Categorizing Perceptions of Indoor Rooms Using 3D Features

Agnes Swadzba and Sven Wachsmuth

Applied Computer Science, Faculty of Technology, Bielefeld University,
Universitätsstraße 25, 33615 Bielefeld, Germany
{aswadzba, swachsmu}@techfak.uni-bielefeld.de

Abstract. In this paper, we propose a holistic classification scheme for different room types, like office or meeting room, based on 3D features. Such a categorization of scenes provides a rich source of information about potential objects, object locations, and activities typically found in them. Scene categorization is a challenging task. While outdoor scenes can be sufficiently characterized by color and texture features, indoor scenes consist of human-made structures that vary in terms of color and texture across different individual rooms of the same category. Nevertheless, humans tend to have an immediate impression in which room type they are. We suggest that such a decision could be based on the coarse spatial layout of a scene. Therefore, we present a system that categorizes different room types based on 3D sensor data extracted by a Time-of-Flight (ToF) camera. We extract planar structures combining region growing and RANSAC approaches. Then, feature vectors are defined on statistics over the relative sizes of the planar patches, the angles between pairs of (close) patches, and the ratios between sizes of pairs of patches to train classifiers. Experiments in a mobile robot scenario study the performance in classifying a room based on a single percept.

1 Introduction

Context is a rich source of information for interpreting tasks in complex scenes. This has already been recognized for a long time. Systems like CONDOR [1] or SPAM [2] coded explicit contextual rules that triggered other image operations. The main drawback of such kind of models was the complex knowledge engineering task and the semantic deficiencies of extensional systems dealing with uncertainty. More recently, graphical models have been applied in order to provide a more concise model relating objects and aspects of the scene [3,4]. Murphy, Torralba, and Freeman estimate global contexts, like persons, vehicles, furniture and vegetation from low-level image features [3]. These provide constraints on object categories and object scales. Hoiem, Efros, and Herbert first extract a 3D surface geometry from 2D images and relate the estimated local geometries to object classes predicted by a window-based object detector [4].

The work discussed so far mainly deals with 2D image information. Murphy et al. demonstrate that many different scene categories can be distinguished by

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 734–744, 2008.

purely considering 2D image statistics on texture and edges. There have also been other approaches using additional features like color histograms that successfully distinguished indoor/outdoor [5], sky/no-sky, vegetation/no-vegetation [6]. However, this does not necessarily extrapolate to more finely graded scene categories like different types of rooms, e.g. office or meeting room. Here, typical furniture like tables, chairs, and shelves reoccur, but in different layouts. Furthermore, furniture in the same type of room may have changing colors and textures or be viewed from different directions.

In these cases, a 3D description of the scene is much more invariant with regard to inner-class variations. However, strategies that provide a complete semantic interpretation of the 3D scene suffer from very constraint settings and extensive modeling efforts. Therefore, we aim at a more holistic 3D approach to scene classification in the spirit of the gist approach used by Torralba [7]. In the following, we describe the scene by a collection of planar structures and analyze whether it is possible to compute proper feature vectors for the classification of different room types (here: office, hall, and meeting room). The challenge faced is to categorize rooms only based on the information of one frame. Section 2 presents the 3D Time-of-Flight (ToF) sensor for acquiring 3D information in real time and introduces preprocessing steps specialized on this data. Section 3 presents necessary steps for determining sets of planar structures in this data, and in Section 4 and 5 features and classifiers are chosen and examined with regard to their performance in categorizing room percepts to room types.

2 Acquiring and Preprocessing 3D ToF Data

(a) (b) (c)

Fig. 1. (a) Swissranger SR3000, (b) example amplitude image, and (c) example 3D point cloud preprocessed

Our system uses the Swissranger SR3000 (Fig. 1(a)) provided by Swiss Center for Electronics and Microtechnology (CSEM) [8] delivering a matrix of distance measurements independent from texture and lighting conditions. It consists of 176 × 144 CMOS pixel sensors which are able to determine actively the distance between the optical center of the camera and the real 3D world point via measuring the time-of-flight of a near-infrared signal. Besides the distance value matrix (Fig. 1(c)), the camera provides a matrix containing amplitude values (Fig. 1(b)) for each frame. The amplitude value indicates the amplitude of the reflected near-infrared signal received by the sensor and implies therefore the amount of light reflected by a world point. A small amplitude corresponds to a small amount of light reflected indicating a weak signal.

To deal with noise arising from the different reflection properties several preprocessing techniques proposed in [9] are applied. The distance image is smoothed with a distance-adaptive median filter, which uses a different mask

size (e.g. 3×3, 5×5, or 7×7) depending on the distance value for each pixel. As the amplitude value refers to the quality of the distance measurement, points with a small amplitude value are removed from the final 3D point cloud. The threshold needed adapts automatically to different reflection properties in different scenes. Further, edge points arising in the case when light from the fore- and the background hits the same pixel simultaneously are rejected. Finally, 3D coordinates are generated out of the distances with regard to a 3D camera coordinate system. With the assumption of ideal perspective projection, the known position of the principal point, pixel sizes, and focal length, the 3D coordinates can be computed from the distances via ray proportions in triangles. As a result the computed 3D points are organized regularly in a 2D matrix. This 2D arrangement enables us to apply 2D preprocessing and search methods nicely to 3D data saving computation time and complexity. Nevertheless, all methods proposed in the following are applicable on any type of 3D data.

3 Planes – Meaningful Structures

For many applications it is necessary to extract meaningful structures which enable a meaningful description of complex scenes. As the Swissranger camera provides $2\frac{1}{2}$D data from which 3D points are computed it can be focused on geometric aspects. Human-made environments – like walls, floors, and furniture – consist of large planar structures. Therefore, it is a reasonable step to find planar surfaces within a given 3D point set. It is assumed that preceptions of halls, offices, and meeting rooms can be categorized in a proper way using planar structures, because they provide more stable features compared to colors, textures, and materials occurring in different indoor scenarios.

In principal there are three possibilities to extract planes from a 3D point set. First, the Random Sample Consensus (RANSAC) algorithm [10] can be used to fit robustly plane models in 3D data, possibly combined with the Iterative Closest Points (ICP) algorithm [11] or SIFT features for refining the planes [12,13]. Second, the Expectation Maximization (EM) algorithm can be used to adjust the number of planes and estimates the locations and orientations by maximizing the expectation of a log-likelihood function [14,15]. Finally, region growing based approaches start from an initial triangle mesh and merge adjacent planar triangles iteratively [16]. These methods mentioned have several disadvantages. For example, RANSAC might lead to non-compact planes e.g. containing points of a table and a wall, simultaneously, or the EM algorithm has to run every time when the number of planes was updated.

In the following, a combination of seeded region growing [17] and RANSAC is introduced based on special values holding the correlating arrangement of points. This planar patch extraction works directly on 3D points instead on triangles as proposed by Hähnel saving computation time needed for triangle generation. The main idea is to decompose the point cloud into planarly connected regions and to extract planes in these regions for refinement.

Oriented Particles. First, oriented particles similar to Fua's approach [18] are defined for each point. A point's normal $\boldsymbol{n}_c$ is computed using a point set $\{\boldsymbol{p}_i \mid \boldsymbol{p}_i \in \mathcal{N}_{3\times3} \text{ of } \boldsymbol{p}_c\}$ defined on the 8-neighborhood of the Swissranger image plane. The normal $\boldsymbol{n}_c$ of the current point $\boldsymbol{p}_c$ is determind by the principal component analysis of the points ($\{\boldsymbol{p}_i\} \in \mathcal{P}_c$). The deviation σ_c of the point $\boldsymbol{p}_c$ to the fitted plane can be used as a measure of the quality of the fit. Points with a deviation below a threshold θ_σ are classified as *locally planar*, otherwise as *nonplanar* [19]. Each plane $\{\mathcal{P}_c(\boldsymbol{n}_c, d_c)\}$ is described by the Hessian Normal Form

$$\mathcal{P}_c : \boldsymbol{n}_c \cdot \boldsymbol{x} - d_c = 0 \tag{1}$$

where d_c is equal to the Euclidean distance between the centroid $\boldsymbol{m}_c = \frac{1}{|\{\boldsymbol{p}_i\}|} \sum_i \boldsymbol{p}_i$ and the origin of the given world coordinate system.

Extracting Planar Patches. Here, the point cloud is decomposed into connected regions using region growing. Iteratively, points are selected randomly as seed of a region and extended with points of the 8-neighborhood $\mathcal{N}_{3\times3}$ if four criteria are fulfilled. Two criteria are defined on the particles themselves, which are the validation of the points generated by the preprocessing and the local planarity as defined above. The other two criteria are computed on pairs of particles like the conormality and coplanarity measurement defined by Stamos and Allen [19]. Two points $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are conormal, when their normals $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ hold:

$$\alpha = \cos^{-1}(\boldsymbol{n}_1 \cdot \boldsymbol{n}_2) < \theta_\alpha \tag{2}$$

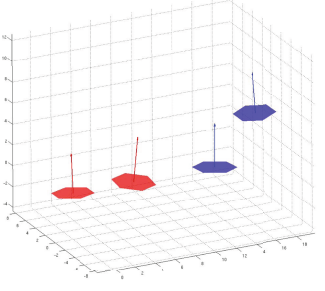

Fig. 2. (red) coplanar example of two patches, (blue) non coplanar pair of patches

Two points $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are coplanar (Fig. 2) when the distance d

$$d = \max(|\, \boldsymbol{r}_{12} \cdot \boldsymbol{n}_1 \,|, |\, \boldsymbol{r}_{12} \cdot \boldsymbol{n}_2 \,|),$$
$$\boldsymbol{r}_{12} = \boldsymbol{p}_1 - \boldsymbol{p}_2 \tag{3}$$

is smaller than a threshold θ_d. The distance is computed with respect to the orientation and the distance of the oriented particles. As a result a set of mainly planar connected patches is provided. On each of these regions several runs of the RANSAC algorithm extract the largest and smoothest planes. This step can be seen as a postprocessing step where basically the parameters of the planes $\{\mathcal{P}_c \mid \boldsymbol{n}_c, d_c\}$ are refined.

Merging Planar Patches. Due to oversegmentation neighboring planar patches which are close to each other (so-called *close* patches) and belong to the same infinite plane have to be merged. In order to realize an efficient merging strategy, first, a patch is chosen randomly and related patches fulfilling the angle condition (Eq. 2) are determined. These selected patches are forwarded to the next step, where around the current patch a region of interest (ROI) is determined. Patch

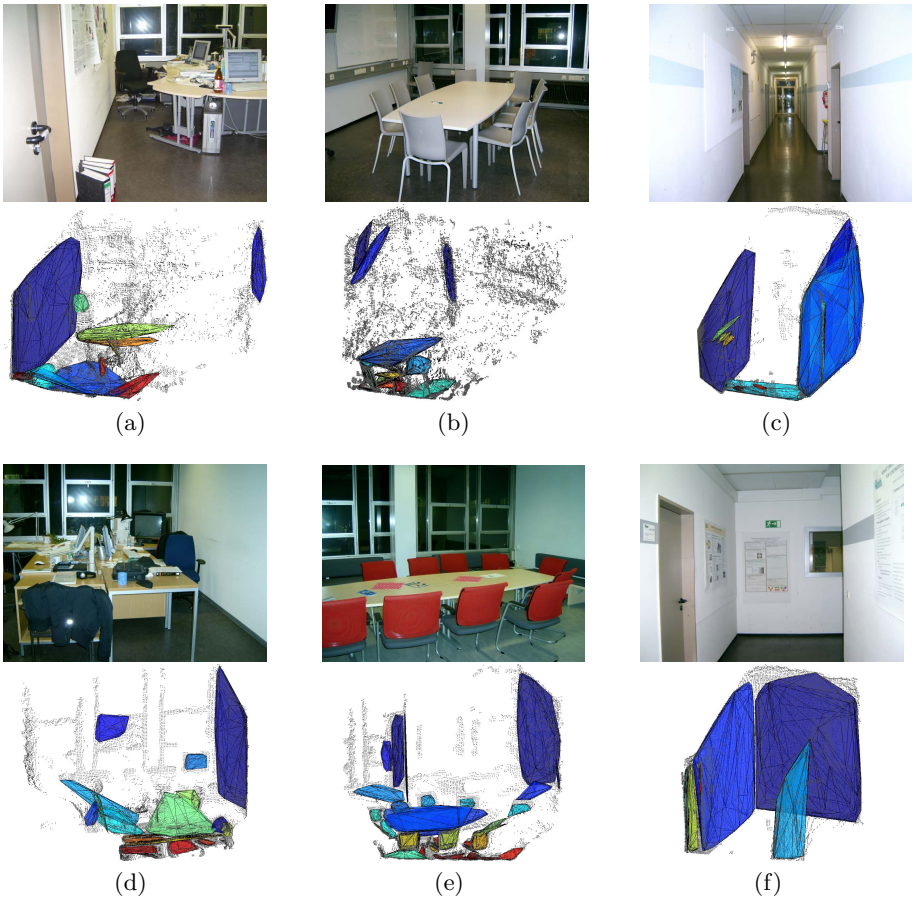

Fig. 3. First, exemplary photos and 3D point clouds of the training set are shown: the trained (a) office, (b) meeting room, and (c) hall. Second, the rooms for testing are displayed: the tested (d) office, (e) meeting room, and (f) hall.

candidates for merging are those patches which contain points that are inside this defined ROI. The candidates together with the current patch are merged to one plane containing all points of these patches. The parameters $(\boldsymbol{n}_c, d_c)$ of this new plane are recomputed using all 3D points. This merging procedure is repeated until the number of planes becomes stable.

Figure 3 presents exemplary photos of the six room scenarios – two office, two halls, and two meeting rooms – and planar patches produced by the algorithm introduced above using 3D point clouds provided by the Swissranger SR3000. The thresholds mentioned here were set to $\theta_\alpha = 10°$, $\theta_d = 0.2 \cdot z_c$ (the absolute Euclidean distances between neighboring pixel points vary over the distance to

the camera), and $\theta_\sigma = \bar{\sigma} + \sqrt{\frac{1}{n}\sum_{c=1}^{n}(\sigma_c - \bar{\sigma})^2}$ with n is the number of valid points per frame and $\bar{\sigma} = \frac{1}{n}\sum_{c=1}^{n}\sigma_c$ the mean deviation.

4 Feature Extraction

For classification an extraction of meaningful features from the given planes is required. The aim is to classify a perception of a room (here: one frame of the 3D ToF sensor) while e.g. a robot enters the room. The result of the classification should be a hypothesis which room type was entered, even if the robot has not seen this specific room before.

As well defined feature vectors have to fulfill several conditions it is not suitable to use all plane parameters merged to one vector as features for classification as proposed by Lourenco [20]. The features should not only be independent from colors and textures in the scene, which is implemented by the extracted planar structures, but they should also be invariant with respect to changes in the absolute number of planes, changes in view angle and view direction of the camera, and invariant to in-class variation of the furniture configuration. In the following, different aspects of the planar patches $\{\mathcal{P}_i\}$ are examined as first simple features for classification concerning the conditions listed above.

(i) *Number of Points per Patch.* First, a feature vector is computed that encodes the size of patches in a frame, e.g. whether it contains large patches or a lot of small planar structures: $\forall i : n_i = \frac{|\mathcal{P}_i|}{\sum_j |\mathcal{P}_j|}$. The resulting terms have values between zero and one with a concentration in the region close to zero. As a feature vector (FV1) a histogram is computed using bins of different size – small close to zero and becoming large towards one.

(ii) *Angles between Normals of Patches.* Here, the orientation between patches is considered: $\forall i \neq j : \alpha_{ij} = \cos^{-1}(\boldsymbol{n}_i \cdot \boldsymbol{n}_j)$ normalized by the maximal possible value which is $\frac{\pi}{2}$. The feature vector (FV2) is created as a histogram with five intervals uniformly distributed over the values between zero and one. In Section 5, it is shown that for classification it is sufficient to compute the median of these angles to encode their information. Both, histograms over number of points per patch and angles between pairs of patches do not contain any structural information about the rooms. This information can be introduced by computing the feature histogram (FV3) over the angles α'_{ij} between pairs of close patches leading to better classification results.

(iii) *Ratios between Sizes of Patches.* This feature (FV4) encodes whether a frame contains a lot of patches of similar or different size: $\forall i \neq j : r_{ij} = \frac{\min(|\mathcal{P}_i|,|\mathcal{P}_j|)}{\max(|\mathcal{P}_i|,|\mathcal{P}_j|)}$, while the feature vector (FV1) over the number of points per patch refers to the absolute sizes of the patches.

The values in the bins of the feature histograms (FV1, FV2, FV3, FV4) are normalized to the range $[0, 1]$ by dividing the entries by the sum over all values in the bins per histogram.

5 Experiments and Discussion

For the following experiments 300 frames of two different offices, two halls, and two meeting rooms were acquired. The camera was positioned at a height of 145 cm (robot's camera head) and rotated horizontally 30° left/right and vertically 10° up/down in order to simulate a more or less random view on the room while entering. Also, the rooms chosen have significant differences in the layout within a room type as shown in Figure 3. The office tested is mirrored concerning the arrangement of furniture compared to the trained one, the table of the meeting room tested is rotated about 90 degree compared to the trained one, and the hall tested is not a straight corridor but a corridor following a corner. One office, one meeting room, and one hall (Fig. 3(a), 3(b), 3(c)) form the training set where 270 frames per room are used to train the classifiers and the remaining 30 frames to test the quality of the performance in recognizing an already seen room. 300 frames per room of the three other rooms (Fig. 3(d), 3(e), 3(f)) form the *main* test set for examining the performance of our system in categorizing percepts of rooms which have not be seen before. We intentionally started with a very small training set containing a single room per category in order to show the generalizability of the learned model.

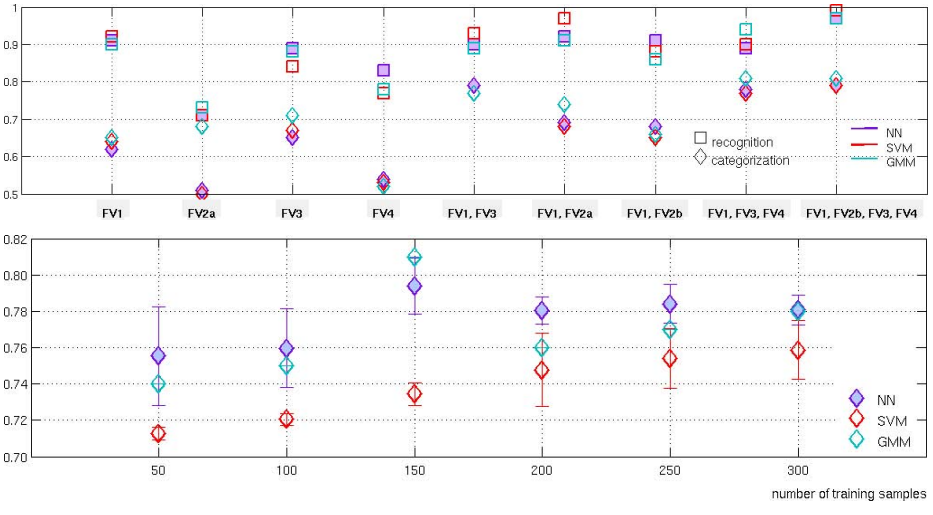

Fig. 4. (top) This plot presents results of the recognition and categorization using different combinations of the feature vectors (FV1, FV2, FV3, FV4: FV2a the histogram of angles between all patches, FV2b the median of these angles). Three classifiers are tested: a neuronal network (NN), a support vector machine (SVM), and a gaussian mixture model (GMM). (bottom) This plot shows the influence of the number of trainings data on the categorization result using [FV1, FV2b, FV3, FV4]. The vertical bars encode the standard deviation of the rates over 10 training runs of the NN, SVM or GMM with identical parameters (results using GMMs stay unchanged over 10 runs).

Three different classifiers are used to examine the proposed features in Section 4. The examined feature vectors are the number of points (FV1), the angles between patches (FV2a) and the median over these angles (FV2b), the angles between close patches (FV3), and the ratio of number of points between pairs of patches (FV4). The features are tested separately and in combination. A neural network (NN) with one hidden layer based on the Neural Network Toolbox of MatLab using backpropagation [21], the support vector machine SVM^{light} (SVM) [22,23] with a 5th-degree polynomial, and a gaussian mixture model (GMM) with five mixed distributions implemented in the toolkit ESMERALDA [24] are used for the classification task. Screening experiments provided five mixed distribution for GMM and a 5th-degree polynomial for SVM as quite suitable to deal with the proposed feature vectors.

Figure 4(top) presents all classification results from different feature vectors and combinations of them. The first four columns examine the feature vectors FV1, FV2a, FV3, and FV4 separated from each other. FV1 and FV3 turn out as features which contribute most to a good feature vector (recognition rate: 0.90, categorization rate: 0.65). The combination of these two features (FV1 and FV3) leads to a feature vector which provides promising categorization results up to 0.79 and recognition results up to 0.93. An additional test is executed to study the influence of FV2a compared to FV2b. FV2b performs similar to FV2a therefore it is assumed that the median of all angles can replace a histogram over all angles. The categorization can be improved up to 0.81 if the feature vector FV4 is added while the recognition rate stays on the level of 0.90. This rate can be further increased up to 0.99 using FV2b. As an assumption it can be stated that GMMs provide the most stable and proper classifiers using [FV1 FV2b FV3 FV4] as a feature vector. Round about 75% of the false classified vectors contains a mix up between meeting room and office. Since both room categories have analogies like a large table area in the middle of the room, this is an expected result.

Figure 4(bottom) shows the influence of the amount of training data on the classification rates. The vertical bars encode for each set of training data how reliable this classification rate is, since every trainings run with the same set of features and identical parameters leads to classifiers producing different classification rates on the same test data. Especially the NN and GMM classifiers seem to become saturated when using more than 150 training samples.

For additional experiments extra offices are recorded. Four of the now six different offices have a similar layout with two opposing work places while the other two rooms contain only a single work place. A categorization of the four double-place offices with the classifiers – trained with the train data introduced above – provides a rate of at least 0.69 right categorized percepts while only 0.34 to 0.51 of the percepts of the two single-place offices are classified correctly. If the train data is extended with frames of a single-place room the categorization rate of all offices can be increased to 0.88 on average.

Eighty percent of successful room categorization indicates that these planar structures extracted from the 3D point clouds provide meaningful information

about categories of rooms whereon feature vectors can be defined suitable for classification. The categorization only based on the given 3D data provides promising first results that may be even further improved via applying more different statistics to the set of planar patches, like e.g. histograms over the smallest distances between pairs of patches, number of close patches, the shapes of patches, or the convex hulls.

6 Conclusion and Outlook

In this paper an approach for classifying and categorizing rooms is introduced. This is especially challenging as individual components reappear in different room categories and vary in the same category with regard to color and texture. We propose a holistic classification scheme of room types using 3D spatial information. The approach is designed to work in a real world setting combining different innovative techniques in a whole processing chain from sensor to classification result. First, planar structures on 3D ToF data are extracted. Oriented particles for each point over its 8-neighborhood are defined. Afterwards using region growing the current region is extended by points that are planar, valid with regard to the preprocessing, conormal and coplanar. On the resulting connected planar regions RANSAC refines smooth planar patches while a merging step fuses close patches that belong to the same infinite plane. Several statistics over these patches are computed like number of points per patch, angles between all pairs of patches, angles between pairs of close patches, and ratios between sizes of patch pairs. Histograms over these statistics define feature vectors that show a good performance in categorizing room percepts of offices, halls, and meeting rooms.

We show that the features defined can be utilized for room categorization providing context information important for a mobile robot acting in a home tour scenario [25,26]. To cope with such a scenario, the next step is to apply the proposed feature extraction and classification to data of other room types – like living rooms, kitchens, and bedrooms which are typical rooms for this scenario. The aim is to combine room hypotheses on 3D data with hypotheses of other sensors like speech or 2D data. There, it might be necessary to compute a huge amount of simple features and to extract the best ones using feature selection techniques like AdaBoost. As our approach is purely data driven and represents a bottom-up description of rooms, a second interesting field for further research is to generate models of different room categories out of planar structures. These models could be triggered by the holistic classification result and support the interpretation as top-down world knowledge. It further can help to build up a scene model of the robot's environment.

Acknowledgement

This work is funded by the Cooperative Research Center CRC673 "Alignment in Communication".

References

1. Strat, T.M., Fischler, M.A.: Context-based vision: Recognizing objects using both 2D and 3D imaging. PAMI 13, 1050–1065 (1991)
2. McKeown, D.M., Harvey, W.A., McDermott, J.: Rule-based interpretation of aerial imagery. In: Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, pp. 415–430 (1987)
3. Murphy, K., Torralba, A., Freeman, W.T.: Using the forest to see the trees: A graphical model relating features, objects, and scenes. In: Advances in Neural Information Processing Systems, vol. 16 (2003)
4. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. CVPR 2, 2137–2144 (2006)
5. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: CAIVD, pp. 42–51 (1998)
6. Paek, S., Chang, S.F.: A knowledge engineering approach for image classification based on probabilistic reasoning systems. ICME 2, 1133–1136 (2000)
7. Torralba, A.: Contextual priming for object detection. IJCV 53, 153–167 (2003)
8. Weingarten, J., Gruener, G., Siegwart, R.: A state-of-the-art 3D sensor for robot navigation. In: IROS (2004)
9. Swadzba, A., Liu, B., Penne, J., Jesorsky, O., Kompe, R.: A comprehensive system for 3D modeling from range images acquired from a 3D ToF sensor. In: ICVS (2007)
10. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395 (1981)
11. Besl, P.J., McKay, N.D.: A method for registration of 3D shapes. PAMI 14, 239–256 (1992)
12. Nüchter, A., Surmann, H., Hertzberg, J.: Automatic model refinement for 3D reconstruction with mobile robots. In: 3DIM, pp. 394–401 (2003)
13. Lee, S., Jang, D., Kim, E., Hong, S., Han, J.: A real-time 3D workspace modeling with stereo camera. In: IROS, pp. 2140–2147 (2005)
14. Liu, Y., Emery, R., Chakrabarti, D., Burgard, W., Thurn, S.: Using EM to learn 3D models of indoor environments with mobile robots. In: ICML (2001)
15. Lakaemper, R., Latecki, L.J.: Using extended EM to segment planar structures in 3D. In: ICPR, pp. 1077–1082 (2006)
16. Hähnel, D., Burgard, W., Thrun, S.: Learning compact 3D models of indoor and outdoor environments with a mobile robot. Robotics and Autonomous Systems 44, 15–27 (2003)
17. Adams, R., Bischof, L.: Seeded region growing. PAMI 16, 641–647 (1994)
18. Fua, P.: From multiple stereo views to multiple 3D surfaces. IJCV 24, 19–35 (1997)
19. Stamos, I., Allen, P.K.: Geometry and texture recovery of scenes of large scale. CVIU 88, 94–118 (2002)
20. Lourenco, A., Freitas, P., Ribeiro, M.I., Marques, J.S.: Detection and classification of 3D moving objects. In: Mediterranean Conf. on Control and Automation (2002)
21. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. Parallel Data Processing 1, 318–362 (1986)
22. Vapnik, V.N.: The Nature of Statistical Learning Theory (1995)
23. Joachims, T.: Learning to Classify Text Using Support Vector Machines. PhD thesis, Cornell University (2002)

24. Fink, G.A.: Developing HMM-based recognizers with ESMERALDA. In: Matoušek, V., Mautner, P., Ocelíková, J., Sojka, P. (eds.) TSD 1999. LNCS (LNAI), vol. 1692, pp. 229–234. Springer, Heidelberg (1999)
25. Haasch, A., Hohenner, S., Hüwel, S., Kleinehagenbrock, M., Lang, S., Toptsis, I., Fink, G.A., Fritsch, J., Wrede, B., Sagerer, G.: BIRON – The Bielefeld Robot Companion. In: ASER, pp. 27–32 (2004)
26. COGNIRON: The cognitive robot companion (FP6-IST-002020) (2004), http://www.cogniron.org