Real-Time Visual Ground-Truth System for Indoor Robotic Applications

André Dias, Jose Almeida, Alfredo Martins, and Eduardo Silva

INESC TEC - INESC Technology and Science
ISEP/IPP - School of Engineering,
Porto, Portugal
{adias,jma,aom,eaps}@lsa.isep.ipp.pt

Abstract. The robotics community is concerned with the ability to infer and compare the results from researchers in areas such as vision perception and multi-robot cooperative behavior. To accomplish that task, this paper proposes a real-time indoor visual ground truth system capable of providing accuracy with at least more magnitude than the precision of the algorithm to be evaluated. A multi-camera architecture is proposed under the ROS (Robot Operating System) framework to estimate the 3D position of objects and the implementation and results were contextualized to the Robocup Middle Size League scenario.

1 Introduction

The robotics community is addressing research areas such as vision perception [1][9], multi-robot cooperative behavior[18] and localization[11]. In all these areas, the ability to evaluate the results is a crucial step to assess the quality of the research presented by the authors. This concern with obtaining better results in robotics was expressed by the European Robotics Research Network (EURON)[12] and cited by [13]. A ground truth system based on vision is frequently used in nearly all robotic labs in response to this uneasiness and the solutions available can be classified in two major categories: commercial and self-developed solutions. In the commercial category, some of the possible solutions are the Vicon[9][13] and Optitrac[2], which provide enough accuracy for most robotics applications, but with higher costs. More than the price itself, these systems have limitations that could make them impracticable in some scenarios, such as the pulsed infra-red light limit range to the reflective markers and the fact that some objects to track cannot carry the markers due to the geometry (for example, a ball). Some examples of the implementation of commercial solutions are in the $6D$ data collection for the biped humanoid robot Aldebaran Nao from the Robocup Standard Platform League (SPL)[13] and the ground truth to benchmark visual pose estimation algorithms implemented in a standard quadrotor platform[9].

Regarding the self-developed solution, there are a wide range of implementations that were written from the scratch and are, with some exceptions, rarely usable by anyone other than the original programmer. To enumerate some of them,

J.M. Sanches, L. Micó, and J.S. Cardoso (Eds.): IbPRIA 2013, LNCS 7887, pp. 304–313, 2013.

there is the XVision[16] and TrackIt[6]. The exceptions are the SwisTrack[10], which is a generic and flexible tool for tracking robots and insects in flocks and formation control experiments with robots such as the Khepera and e-puck; the SSL-Vision[19] system which is used by the Robocup Small Size League to estimate the ground truth position of the robots and ball in 2D, extended in [20] for the Middle Size League and to the Aldebaran Nao by [15]. It is important to stress that all solutions available are limited to 2D tracking of known markers using a vision system located above the tracking area. More recently, a real-time solution was developed using depth information (3D tracking) obtained with a Microsoft Kinect RGB-D Sensor to track the robots from the SPL[7]. The main limitation of this approach is the scenario constraints, a max area of $6m^2$ due to the Kinect sensor range limit of $3.5m$, the field of view (57° horizontally and 43° vertically) and camera resolution (640×480 pixels).

This work proposes a vision-based Ground Truth capable of performing 3D tracking of multiple targets and overcoming the previously mentioned limitations. The proposed Visual 3D Ground Truth system was developed for indoor scenarios with the following goals: providing an easy method to obtain the required calibration parameters, which is able to ensure a good accuracy without special markers, supporting an open-source implementation with low-cost hardware and providing a framework capable of receiving add-ons from the robotics community.

The visual ground truth evaluation scenario will be contextualized to the Robocup Middle Size League (MSL) as this is an important testbed for robotics applications. Here, a team of five robots play together sharing and combining information in order to present cooperative behaviors and to achieve a final goal, which is playing soccer. In this specific and complex scenario, there are two works[21][8] in the state of art that were considered references to characterize which should be the requirements for the proposed system applicable to the envisioned Robocup scenario. These references were chosen as they are visual tools (besides being visual ground truth systems) for the MSL developed for monitoring data logs, providing state estimation (localization and target estimation) and path planning for each robot over the video recording of the game.

This paper is outlined as follows: Section II presents the proposal architecture and methodology for target tracking objects by color and based on morphological characteristics, the calibration procedure and the stereo triangulation method to estimate the 3D position. Section III provides the results under the MSL scenario for tracking two teams of robots during a game and the ball trajectory for two types of kicking. Section IV provides the conclusion and outlines future work topics.

2 Proposal Architecture and Methodology

The proposed architecture is organized in three major layers: Data Acquisition, Image Processing and 3D Stereo Triangulation (see figure 1). The layers were implemented under the ROS (Robotic Operating System) framework as it is capable of supporting multiple camera drivers, and it has a large impact on the

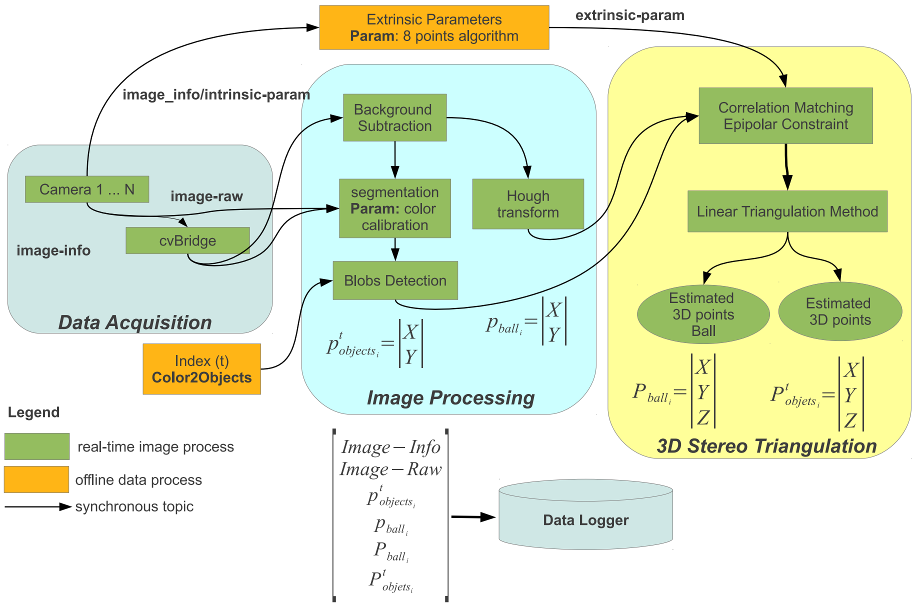

Extrinsic Parameters
Param: 8 points algorithm
extrinsic-param
image_info/intrinsic-param
Camera 1 ... N
image-raw
cvBridge
image-info
Data Acquisition
Index (t)
Color2Objects
Background Subtraction
segmentation
Param: color calibration
Hough transform
Blobs Detection
$p^{t}_{objects_i} = \begin{vmatrix} X \\ Y \end{vmatrix}$
$p_{ball_i} = \begin{vmatrix} X \\ Y \end{vmatrix}$
Image Processing
Correlation Matching
Epipolar Constraint
Linear Triangulation Method
Estimated 3D points Ball
Estimated 3D points
$P_{ball_i} = \begin{vmatrix} X \\ Y \\ Z \end{vmatrix}$
$P^{t}_{objets_i} = \begin{vmatrix} X \\ Y \\ Z \end{vmatrix}$
3D Stereo Triangulation
Legend
real-time image process
offline data process
synchronous topic
$\begin{bmatrix} Image-Info \\ Image-Raw \\ p^{t}_{objects_i} \\ p_{ball_i} \\ P_{ball_i} \\ P^{t}_{objets_i} \end{bmatrix}$
Data Logger

Fig. 1. Proposal Architecture for Multi-Camera Visual Ground Truth System

robotics community, providing a middleware for inter-process communication. Furthermore, it is an open and modular framework suitable for further feature development and integration.

Although the architecture supports multi-cameras, the three layers contextualized to a visual ground truth system will be described in detail. These are composed of two gigabit ethernet cameras in a stereo baseline (see figure 2) and applied to the Middle Size League (MSL) scenario.

2.1 Real-Time Data Acquisition System

The Data Acquisition layer is responsible for the camera acquisition hardware abstraction through a generic image structure provided by the ROS. With this abstraction, the researchers can integrate new features, such as camera drivers inside the proposed architecture and ensure the integration with the other layers available. For the proposal implementation scenario, the ground truth cameras were positioned looking towards the testbed with a baseline of $\sim$13 meters and connected to a machine with Core Intel(R) Core(TM) i7 CPU 2.8GHz, 4GB RAM, running a Linux operating system and connected to an external trigger device to provide a snapshot from both cameras at the same instant. The cameras used were the Basler acA1300-30gc at 15 frames per second (fps), each with a pixel resolution of 1278 × 958. Considering the application scenario, it is important to provide an accurate time synchronization between the ground truth system and the robots in order to correlate the prioceptive and the eteroceptive information acquired by each system. Three clock synchronization protocols were evaluated: the NTP (Network Time Protocol)[4], the PTP (Precision Time Protocol)[22] and the Chrony[3]. The chrony was chosen for the comparison as it presents a steady state with low offset $< 2.6\mu s$, after the system's reboot takes $0.2s$ to stabilize the offset and even when operating under sudden changes (wireless link), the system remains stable.

2.2 Image Processing

The vision system setup is in a large field of view (FOV) and, as a consequence, it becomes exposed to constraints that could lead to errors in the scene analysis (see figure 2). The constraints are: light variations caused by the existing windows, color objects inside the FOV of the system and the possibility of having people moving during a dataset. Adding to these constraints, there are challenges imposed by the Robocup MSL with the color objects (arbitrary ball) which tend to disappear as the competition evolves in order to increase the difficulty posed to the vision algorithms.

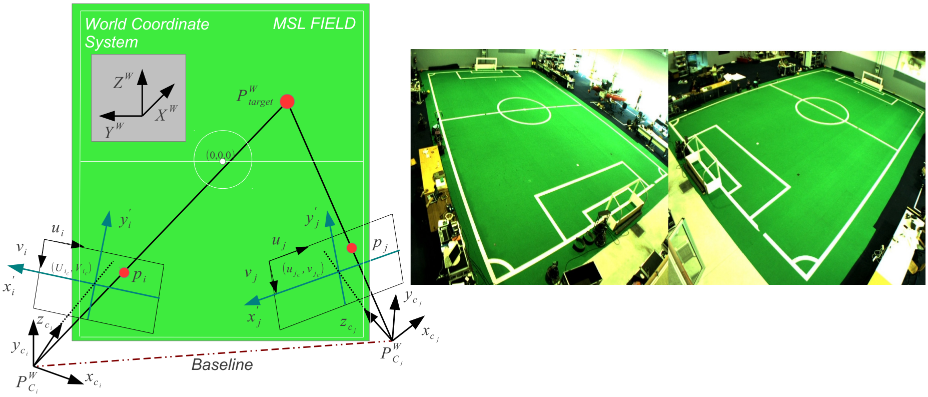
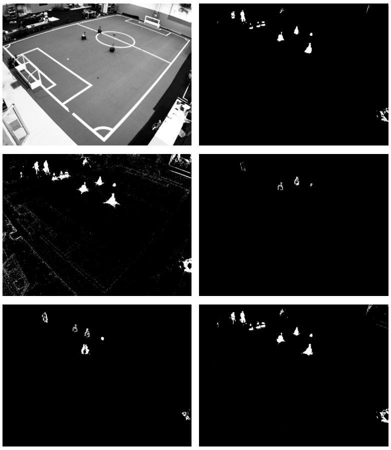


Fig. 2. Left: Geometrical Model for Stereo Triangulation. **Right:** Left and Right camera field of view for the proposal implementation scenario.

In response to all these constraints, a scene analysis with the following steps was implemented:

- Apply background subtraction (BGS) over the image. The objects to be detected should not be present in the field. Then create a statistical model. With this model, the objects can be detected by evidencing the parts of the image that do not fit the model. The figure 3 presents the output from the methods implemented to evaluate the quality of tracking and the computational overload of each method (see table from figure 3). From the snapshot output results presented in figure 3 the method that will be applied to perform real-time tracking of robots and ball will be the BGS Gaussian Average.



Background Method	time (s)
Adaptive Median[14]	0.0487
Adaptive Gaussian Mixure Model (GMM)[24]	0.1238
Temporal Median[14]	0.0660
Gaussian Average[14]	0.0485
Improved Adaptive Gaussian Mixture Model[24]	0.0810

Fig. 3. Left: Top-Left: Original Image. Top-Right: Adaptive Median. Middle-Left: Gaussian Mixure Model (GMM). Middle-Right: Temporal Mean. Bottom-Left: Average GMM. Bottom-Right: Improved Adaptive GMM. **Right:** Results with different Background Subtraction Methods with a fix value of 30 learning frames.

- With the action areas from the BGS, there are two concurrently running blocks, one for detecting the ball and another for detecting the robot teammate. The ball detection is conducted by taking advantage of the morphological characteristics of the ball (round shape) and applying the Hough transform algorithm to extract the ball's position from the image plane p_{ball_i}. This method allows it to become independent of the color of the ball detection, overcoming one of the previously defined constraints. For the robots on the field the detection is performed by finding color blobs over the action areas and indexing (t) the colors to the color defined for each team $p^t_{object_i}$ (see figure 2).

2.3 Stereo Triangulation

With the points p_{ball_i} and $p^t_{object_i}$ obtained from the scene analysis of each camera, it was possible to obtain the 3D estimation of the objects by performing stereo triangulation. This method requires two or more two-dimensional camera views from a point, and it is implemented by a linear least-square fit of the intersection of two rays and defined by the 2D image points p_i, p_j, the intrinsic camera parameters and 3D camera configuration of each of the cameras[5] as shown in figure 2.

The cameras are modeled by the classical pinhole model. Therefore, if we assume a zero skew and a point in the camera $p_i = [u_i, v_i]$ image frame, a point

$P_i = (X_i, Y_i, Z_i)$ will be given by $z \cdot [u_i, v_i, 1]^T = K_i \cdot P_i$, where $K_i = \begin{bmatrix} f_x^i & 0 & u_{i_c} \\ 0 & f_y^i & v_{i_c} \\ 0 & 0 & 1 \end{bmatrix}$ is the intrinsic parameters[23] from the left camera, f_x^i, f_y^i is the focal length in pixels for both directions, (u_{i_c}, v_{i_c}) is the main point and z is the scale factor. The projection model will be the same for the camera on the right. In order to estimate the extrinsic parameters, the common procedure to estimate the rigid transformation parameters requires putting a calibration chessboard in front of both cameras. For the proposed scenario this procedure could not be implemented due to the distance between the cameras and the defined plane π as that would require a large chessboard to ensure that corners are detected with higher accuracy. Considering the constraint, the solution is to extract points in both cameras based on the available MSL field line intersection and triangulate their 3D positions. Then, using the least-squares method it is possible to estimate a 3D plane equation in the stereoscopic coordinate system, assuming the center of the field as the 3D position $(0, 0, 0)$ (see figure 2). This method makes it possible to obtain the extrinsic parameters relating the world coordinate frame and the camera coordinate frame, defined by a rotation R_i and translation T_i.

Both feature attributes and the epipolar constraint will be used to evaluate the correspondence between points detected in each camera. In order to meet the epipolar constraint: $P_i^T K_i^{-T} E K_j^{-1} P_j = 0$ where the essential matrix is $E = \hat{T}_j^i R_j^i$ and the relative rigid transformation parameters between cameras are defined as $R_j^i = R_i R_j^T \; T_j^i = T_i - R_j^i T_j$.

3 Results

In order to evaluate the Ground Truth (GT) system, the MSL is proposed as the implementation scenario as previously stated and justified. Two game situations are demonstrated to ensure a correct validation: two possible types of ball trajectory (see figure 4) and a game situation with more than a robot playing soccer (see figure 5).

Considering the first game situation, in both figures it is possible to observe the ball trajectory tracked by the robot and by the GT system (see figure 4). For more details on the ball trajectory tracked by the robot go to [17]. Focusing on the right figure, it is possible to observe that the robot was not able to track the ball with the same quality of the GT went the ball was far ($\sim 20m$) from the robot in the first parabola and improve its own detection went the ball was near ($\sim 5m$). In both figures, the GT was able to provide the trajectory of the ball with high accuracy.

For the second game situation (see figure 5), it was possible to observe that the ball was intercepted and kicked by a robot near the center of the field. Although the figure on the right only presents one robot kicking a ball, in the image on the left the GT system detects all robots inside the game field.

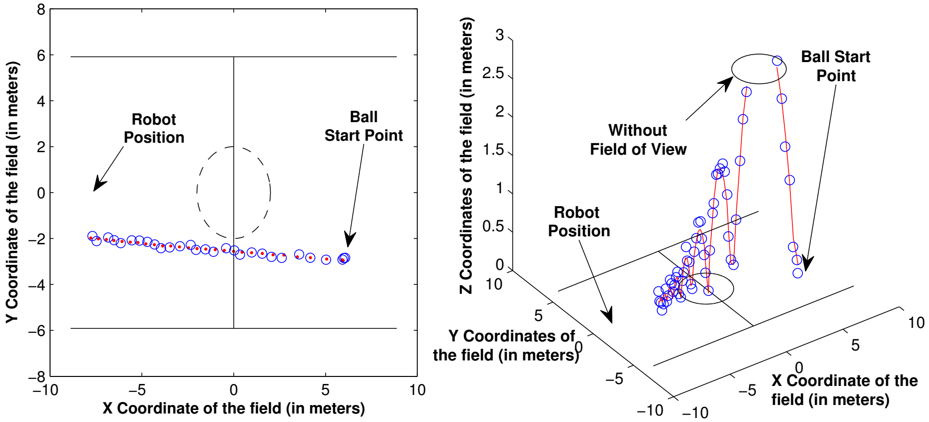


Fig. 4. Comparison of the ball position observed by the robot (blue circle) against the ground truth XY coordinate (red dot). **Left**: Ball kicked over the floor. **Right:** Ball kicked with a parabola trajectory.

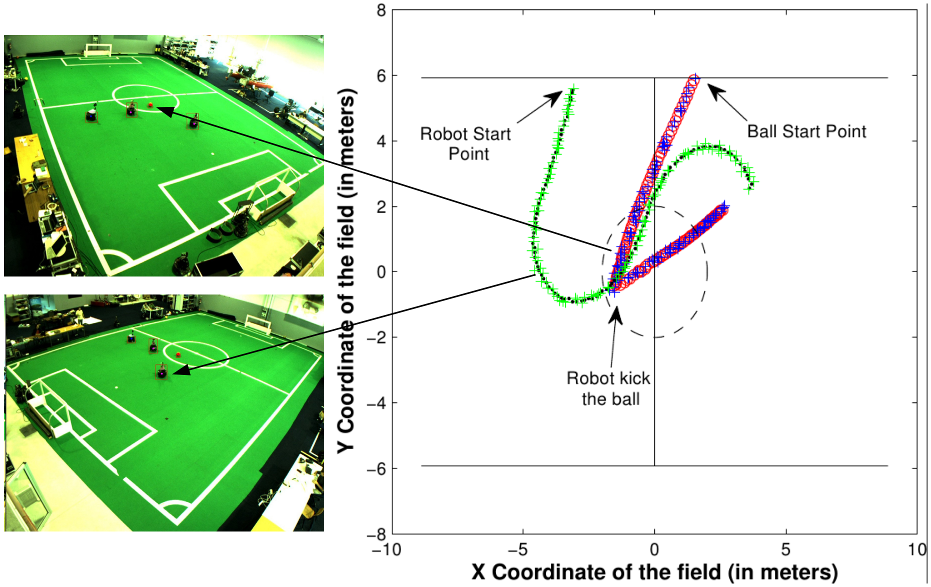


Fig. 5. Left: Snapshot from the stereo vision system installed for ground-truth evaluation of a game situation with three robots. **Right:** Comparison of the position of the ball and robot XY coordinate (blue, green) against the visual ground truth's XY coordinates for the same (red, black).

3.1 Error Analysis

The accuracy of 3D reconstructions was analyzed with two methods. In the first method, the distance between 2D points projected from a 3D estimate derived from the originally extracted two-dimensional points will be the measure of calibration precision. The result obtained from the mean reprojection error was less than one pixel. The second method will be used to determine accuracy, comparing the 3D coordinates and distances between coordinates measured through stereo triangulation to the values measured physically. The ground truth system proved to have high precision (low reprojection error) and a plausible accuracy, even at 20 meters from the cameras' position, and with an error below 0.05 meters.

4 Conclusions and Future Work

This paper proposes a vision based 3D ground truth tracking system for indoor scenarios. The proposed system does not require special markers or illumination such as the most commonly used in the current state of the art, and as a result it can be applied to a wider range of scenarios.

An open development approach was used taking advantage of the widespread ROS infrastructure, allowing further developments and a simple integration within the research community. This factor is strengthened by the fact that the original system has already been replicated and used in two other research laboratories (at ISRLAB of ISR/IST Lisbon and at GroundSys Lab at INESC Porto/ University of Porto).

A three layer architecture was proposed: decoupling data acquisition, image processing and 3D target position determination allowing further developments and modularity, leading to a high degree of adaptability to particular implementation scenarios.

The multi-camera system was characterized in a stereo setup, using standard GigaE digital cameras and implemented in common computational hardware.

The time synchronization required to produce valid ground truth data for multiple robotic research evaluations was taken into account with the analysis of various time synchronization protocols. The overall time offset achieved was under $2.6\mu s$.

The results were evaluated in the RoboCup MSL scenario as this is a highly dynamic and representative benchmark of multi-robot operation scenarios.

Multiple methods of background subtraction were analyzed to identify target areas. A color based and a morphological feature detector were used to track both the ball and the robots on the field.

This system has already been used to evaluate research results providing valuable ground truth data [17].

The following topics can be considered for further work: extension of the area using multiple additional cameras, particularly with the increased operational area and precision to be characterized. Different additional target detection methods are also being analyzeds for other application scenarios. An optimization camera to estimate position will be integrated in order to maximize the field of view of the action area.

We want to receive improvements and add-ons to the system from the robotic research community, and therefore the developed visual ground truth ROS stack will be available at the ROS website.

Extending the system to outdoor operations under more extreme lighting conditions may also be addressed in the future.

Acknowledgment. The authors acknowledge the major support given by the ISEP-IPP Institution, by the INESC TEC, to this project. This work is financed by the ERDF - European Regional Development Fund through the COMPETE Programme and by National Funds through the FCT - Portuguese Foundation for Science and Technology within project FCOMP-01-0124-FEDER-022701 and under Perception-Driven Coordinated Multi-Robot Motion Control Project with reference PTDC/EEA-CRO/100692/2008.

References

1. Achtelik, M., Weiss, S., Chli, M., Dellaert, F., Siegwart, R.: Collaborative stereo. In: Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS (2011)
2. Augenstein, S.: Monocular Pose and Shape Estimation of Moving Targets, for Autonomous Rendezvous and Docking. PhD thesis, Stanford University (2011)
3. Chrony, `http://chrony.tuxfamily.org/`
4. Torres, P.R., Murta, C.D.: Characterizing quality of time and topology in a time synchronization network. In: 49th IEEE Global Telecommunications Conference, IEEE GLOBECOM 2006 (2006)
5. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004) ISBN: 0521540518
6. Computational Interaction and Robotics lab John Hopkins University. A brief tour of xvision (2008), `http://www.cs.jhu.edu/CIPS/xvision`
7. Khandelwal, P., Stone, P.: A low cost ground truth detection system for roboCup using the kinect. In: Röfer, T., Mayer, N.M., Savage, J., Saranlı, U. (eds.) RoboCup 2011. LNCS (LNAI), vol. 7416, pp. 515–527. Springer, Heidelberg (2012)
8. Koch, A., Berthelot, A., Eckstein, B., Zweigle, O., Häussermann, K., Käppeler, U.-P., Tamke, A., Rajaie, H., Levi, P.: Advanced data logging in robocup. In: Dillmann, Beyerer, Stiller, Zöllner (eds.) Proceedings of the AMS 2009, Autonome Mobile Systeme, Karlsruhe, Germany, December 3-4. Informatik Aktuell, pp. 1–8. Springer (2009)
9. Lee, G.H., Achtelik, M., Fraundorfer, F., Pollefeys, M., Siegwart, R.: A benchmarking tool for mav visual pose estimation. In: Proc. 11th Int. Control Automation Robotics & Vision (ICARCV) Conf., pp. 1541–1546 (2010)
10. Lochmatter, T., Roduit, P., Cianci, C., Correll, N., Jacot, J., Martinoli, A.: Swistrack - a flexible open source tracking software for multi-agent systems. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, IROS 2008, pp. 4004–4010 (2008)

11. Michael, N., Shen, S., Mohta, K., Mulgaonkar, Y., Kumar, V., Nagatani, K., Okada, Y., Kiribayashi, S., Otake, K., Yoshida, K., Ohno, K., Takeuchi, E., Tadokoro, S.: Collaborative mapping of an earthquake-damaged building via ground and aerial robots. Journal of Field Robotics 29(5), 832–841 (2012)
12. Euron European Robotics Search Network. Introduction: Benchmarks in robotics research, http://www.robot.uji.es/EURON/en/index.htm
13. Niemüller, T., Ferrein, A., Eckel, G., Pirro, D., Podbregar, P., Kellner, T., Rath, C., Steinbauer, G.: Providing ground-truth data for the nao robot platform. In: Ruiz-del-Solar, J. (ed.) RoboCup 2010. LNCS, vol. 6556, pp. 133–144. Springer, Heidelberg (2010)
14. Parks, D.H., Fels, S.S.: Evaluation of background subtraction algorithms with post-processing. In: IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, AVSS 2008, pp. 192–199 (September 2008)
15. Röfer, T., Laue, T., Müller, J., Fabisch, A., Feldpausch, F., Gillmann, K., Graf, C., de Haas, T.J., Härtl, A., Humann, A., Honsel, D., Kastner, P., Kastner, T., Könemann, C., Markowsky, B., Lars Riemann, O.J., Wenk, F.: B-human team report and code release 2011, p. 79 and p. 80 (2011)
16. Miller, P., Fry, S.N., Bichsel, M., Robert, D.: Tracking of flying insects using pan-tilt cameras. Journal of Neuroscience Methods 101, 59–76 (2000)
17. Silva, H., Dias, A., Almeida, J., Martins, A., Silva, E.: Real-time 3d ball trajectory estimation for robocup middle size league using a single camera. In: Röfer, T., Mayer, N.M., Savage, J., Saranlı, U. (eds.) RoboCup 2011. LNCS, vol. 7416, pp. 586–597. Springer, Heidelberg (2012)
18. Soltero, D.E., Schwager, M., Rus, D.: Generating informative paths for persistent sensing in unknown environments. In: Proc. of the International Conference on Intelligent Robots and Systems, IROS 2012 (October 2012)
19. Zickler, S., Laue, T., Birbach, O., Wongphati, M., Veloso, M.: SSL-vision: The shared vision system for the robocup small size league. In: Baltes, J., Lagoudakis, M.G., Naruse, T., Ghidary, S.S. (eds.) RoboCup 2009. LNCS, vol. 5949, pp. 425–436. Springer, Heidelberg (2010)
20. Stulp, F., Gedikli, S., Beetz, M.: Evaluating multi-agent robotic systems using ground truth. In: Proceedings of the Workshop on Methods and Technology for Empirical Evaluation of Multi-agent Systems and Multi-robot Teams, MTEE (2004)
21. TechUnited Middle Size League Team. Greenfield augmented reality (2010), http://www.techunited.nl/wiki/index.php?title=GreenfieldAugmentedReality
22. Vallat, A., Schneuwly, D.: Clock synchronization in telecommunications via ptp (IEEE 1588). In: IEEE International Frequency Control Symposium, 2007 Joint with the 21st European Frequency and Time Forum (2007)
23. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. 22(11), 1330–1334 (2000)
24. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. 27(7), 773–780 (2006)