Active Vision-based Localization For Robots In A Home-Tour Scenario

Falk Schubert, Thorsten Spexard, Marc Hanheide, and Sven Wachsmuth

Bielefeld University, D-33615 Bielefeld, Germany,
falk.schubert@eads.net
{tspexard, mhanheid, swachsmu}@techfak.uni-bielefeld.de,
WWW home page: http://www.techfak.uni-bielefeld.de/ags/ai/

Abstract. Self-Localization is a crucial task for mobile robots. It is not only a requirement for auto navigation but also provides contextual information to support human robot interaction (HRI). In this paper we present an active vision-based localization method for integration in a complex robot system to work in human interaction scenarios (e.g. *home-tour*) in a real world apartment. The holistic features used are robust to illumination and structural changes in the scene. The system uses only a single pan-tilt camera shared between different vision applications running in parallel to reduce the number of sensors. Additional information from other modalities (like laser scanners) can be used, profiting of an integration into an existing system. The camera view can be actively adapted and the evaluation showed that different rooms can be discerned.

1 Introduction

Self-Localization is concerned with determining a robot's spatial position with respect to the environment. This position is not only relevant to facilitate autonomous navigation but also to enable situation-aware interaction between a robot and its users. While for the first mostly exact metric position information within a known map is required to allow path planning, for the latter case a coarse and symbolic localization might already be sufficient. Compared to human-human interaction room names (living room, hallway) are much more appropriate as location information than position coordinates on a map and allow direct communication between the robot and the user about the environment for a joint exploration. Further advantages of such a coarse localization are faster learning and classification without requiring a pre-known map. For a complete localization multi modal mobile robots can incorporate different localization methods to refine each other or being activated separately depending on the desired level of position information.

As an example scenario benefiting from symbolic, coarse position, we identified the *home-tour* scenario. In this scenario, a human user starts a robot for the first time in her apartment and shows it around to familiarize it with this new environment. The human introduces particular parts of the environment and tells the robot "Look robot, this is my living room", for instance. After the *home-tour* the robot should be able to introduce other people to the apartment in the same way.

For such a real world scenario localization methods have to be robust in different ways. A desired property of localization methods is the robustness concerning minor

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

changes in the structure of the scene caused by rearranged furniture, new items like plants or pictures on the wall for instance. Depending on the sensor used for localization this can be hard to accomplish. The laser scans for a rearrangement of furniture in one room (see Fig. 4b and 4c) look totally different. Furthermore localization methods often work very well in predefined static areas like labs or offices. In those scenes disturbances caused by dynamic changes in the structural layout of a scene like moving objects or walking people talking to the robot are rare. When using a mobile robot in a real life scenario however one has to adapt to those dynamic changes in the scene. Additional to structural changes, the system should also be invariant to small illumination changes. Many vision-based systems using histograms have troubles with unstable light conditions [1].

We developed a vision-based localization method that uses so called gist-based features [2,3] capturing the holistic structure of a scene. The approach features the properties mentioned above and is designed for efficient integration into a complex mobile robot system BIRON. The localization module can also use additional information from other sensors and uses a shared camera. In the robot system the localization part can be fully controlled by speech. Thus localization training and application can be investigated from a user centered point of view.

2 Comparing Alternative Localization Methods

In this section we give a short overview of existing vision-based localization methods. Many different methods exist to localize a mobile robot. They can be divided up into three main categories: geometric, topological and hybrid [1]. The first one tries to find a metric position for the robot mostly based on either a given map, based on the knowledge of the start position [4] or by continuously building a map during localization (SLAM) [5]. Methods belonging to the second group only estimate the coarse location often only knowing the neighbor relations of possible locations stored in an adjacency map or even without a pre-known map [1]. Location methods can also be divided into different groups based on the sensors used. Many geometric variants use more or less accurate sensors like GPS, compass, odometry information, laser scanners or sonar sensor to name the most common ones [6]. For a coarse localization low cost sensors like a camera are often sufficient. Quite some research has been done in this field of so called vision-based localization approaches [7,8,9,10]. These methods differ in the features computed for each image and in the classifier used to match them to representative images prerecorded for each possible location. Most of them use color histograms computed on images recorded with an omni-directional camera [1]. The use of histograms can be problematic as they are sensitive to changes in illumination. Hence the features need to be extended to capture structural properties as well. Classifying a single image of a complete scene into holistic classes has been tried in many different ways. Most methods are example-based classification approaches using features based on color and structure [11,12,13,14]. [14] also evaluated the application for place recognition, however with rather a focus on context-based object recognition than on integration on a mobile robot.

None of the vision-based methods mentioned above uses an active vision approach to change the field of view in cases of low confidence for a location. These can occur

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8

through dynamic objects (e.g. moving people) occluding the viewfield or through other applications sharing the same camera and focusing it on areas not specific for localization (e.g. face detector focuses on heads of humans). Many vision-based methods require an omni-directional camera solely for the localization task in order to receive stable results in respect to rotational variances and small dynamical changes in the scene. We propose a new localization method based on existing methods adapted to work with only a single pan-tilt camera which is also used for other detection purposes. Active sensors used during localization have already been evaluated for a geometric map-based localization method [4].

Only few vision-based approaches combine geometric and topological localization methods like done in [15]. We integrated the localization method in a framework, which is easily extendable by different localization algorithms sharing same sensors. In the following section we describe how the system classifies an image of the current view into a scene.

3 Holistic Scene Classification

A main problem of classifying holistic scenes lies in finding a compact and yet descriptive representation of scenes. How can we extract enough information from as few images as possible to retrieve information about the class the scene belongs to? Oliva and Torralba [16] suggest to use an holistic view on a scene to entirely describe its character. This representation corresponds to the perceptual gist [3], which according to the author is similar to the way humans perceive scenes. Based on successful work using this theory [2], we compute a feature vector using neighborhood filters. This way the features can be computed very efficiently and independently from each other. We use a boosting algorithm to select scene specific features that separate a scene from all others. This selection process helps to speed up the classification process as suggested in [17].

3.1 Feature Selection

We compute 12 different filter responses (11 edge filters with different orientations, 1 corner filter) as well the image intensity itself from 46 differently sized and located regions of the image. These filtered patches are reduced to two values by taking the sum of the pixels for each patch to the power of 2 (energy) and 4 (kurtosis) respectively. Because of the nature of the boosting algorithm the number of patches and filters can not be increased arbitrarily as a single, too specific feature will not be discriminative enough. Too few filters and patches on the other side will decrease the flexibility concerning class types that can be trained using this feature set. Compared to the perceptual gist, edges and corners seem a good computational representation as they also capture the coarse structure of a scene as shown with some typical filter responses depicted in Fig. 2. Because of the feature selection boosting performs during training, we can compute many features ($1196 = 13 * 46 * 2$; 12 filters and image intensity, 46 patches, energy and kurtosis) and receive the most discriminative ones for the training set. In Fig. 1 the most discriminative feature for each scene of the training set are shown. For example for the living room the big window and the shelve result in higher horizontal

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

edge responses than in other classes. For the hallway the lack of many horizontal edges separates this scene from the others.

Fig. 1: Best features selected for each scene, from left to right: living room with 5px horizontal edge filter ($p = -1$), hallway with 7px horizontal edge filter ($p = 1$), dinner room with light intensity ($p = 1$), kitchen with light intensity ($p = -1$).

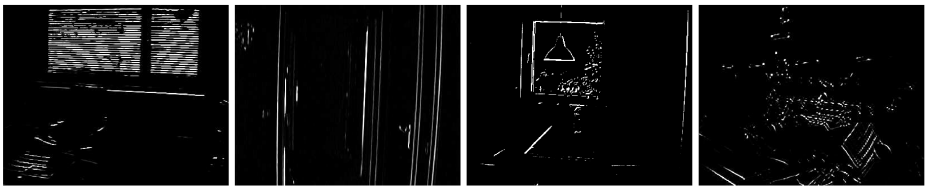


Fig. 2: Typical filter responses for different scenes, from left to right: horizontal edges in the living room, vertical edges in the hallway, slanted edges in the dinner room and corners in the kitchen.

3.2 Classification

AdaBoost [18] has been successfully applied to many classification problems [17]. As mentioned above, boosting can be used to perform a selection of most discriminative features, thus reducing the computational cost during the classification process. Each selected feature can be seen as a weak binary classifier separating a class from the rest. We use discrete AdaBoost [19] to train a set of such weak binary classifiers $h_i(\boldsymbol{x})$ based on thresholds σ_i and polarities p_i (specifying whether x_i has to be greater than σ_i to belong to the class or vice versa) using only the i-th dimension of the feature vector $\boldsymbol{x}$:

$$h_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } p_i \cdot x_i < p_i \theta_i \\ 0 & \text{otherwise} \end{cases}$$

A weighted majority vote of the weak classifiers using the boosted weights α_r defines the strong classifier:

$$H(\boldsymbol{x}) = \sum_r \alpha_r h_r(\boldsymbol{x})$$

For each boosting round the algorithm computes one weak classifier (h_r) which corresponds to the selection of one feature. The number of rounds defines the number of features selected. In a one-vs-all scheme we compute n strong classifiers $H_u(\boldsymbol{x})$ that

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

separate class u from the rest. The strongest classifier H determines the final class $c = \text{argmax}_u H_u$. Since we are classifying single frames in a continuous image sequence, neighbored frames are likely to belong to the same class, most the time. Hence we average the final class decision over a multi staged frame window (e.g. 5,10 and 20 frames) using a majority voting scheme.

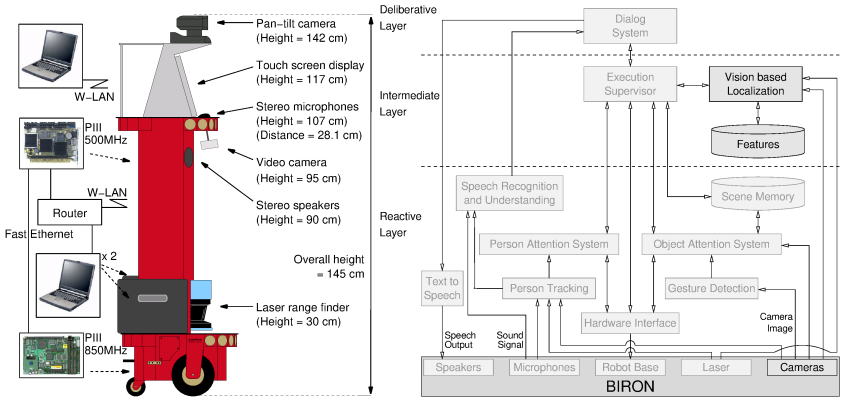
3.3 Rejection Scheme and Active Vision

A very simple rejection scheme can be achieved by setting a threshold t_{max} for the winning classifier H_c. This however may cause falsely rejected images, when classes are not very distinct. We adapt the thresholding to the classes itself by using a separate classifier that learns to classify the output H_c into either *class* or *rejection*. We use a collection of heuristically chosen confidence features as a feature vector. Among them are the distances $\boldsymbol{x} - \theta_i$ of the data features to the thresholds of the 4 best weak classifiers, the sum of those distances, the output of the strong classifier for the winning class H_c as well as its differences to the other strong classifiers $H_u - H_c$. A perceptron is trained to find a hyperplane $\boldsymbol{w}$ that splits misclassified examples from correct ones. This allows to compute a confidence value $c = \boldsymbol{w}^T \boldsymbol{x}$ on a test set. Examples yielding a confidence $c < 0$ are rejected. We average the confidence values over multiple frame windows using majority voting. Instead of just ignoring an image when it gets rejected, we interpret this information that somehow the scene is occluded (e.g. by a person in the field of view or by a bad view caused by a different application sharing the same camera). Thus we implemented 3 attempts to resolve this occlusion by panning the camera away from its current field of view. This can be done either by panning around a fixed angle until no rejection occurs, by moving along a trajectory trying to avoid typically moving humans in front of the robot or by moving the camera to the angle where the laser ranger (if available) detects the deepest area in the scene.

4 Building the Integrated System BIRON

BIRON- the Bielefeld Robot Companion was developed as a demonstrator in home-like surroundings regarding to the *home-tour* scenario. It bases on the *Pioneer PeopleBot*^{TM} from *ActiveMedia* (see Fig. 3). The robot has a height of 145 cm and is equipped with several sensors to obtain information about the environment and the surrounding humans: a Sick laser range finder for a frontal area perception and leg detection with a range of $180°$, two farfield microphones for sound source localization, and a touch screen display for direct user input. The pan-tilt color camera mounted on the top of the robot acquires detailed images of the scene like furniture, hand-size objects, and the upper body part, especially the face, of human interaction partners. A second color video camera is mounted below the microphones and used for less detailed scene perception and deictic gesture detection. Within the robot chassis two PCs are embedded connected by Fast-Ethernet for controlling the drive, on-board sensors, sound localization and person tracking with interaction attention. Paying tribute to the increasing demand of computational power based on the growing amount of software modules integrated on BIRONtwo additional notebooks mainly used for image processing as

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de



(a) Hardware (b) Architecture

Fig. 3: Hardware composition (a) and integration of the localization module (darkened) into the existing (bright) robot control architecture (b) of BIRON.

face detection, object recognition and the visual localization are fixed at the bottom platform. These are connected to the internal PCs by LAN via router and by wireless LAN to a third notebook (Pentium Mobile 1.5GHz 768MB Ram) running the speech recognition and dialog management. For the successful and efficient integration of the different modules not only sufficient calculation power is needed but also a well designed integration architecture [20] is required, enabling an easy and fast integration of new modules like the presented vision-based localization. We chose a hybrid architecture as proposed by [21] containing one layer for deliberative modules, one layer for reactive components and one intermediate layer for synchronizing the results of the different layers and modules. The synchronization becomes necessary as all modules are operating independently, exchanging information by XML. The XML communication is supported by the XML Communication Framework (XCF) [22] capable of handling both, Remote Method Invocation (RMI) via function server for irregular data exchange on certain events and 1:n data streaming via one publisher and n subscribers for frequent and periodically data exchange, e.g. for publishing laser data. For laser data and other uninterpreted data like video or audio information a direct data connection between the concerning modules is established. Preprocessed data is exchanged via the *Execution Supervisor* (ESV) as depicted in Fig. 3. As the central module of the intermediate layer the ESV is responsible for the previously mentioned synchronization of the high level data flow between the modules. Therefore the ESV is in principle a finite state machine using the data input of the different modules as *Events* for a transition from state A to a state B. For any transition a number of *Orders* can be sent to the modules of the reactive layer containing new operating parameters (e.g. the Person Attention System now fixates a certain person instead of looking around). The modules of the deliberative layer may also get new information from the ESV during a transition, but they are only informed about the current *State* of the system and are ”free” to react depending only on their task. If an event from a deliberative module is received by the ESV that can not be processed at the current state the event is queued until an according state is entered or the event becomes too old. In the latter case the corresponding deliberative module will be notified that the sent data was not processed by a *Reject* message.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Using these predefined interface structures of the ESV the localization module is able to exchange information with many of the remaining modules in the overall system. Therefore we will shortly describe the remaining modules and how they can interact with the localization module:

Speech Recognition, Speech Understanding, and Dialog The speech recognition system [23] translates the user utterances into words, which are feed to the speech understanding [24]. The speech understanding uses not only grammatical information for the syntactical analysis of the speech recognition results, but can also enhance these results taking into account the semantical meaning of an utterance as well. Finally the information is given to the Dialog [25] converting both natural spoken language into processable events and the current system procedures into human-oriented speech. Given an user utterance like: "This is the kitchen." a label for the current position will be sent to the ESV for transmitting it to the localization module.

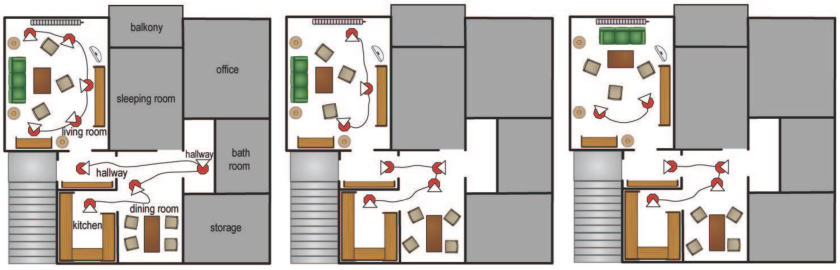
Gesture Detection and Object Attention System While the Dialog may provide label information for the Localization the Gesture Detection and Object Attention System (OAS) [26] can use the localization information. The OAS gathers information about an object at which a person points. After a pointing gesture is detected by the Gesture Detection it sends the estimated target coordinates to the OAS. Without the Localization the OAS receives only the relative coordinates from Gesture Detection. To store which objects are common for a certain area like 'the glasses are at the living room' and such to increase the robot's abilities, the information from the localization module can be used.

Person Tracking and Attention System Besides label information from the Dialog the presented localization is able to use additional information from other modules like the Person Attention System (PTA) if available to increase its robustness. The PTA [27] tracks multiple humans by fusing information from different sensors. Therefore the anchoring approach by Coradeschi and Saffiotty [28] is used for each modality: Laser scans are used for classifying pairs of legs, a speaker localization is performed on audio data, and both position and gazing direction of faces are detected on video images. Subsequently the person most likely to interact with the robot is selected and data concerning this person is delivered. This data also contains the current person position which can be used by the localization module for the previously mentioned active vision. In case of occlusion by a person the camera can be controlled to look away from the person position given by the PTA. Even if no PTA was available the Localization can directly read the laser distances from the Hardware Interface module aligning robot and camera to a direction with a minimum distance to the next obstacle.

5 Experimental Evaluation

We focus our system to be usable in a real world *home-tour*. Hence, we simulated such a scenario by showing the robot 4 different rooms (living room, hallway, dinner room, kitchen) for various viewpoints selected by a user who is familiar with the apartment and had no idea about the methods used. The tour and the points for training are shown in Fig. 4a. We captured the image sequences from the robot viewpoint at those locations (marked as red circles in Fig. 4) by panning the camera about 90-120 degrees (denoted by the triangle on each red circle in Fig. 4) and without manually selection. Depending

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de



(a) training (b) testing (c) rearranged

Fig. 4: Different tours through the apartment for training, testing and with rearranged living room.

on the number of viewpoints and degrees for panning through the scene for each room, we recorded 3075, 1164, 726 and 471 images for the living room, hallway, dining room and kitchen. Example images are shown in Fig. 1. We used AdaBoost as described in section 3 to train the classifier 50 boosting rounds on those images. After the training we recorded images in the same way for the testing in the same way as described above. The scenes where recorded at different locations (see Fig. 4b) in the rooms, without moving objects (e.g. humans) in the view field and with minor changes like slightly moved chairs and accessories on furniture. For the test images with a clear view on the scene (e.g. no people walking in front of the robot), the classification rates are applicable for localization, as the results show in Tab. 1. Even for the rearranged living room (Fig. 4c) the classification is about 90% correct. Algorithms solely based on the laser scans might fail at this point since the depth profiles after rearrangement look totally different. For a real *home-tour* with humans interacting with the robot, the classification rates are only acceptable for the living room and dinner room (see Tab. 1). Because of the small space in the hallway and kitchen the human had to stand very close to the robot occluding most of the scene. However both rejection schemes show high rejection rates for these small rooms. In those cases the active vision module can pan the camera away from the user using one of the three modes in the hope to capture a clear view onto the scene (although this was not used in the evaluation). By increasing the thresholds for both rejection schemes it possible to decrease the false-positive rate as only absolute certain classification results will be accepted. However the trade-off will be more cases where the robot has to look thus increasing the response time. Because of the use of edge and corner filters, the classification results are also stable to illumination changes to some extend.

6 Summary & Outlook

We proposed an active vision-based localization approach for an integrated, mobile robotic system using a pan-tilt camera shared with other vision components running on the platform. It applies holistic visual features allowing the approach to directly facilitate human-robot interaction in real world settings by means of symbolic localization information like room names. Being integrated in the system architecture of the robotic system, our localization approach can be interactively trained. Also additional informa-

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

	no humans	humans, $H < 0$		humans, $c < 0$	
room	correct (Fig.4b / Fig.4c)	correct	rejected	correct	rejected
living	89% (88% / 90%)	99%	14%	100%	7%
hallway	76%	16%	43%	0%	27%
dinner	89%	61%	52%	61%	12%
kitchen	75%	25%	35%	18%	3%

Table 1: Classification rate during *home-tour*. Left table shows percentage of correct classified images without humans in the field of view, the middle table shows rates for images with human interacting with the robot and using the simple rejection scheme (rejected images could not be classified in any of the classes), the right table one shows results using the advanced rejection scheme instead.

tion from other modalities like laser scanner can be used. The overall localization can be easily extended by alternative localization methods.

Different rejection schemes are being used to trigger active control of the camera in cases of uncertainty of classification to make the approach feasible even in challenging settings. The localization method has been successfully tested in a real world apartment in a typical human-robot interaction scenario (*home-tour*). The system turned out to be quite robust to changes in illumination and restructuring of the scene.

Though the currently implemented active vision strategies already allow to make the classification more robust, further elaborated approaches promise to increase the performance. As part of ongoing research a joint logic between different localization results would be of interest to receive better and more stable results. Also a more exhaustive evaluation of the human interaction scenario would help to assess the amount of restructuring the scene the system can deal with.

References

1. Ulrich, I., Nourbakhsh, I.R.: Appearance-based place recognition for topological localization. In: Proceedings of IEEE International Conference on Robotics and Automation. (2000) 1023–1029
2. Murphy, K., Torralba, A., Freeman, W.T.: Using the forest to see the tree: a graphical model relating features, objects and the scenes. In: Advanced Neural Information Processing Systems, Vancouver, BC, MIT Press (2003)
3. Oliva, A.: Gist of a scene. In: Neurobiology of attention. Academic Press, Elsevier, San Diego, CA (2005) 251–256
4. Burgard, W., Fox, D., Thrun, S.: Active mobile robot localization. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI), San Mateo, CA, Morgan Kaufmann (1997)
5. Dissanayake, M., Newman, P., Clark, S., Durrant-Whyte, H., Csorba, M.: A solution to the simultaneous localization and map building (slam) problem. In: Robotics and Automation, IEEE Transactions on. Volume 17. (2001) 229–241
6. Gutmann, J., Burgard, W., Fox, D., Konolige, K.: An experimental comparison of localization methods (1998)
7. Blaer, P., Allen, P.K.: Topological mobile robot localization using fast vision techniques. In: ICRA. (2002) 1031–1036

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8

8. Kosecka, J., Zhou, L., Barber, P., Duric, Z.: Qualitative image based localization in indoors environments. CVPR **02** (2003) 3
9. Wolf, J., Burgard, W., Burkhardt, H.: Robust vision-based localization for mobile robots using an image retrieval system based on invariant features. In: ICRA. (2002) 359–365
10. Dudek, G., Jugessur, D.: Robust place recognition using local appearance based methods. In: ICRA. (2000) 1030–1035
11. Serrano, N., Savakis, A.E., Luo, J.: A computationally efficient approach to indoor/outdoor scene classification. In: 16th International Conference on Pattern Recognition (ICPR). Volume 4. (2002) 146–149
12. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98. (1998) 42–51
13. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.J.: Image classification for content-based indexing. In: IEEE Transactions on Image Processing. Volume 10. (January 2001) 117–130
14. Torralba, A., Murphy, K., Freeman, W., Rubin, M.: Context-based vision system for place and object recognition. In: ICCV. (2003) 273–280
15. Georgiev, A., Allen, P.K.: Localization methods for a mobile robot in urban environments. IEEE Transactions on Robotics **20**(5) (October 2004) 851–864
16. Oliva, A., Torralba, A.B.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42**(3) (2001) 145–175
17. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1., Kauai, HI, USA (2001) 511–518
18. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**(1) (1997) 119–139
19. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. Technical report, Dept. of Statistics, Stanford, University Technical Report (1998)
20. Kleinehagenbrock, M., Fritsch, J., Sagerer, G.: Supporting advanced interaction capabilities on a mobile robot with a flexible control system. In: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems. Volume 3., Sendai, Japan (September/October 2004) 3649–3655
21. Arkin, R.C.: Path planning for a vision-based autonomous robot. In: Proc. SPIE Conf. on Mobile Robots, Cambridge, MA (October 1986) 240–249
22. Wrede, S., Fritsch, J., Bauckhage, C., Sagerer, G.: An XML Based Framework for Cognitive Vision Architectures. In: Proc. Int. Conf. on Pattern Recognition. Number 1 (2004) 757–760
23. Fink, G.A.: Developing HMM-based recognizers with ESMERALDA. In Matoušek, V., Mautner, P., Ocelíková, J., Sojka, P., eds.: Lecture Notes in Artificial Intelligence. Volume 1692., Berlin Heidelberg, Springer (1999) 229–234
24. Hüwel, S., Wrede, B.: Situated speech understanding for robust multi-modal human-robot communication. In: Proceedings of the International Conference on Computational Linguistics (COLING/ACL), ACL Press (2006)
25. Li, S., Wrede, B., Sagerer, G.: A computational model of multi-modal grounding. In: Proc. ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006, ACL Press (2006)
26. Haasch, A., Hofemann, N., Fritsch, J., Sagerer, G.: A multi-modal object attention system for a mobile robot. In: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Edmonton, Canada, IEEE (2005) 1499–1504
27. Lang, S., Kleinehagenbrock, M., Hohenner, S., Fritsch, J., Fink, G.A., Sagerer, G.: Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In: Proc. Int. Conf. on Multimodal Interfaces, Vancouver, Canada, ACM Press (2003) 28–35
28. Coradeschi, S., Saffiotti, A.: Perceptual anchoring of symbols for action. In: Proc. of the 17th IJCAI Conference, Seattle, Washington (2001) 407–412

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de