proceed through the remainder of a robotics course without avoiding issues in vision, while also not allowing these issues to become overwhelming. In addition to educational work, our systems have been in use every year by a number of teams from around the world in the F-180 (small-size) league at RoboCup.

We are currently extending the concept of global vision to include mixed reality environments. Since recent IR platforms have become smaller and smaller (we currently employ 2” IR tank toys in our undergraduate classes), and LCD technology cheaper, it has become feasible to employ horizontally mounted LCD panels as field surfaces, allowing a virtual environment to interact with a physical one. Figure 7 shows a 42” television supporting a mixed-reality Pac Man game developed by our undergraduate students, where the ghosts and Pac-man are real robots based on toy tanks. The television displays the virtual world (labyrinth walls, power-ups, and other game elements) with which the robots interact. This setup is small enough to be easily portable and can employ off-the-shelf IR emitters rather than the custom-made emitters that would be required for larger environments. A field of this size, along with the vision systems described here, would allow a robotics group to use entirely consumer-grade components to learn and explore.

Acknowledgement

The authors would like to thank all of the students in our laboratories in Manitoba and Auckland that have worked on global vision software in the past, including Paul Furgale and Ben Vosseteig, as well as all of the students in our robotics courses that stress-test this software and continue to drive ongoing development.

Figure 7: Ergo in use with a Pac Man mixed-reality environment

References

Anderson, J., and Baltes, J. 2002. Doraemon user’s manual. http://robocup-video.sourceforge.net.

Ball, D.; Wyeth, G.; and Nuske, S. 2004. A global vision system for a robot soccer team. In *Proceedings of the 2004 Australasian Conference on Robotics and Automation (ACRA)*.

Baltes, J., and Anderson, J. 2006. Doraemon, Ergo, and related global vision systems. http://robocup-video.sourceforge.net.

Baltes, J. 2002. Doraemon: Object orientation and id without additional markers. In *2nd IFAC Conference on Mechatronic Systems*, 845–850. Berkeley, CA: American Automatic Control Council.

Browning, B.; Bowling, M.; Bruce, J.; Balasubramanian, R.; and Veloso, M. 2002. Cm-dragons01 - vision-based motion tracking and heterogenous robots. In Birk, A.; Coradeschi, S.; and Tadokoro, S., eds., *RoboCup-2001: Robot Soccer World Cup V*. Berlin: Springer-Verlag. 567–570.

Bruce, J., and Veloso, M. 2003. Fast and accurate vision-based pattern detection and identification. In *Proceedings of Proceedings of the IEEE International Conference on Robotics and Automation (ICRA-03)*, 567–570.

Furgale, P.; Anderson, J.; and Baltes, J. 2005. Real-time vision-based pattern tracking without predefined colors. In *Proceedings of the Third International Conference on Computational Intelligence, Robotics, and Autonomous Systems (CIRAS)*.

Manzanera, A., and Richefeu, J. 2004. A robust and computationally efficient motion detection algorithm based on sigma-delta background estimation. In *Proceedings of the 4th Indian Conference on Computer Vision, Graphics and Image Processing*, 46–51.

Simon, M.; Behnke, S.; and Rojas, R. 2001. Robust real time color tracking. In Stone, P.; Balch, T.; and Kraetszchmar, G., eds., *RoboCup-2000: Robot Soccer World Cup IV*. Berlin: Springer Verlag. 239–248.

Thrun, S.; Fox, D.; Burgard, W.; and Dellaert, F. 2001. Robust monte carlo localization for mobile robots. *Artificial Intelligence* 128(1-2):99–141.

Tsai, R. Y. 1986. An efficient and accurate camera calibration technique for 3d machine vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 364–374. Miami Beach, FL: IEEE Computer Society Press.