Improved Mapping and Image Segmentation by Using Semantic Information to Link Aerial Images and Ground-Level Information

Martin Persson[1], Tom Duckett[2], and Achim Lilienthal[1]

[1] Centre of Applied Autonomous Sensor Systems, Örebro University, Sweden
`martin.persson@tech.oru.se, achim@lilienthals.de`
[2] Department of Computing and Informatics, University of Lincoln, UK
`tduckett@lincoln.ac.uk`

Summary. This paper investigates the use of semantic information to link ground-level occupancy maps and aerial images. A ground-level semantic map is obtained by a mobile robot equipped with an omnidirectional camera, differential GPS and a laser range finder. The mobile robot uses a virtual sensor for building detection (based on omnidirectional images) to compute the ground-level semantic map, which indicates the probability of the cells being occupied by the wall of a building. These wall estimates from a ground perspective are then matched with edges detected in an aerial image. The result is used to direct a region- and boundary-based segmentation algorithm for building detection in the aerial image. This approach addresses two difficulties simultaneously: 1) the range limitation of mobile robot sensors and 2) the difficulty of detecting buildings in monocular aerial images. With the suggested method building outlines can be detected faster than the mobile robot can explore the area by itself, giving the robot an ability to "see" around corners. At the same time, the approach can compensate for the absence of elevation data in segmentation of aerial images. Our experiments demonstrate that ground-level semantic information (wall estimates) allows to focus the segmentation of the aerial image to find buildings and produce a ground-level semantic map that covers a larger area than can be built using the onboard sensors.

1 Introduction

A mobile robot has a limited view of its environment. Mapping of the operational area is one way of enhancing this view for visited locations. In this paper we explore the possibility to use information extracted from aerial images to further improve the mapping process. Semantic information (classification of buildings versus non-buildings) is used as the link between the ground level information and the aerial image. The method allows to speed up exploration or planning in areas unknown to the robot.

Colour image segmentation is often used to extract information about buildings from an aerial image. However, it is hard to perform automatic detection of buildings in monocular aerial images without elevation information [15]. Buildings can not easily be separated from other man-made structures such as driveways, tennis courts, etc. due to the resemblance in colour and shape. We show that wall estimates found by a mobile robot can compensate for the absence of elevation data. In the approach proposed in this paper, wall estimates detected by a mobile robot are matched with edges extracted from an aerial image. A virtual sensor[3] for building detection is used to identify parts of an occupancy map that belong to buildings (wall estimate). To determine potential matches we use geo-referenced aerial images and an absolute positioning system on board of the robot. The matched lines are then used in region- and boundary-based segmentation of the aerial image for detection of buildings. The purpose is to detect building outlines faster than the mobile robot can explore the area by itself. Using a method like this, the robot can estimate the size of found buildings and using the building outline it can "see" around one or several corners without actually visiting the area. The method does not assume a perfectly up-to-date aerial image, in the sense that buildings may be missing although they are present in the aerial image, and vice versa. It is therefore possible to use globally available[4] geo-referenced images.

1.1 Related Work

Overhead images in combination with ground vehicles have been used in a number of applications. Oh *et al.* [10] used map data to bias a robot motion model in a Bayesian filter to areas with higher probability of robot presence. Mobile robot trajectories are more likely to follow paths in the map and using the map priors, GPS position errors due to reflections from buildings were compensated. This work assumed that the probable paths were known in the map. Pictorial information captured from a global perspective has been used for registration of sub-maps and subsequent loop-closing in SLAM [2].

Silver *et al.* [14] discuss registration of heterogeneous data (e.g. data recorded with different sampling density) from aerial surveys and the use of these data in classification of ground surface. Cost maps are produced that can be used in long range vehicle navigation. Scrapper *et al.* [13] used heterogeneous data from, e.g., maps and aerial surveys to construct a world model with semantic labels. This model was compared with vehicle sensor views providing a fast scene interpretation.

For detection of man-made objects in aerial images, lines and edges together with elevation data are the features that are used most often. Building

---

[3] A virtual sensor is understood as one or several physical sensors with a dedicated signal processing unit for recognition of real world concepts.

[4] E.g. Google Earth, Microsoft Virtual Earth, and satellite images from IKONOS and its successors.

detection in single monocular aerial images is very hard without additional elevation data [15]. Mayer's survey [8] describes some existing systems for building detection and concludes that scale, context and 3D structure were the three most important features to consider for object extraction, e.g., buildings, roads and vegetation, in aerial images. Fusion of SAR (Synthetic Aperture Radar) and aerial images has been employed for detection of building outlines [15]. The building location was established in the overhead SAR image, where walls from one side of buildings can be detected. The complete building outline was then found using edge detection in the aerial image. Parallel and perpendicular edges were considered and the method belongs to edge-only segmentation approaches. The main difference to our work regarding building detection is the use of a mobile robot on the ground and the additional roof homogeneity condition.

Combination of edge and region information for segmentation of aerial images has been suggested in several publications. Mueller *et al.* [9] presented a method to detect agricultural fields in satellite images. First, the most relevant edges were detected. These were then used to guide both the smoothing of the image and the following segmentation in the form of region growing. Freixenet *et al.* [4] investigated different methods for integrating region- and boundary-based segmentation, and also claim that this combination is the best approach.

1.2 Outline and Overview

The presentation of our proposed system is divided into three main parts. The first part, Sect. 2, concerns the estimation of walls by the mobile robot and edge detection in the aerial image. The wall estimates are extracted from a probabilistic semantic map. This map is basically an occupancy map built from range data and labelled using a virtual sensor for building detection [11] mounted on the mobile robot. The second part describes the matching of wall estimates from the mobile robot with the edges found in the aerial image. This procedure is described in Sect. 3. The third part presents the segmentation of an aerial image based on the matched lines, see Sect. 4. Details of the mobile robot, the experiments performed and the obtained result are found in Sect. 5. Finally, the paper is concluded in Sect. 6 and some suggestions for future work are given.

2 Wall Estimation

A major problem for building detection in aerial images is to decide which of the edges in the aerial image correspond to building outlines. The idea of our approach, to increase the probability that correct segmentation is performed, is to match wall estimates extracted from two perspectives. In this section we describe the process of extracting wall candidates, first from the mobile robot's perspective and then from aerial images.

2.1 Wall Candidates from Ground Perspective

The wall candidates from the ground perspective are extracted from a semantic map acquired by a mobile robot. The semantic map we use is a probabilistic occupancy grid map with two classes: buildings and non-buildings [12]. The probabilistic semantic map is produced using an algorithm that fuses different sensor modalities. In this paper, a range sensor is used to build an occupancy map, which is converted into a probabilistic semantic map using the output of a virtual sensor for building detection based on an omnidirectional camera.

The algorithm consists of two parts. First, a local semantic map is built using the occupancy map and the output from the virtual sensor. The virtual sensor uses the AdaBoost algorithm [5] to train a classifier that classifies close range monocular grey scale images taken by the mobile robot as buildings or non-buildings. The method combines different types of features such as edge orientation, grey level clustering, and corners into a system with high classification rate [11]. The classification by the virtual sensor is made for a whole image. However, the image may also contain parts that do not belong to the detected class, e.g., an image of a building might also include some vegetation such as a tree. Probabilities are assigned to the occupied cells that are within a sector representing the view of the virtual sensor. The size of the cell formations within the sector affects the probability values. Higher probabilities are given to larger parts of the view, assuming that larger parts are more likely to have caused the view's classification [12].

In the second step the local maps are used to update a global map using a Bayesian method. The result is a global semantic map that distinguishes between buildings and non-buildings. An example of a semantic map is given in Fig. 1. From the global semantic map, lines representing probable building outlines are extracted. An example of the extracted lines is given in Fig. 2.

2.2 Wall Candidates in Aerial Images

Edges extracted from an aerial image are used as potential building outlines. We limit the wall candidates used for matching in Sect. 3 to straight lines extracted from a colour aerial image taken from a nadir view. We use an output fusion method for the colour edge detection. The edge detection is performed separately on the three RGB-components using Canny's edge detector [1]. The resulting edge image I_e is calculated by fusing the three binary images obtained for the three colour components with a logical OR-function. Finally a thinning operation is performed to remove points that occur when edges appear slightly shifted in the different components. For line extraction in I_e an implementation by Peter Kovesi[5] was used. The lines extracted from the edges detected in the aerial image in Fig. 3, are shown in Fig. 4.

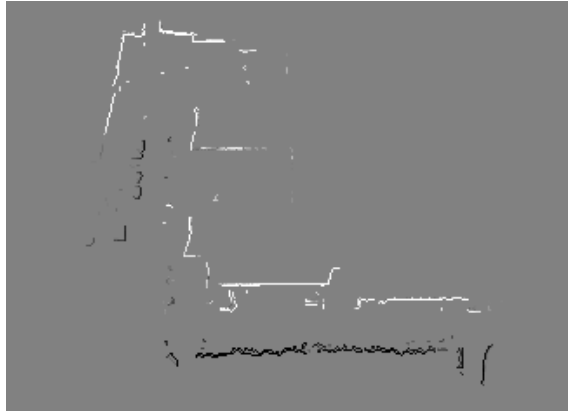[5] http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/, University of Western Australia, Sep 2005

Fig. 1. An example of a semantic map where white lines denote high probability of walls and dark lines show outlines of non-building entities
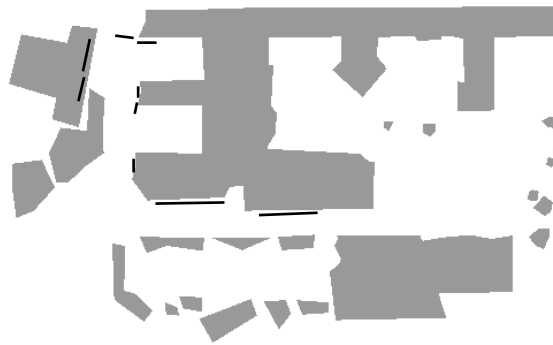

Fig. 2. Illustration of the wall estimates (*black lines*) calculated from the semantic map. The grey areas illustrate building and nature objects (manually extracted from Fig. 3). The semantic map in Fig. 1 belongs to the upper left part of this figure.

3 Wall Matching

The purpose of the wall matching step is to relate a wall estimate, obtained at ground-level with the mobile robot to the edges detected in the aerial image. In both cases the line segments represent the wall estimates. We denote a wall estimate found by the mobile robot as L_g and the N lines representing the
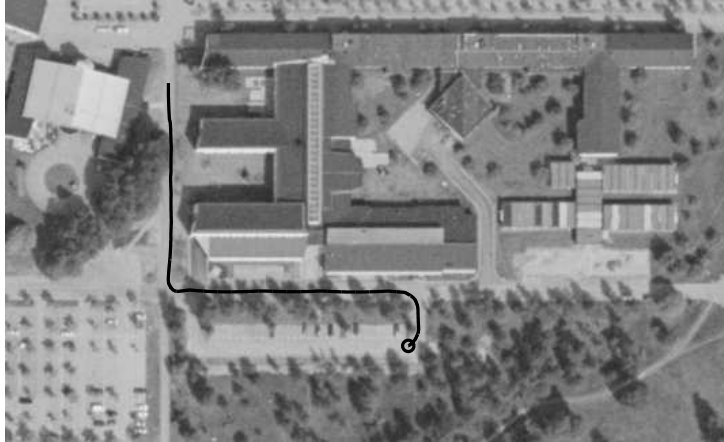
Fig. 3. The trajectory of the mobile robot (*black line*) and a grey scale version of the used aerial image
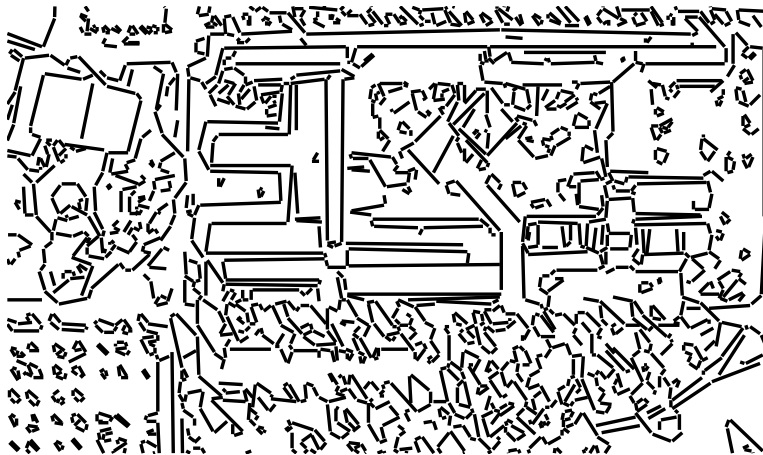

Fig. 4. The lines extracted from the edge version of the aerial image

edges found in the aerial image by L_a^i with $i \in \{1, \ldots, N\}$. Both line types are geo-referenced in the same Cartesian coordinate system.

The lines from both the aerial image and the semantic map may be erroneous, especially concerning the line endpoints, due to occlusion, errors in the semantic map, different sensor coverage, etc. We therefore need a metric
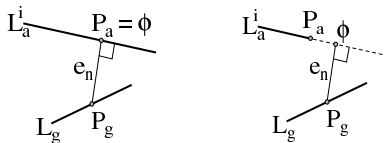
Fig. 5. The line L_g with its midpoint $P_g = (x_m, y_m)$, the line L_a^i, and the normal to L_a^i, e_n. To the left, $P_g = \phi$ since ϕ is on L_a^i and to the right, P_g is the endpoint of L_a^i since ϕ is not on L_a^i

for line-to-line distances that can handle partially occluded lines. We do not consider the length of the lines and restrict the line matching to the line directions and the distance between two points, one point on each line. The line matching calculations are performed in two sequential steps: 1) decide which points on the lines are to be matched, and 2) calculate a distance measure to find the best matches.

3.1 Finding the Closest Point

In this section we define which points on the lines are to be matched. For L_g we use the line midpoint, P_g. Due to the possible errors described above we assume that the point P_a on L_a^i that is closest to P_g is the best candidate to be used in our 'line distance metric'.

To calculate P_a, let e_n be the orthogonal line to L_a^i that intersects L_g in P_g, see Fig. 5. We denote the intersection between e_n and L_a^i as ϕ where $\phi = e_n \times L_a^i$ (using homogenous coordinates). The intersection ϕ may be outside the line segment L_a^i, see right part of Fig. 5. We therefore need to check if ϕ is within the endpoints and then set $P_a = \phi$. If ϕ is not within the endpoints, then P_a is set to the closest endpoint on L_a.

3.2 Distance Measure

The calculation of a distance measure is inspired by [7], which describes geometric line matching in images for stereo matching. We have reduced the complexity in those calculations to have fewer parameters that need to be determined and to exclude the line lengths. Matching is performed using L_g's midpoint P_g, the closest point P_a on L_a^i and the line directions, θ_i. First, a difference vector is calculated as

$$\mathbf{r_g} = [P_{g_x} - P_{a_x}, P_{g_y} - P_{a_y}, \theta_g - \theta_a]^T. \tag{1}$$

Second, the similarity is measured as the Mahalanobis distance

$$\mathbf{d_g} = \mathbf{r_g}^T \mathbf{R}^{-1} \mathbf{r_g} \tag{2}$$

where the diagonal covariance matrix $\mathbf{R}$ is defined as

$$\mathbf{R} = \begin{bmatrix} \sigma_{Rx}^2 & 0 & 0 \\ 0 & \sigma_{Ry}^2 & 0 \\ 0 & 0 & \sigma_{R\theta}^2 \end{bmatrix} \tag{3}$$

with σ_{Rx}, σ_{Ry}, and $\sigma_{R\theta}$ being the expected standard deviation of the errors between the ground-based and aerial-based wall estimates.

4 Aerial Image Segmentation

This section describes how local segmentation of the colour aerial image is performed. Segmentation methods can be divided into two groups; discontinuity- and similarity-based [6]. In our case we combine the two groups by first performing an edge based segmentation for detection of closed areas and then colour segmentation based on a small training area to confirm the areas' homogeneity. The following is a short description of the sequence that is performed for each line L_g:

1. Sort L_a^N based on $\mathbf{d_g}$ from (2) in increasing order and set $i = 0$.
2. Set $i = i + 1$.
3. Define a start area A_{start} on the side of L_a^i that is opposite to the robot (this will be in or closest to the unknown part of the occupancy grid map).
4. Check if A_{start} includes edge points (parts of edges in I_e). If yes, return to step 2.
5. Perform edge controlled segmentation.
6. Perform homogeneity test.

The segmentation based on L_g is stopped when a region has been found. Step 4 makes sure that the regions have a minimum width. Steps 5 and 6 are elaborated in the following paragraphs.

4.1 Edge Controlled Segmentation

Based on the edge image I_e constructed from the aerial image, we search for a closed area. Since there might be gaps in the edges bottlenecks need to be found [9]. We use morphological operations, with a 3×3 structuring element, to first dilate the interesting part of the edge image in order to close gaps and then search for a closed area on the side of the matched line that is opposite to the mobile robot. When this area has been found the area is dilated in order to compensate for the previous dilation of the edge image. The algorithm is illustrated in Fig. 6.
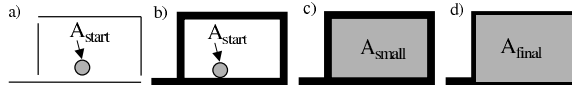
Fig. 6. Illustration of the edge-based algorithm. (**a**) shows a small part of I_e and A_{start}. In (**b**) I_e has been dilated and in (**c**) A_{small} has been found. (**d**) shows A_{final} as the dilation of A_{small}

4.2 Homogeneity Test

Classical region growing allows neighbouring pixels with properties according to the model to be added to the region. The model of the region can be continuously updated as the region grows. We started our implementation in this way but it turned out that the computation time of the method was quite high. Instead we use the initial starting area A_{start} as a training sample and evaluate the rest of the region based on the corresponding colour model. This means that the colour model does not gradually adapt to the growing region, but instead requires a homogeneous region on the complete roof part that is under investigation. Regions that gradually change colour or intensity, such as curved roofs, might then be rejected.

Gaussian Mixture Models, GMM, are popular for colour segmentation. Like Dahlkamp *et al.* [3] we tested both GMM and a model described by the mean and the covariance matrix in RGB colour space. We selected the mean/covariance model since it is faster and we noted that the mean/covariance model performs approximately equally well as the GMM in our case.

5 Experiments

5.1 Data Collection

The above presented algorithms have been implemented in Matlab for evaluation and currently work off-line. Data were collected with a mobile robot, a Pioneer P3-AT from ActivMedia, equipped with differential GPS, laser range scanner, cameras and odometry. The robot is equipped with two different types of cameras; an ordinary camera mounted on a PT-head and an omni-directional camera. The omni-directional camera gives a 360° view of the surroundings in one single shot. The camera itself is a standard consumer-grade SLR digital camera (Canon EOS350D, 8 megapixels). On top of the lens, a curved mirror from 0-360.com is mounted. From each omni-image we compute 8 (every 45°) planar views or sub-images with a horizontal field-of-view of 56°. These sub-images are the input to the virtual sensor. The images were taken with ca. 1.5 m interval and were stored together with the corresponding
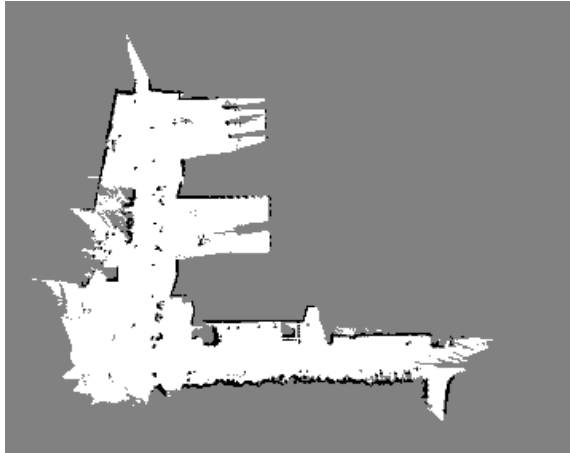
Fig. 7. Occupancy map used to build the semantic map presented in Fig. 1

robot's pose, estimated from GPS and odometry. The trajectory of the mobile robot is shown in Fig. 3.

5.2 Tests

The occupancy map in Fig. 7 was built using the horizontally mounted laser range scanner. The occupied cells in this map (marked in black) were labelled by the virtual sensor giving the semantic map presented in Fig. 1. The semantic map contains two classes: buildings (values above 0.5) and non-buildings (values below 0.5). From this semantic map we extracted the grid cells with a high probability of being a building (above 0.9) and converted them to the lines L_g^M presented in Fig. 2. Matching of these lines with the lines extracted from the aerial image L_a^N, see Fig. 4, was then performed. Finally, based on best line matches the segmentation was performed according to the description in Sect. 4.

In the experiments, the three parameters in **R** (3) were set to $\sigma_{Rx} = 1$ m, $\sigma_{Ry} = 1$ m, and $\sigma_{R\theta} = 0.2$ rad. Note that it is only the relation between the parameters that influences the line matching.

We have performed two different types of tests. *Tests 1-3* are the nominal cases when the collected data are used as they are. The tests intend to show the influence of a changed relation between σ_{Rx}, σ_{Ry} and $\sigma_{R\theta}$ by varying $\sigma_{R\theta}$. In *Test 2* $\sigma_{R\theta}$ is decreased by a factor of 2 and in *Test 3* $\sigma_{R\theta}$ is increased by a factor of 2. In *Tests 4* and *5* additional uncertainty (in addition to the uncertainty already present in L_g^M and L_a^N) was introduced. This uncertainty is in the form of Gaussian noise added to the midpoints (σ_x and σ_y) and directions σ_θ of L_g^M. The tests are defined in Table 1.

Table 1. Definition of tests and the used parameters

Test	σ_x [m]	σ_y [m]	σ_θ [rad]	$\sigma_{R\theta}$ [rad]	N_{run}
1	0	0	0	0.2	1
2	0	0	0	0.1	1
3	0	0	0	0.4	1
4	1	1	0.1	0.2	20
5	2	2	0.2	0.2	20

5.3 Quality Measure

We introduce two quality measures to be able to compare different algorithms or sets of parameters in an objective way. For this, four sets (A-D) are defined: A is the ground truth, a set of cells/points that has been manually classified as building; B is the set of cells that has been classified as building by the algorithm; C is the set of false positives, $C = B \setminus A$, the cells that have been classified as building B but do not belong to ground truth A; and D are the true positives, $D = B \cap A$, the cells that have been classified as building B and belong to ground truth A. Using these sets, two quality measures are calculated as:

- The true positive rate, $\Phi_{TP} = \#D/\#B$.
- The false positive rate, $\Phi_{FP} = \#C/\#B$.

where $\#D$ denotes the number of cells in D, etc.

5.4 Result

The results of *Test 1* show a high detection rate (96.5%) and a low false positive rate (3.5%), see Table 2. The resulting segmentation is presented in Fig. 8. Four deviations from an ideal result can be noted. At a and b tree tops are obstructing the wall edges in the aerial image, a white wall causes a gap between two regions at c, and a false area, to the left of b, originates from an error in the semantic map (a low hedge was marked as building).

The results of *Test 1-3* are very similar which indicate that the algorithm in this case was not specifically sensitive to the changes in $\sigma_{R\theta}$. In *Test 4* and *5* the scenario of *Test 1* was repeated using a Monte Carlo simulation with introduced pose uncertainty. The result is presented in Table 2. One can note that the difference between the nominal case and *Test 4* is very small. In *Test 5* where the additional uncertainties are higher the detection rate has decreased slightly.
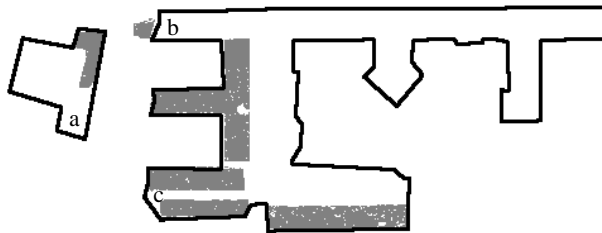
Fig. 8. The result of segmentation of the aerial image using the wall estimates in Fig. 2 (*grey*) and the ground truth building outlines (*black lines*)

Table 2. Results for the tests. The results of Test 4 and 5 are presented with the corresponding standard deviation

Test	Φ_{TP} [%]	Φ_{FP} [%]
1	96.5	3.5
2	97.0	3.0
3	96.5	3.5
4	96.8 $\pm$ 0.2	3.2 $\pm$ 0.2
5	95.9 $\pm$ 1.7	4.1 $\pm$ 1.7

6 Conclusions and Future Work

This paper discusses how semantic information obtained with a virtual sensor for building detection on a mobile robot can be used to link ground-level information to aerial images. This approach addresses two difficulties simultaneously: 1) buildings are hard to detect in aerial images without elevation data and 2) the range limitation of the sensors of mobile robots. Concerning the first difficulty the high classification rate obtained shows that the semantic information can be used to compensate for the absence of elevation data in aerial image segmentation. The benefit from the extended range of the robot's view can clearly be noted in the presented example. Although the roof structure in the example is quite complicated, the outline of large building parts can be extracted even though the mobile robot has only seen a minor part of the surrounding walls.

There are a few issues that should be noted:

- It turns out that we can seldom segment a complete building outline due to, e.g., different roof materials, different roof inclinations and additions on the roof.
- It is important to check several lines from the aerial image since the edges are not always as exact as expected. Roofs can have extensions in other

colours and not only roofs and ground are usually seen in the aerial image. In addition, when the nadir view is not perfect, walls appear in the image in conjunction with the roof outline. Such a wall will produce two edges in the aerial image, one where ground and wall meet and one where wall and roof meet.

6.1 Future Work

An extension to this work is to use the building estimates as training areas for colour segmentation in order to make a global search for buildings within the aerial image. Found regions would then have a lower probability until the mobile robot actually confirms that the region is a building outline.

The presented solution performs a local segmentation of the aerial image after each new line match. An alternative solution would be to first segment the whole aerial image and then confirm or reject the regions as the mobile robot finds new wall estimates.

As can be seen in the result, the building estimates can be parts of large buildings. It could therefore be advantageous to merge these regions. Another improvement would be to introduce a verification step that could include criteria such as:

- The building area should not cover ground that the outdoor robot has traversed.
- The size of the building estimate should exceed a minimum value (in relation to a minimum roof part).
- The found area should be checked using shadow detection to eliminate false building estimates.

References

1. J. Canny. A computational approach for edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(2):279–98, Nov 1986.
2. C. Chen and H. Wang. Large-scale loop-closing with pictorial matching. In *Proceedings of the 2006 IEEE International Conference on Robotics and Automation*, pages 1194–1199, Orlando, Florida, May 2006.
3. H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski. Self-supervised monocular road detection in desert terrain. In *Proceedings of Robotics: Science and Systems*, Cambridge, USA, June 2006.
4. J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi. Yet another survey on image segmentation: Region and boundary information integration. In *European Conference on Computer Vision*, volume III, pages 408–422, Copenhagen, Denmark, May 2002.
5. Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

6. R. C. Gonzales and R. E. Woods. *Digital Image Processing*. Prentice-Hall, 2002.
7. J. Guerrero and C. Sagüés. Robust line matching and estimate of homographies simultaneously. In *Pattern Recognition and Image Analysis: First Iberian Conference, IbPRIA 2003*, pages 297–307, Puerto de Andratx, Mallorca, Spain, June 2003.
8. H. Mayer. Automatic object extraction from aerial imagery – a survey focusing on buildings. *Computer vision and image understanding*, 74(2):138–149, May 1999.
9. M. Mueller, K. Segl, and H. Kaufmann. Edge- and region-based segmentation technique for the extraction of large, man-made objects in high-resolution satellite imagery. *Pattern Recognition*, 37:1621–1628, 2004.
10. S. M. Oh, S. Tariq, B. N. Walker, and F. Dellaert. Map-based priors for localization. In *IEEE/RSJ 2004 International Conference on Intelligent Robotics and Systems*, pages 2179–2184, Sendai, Japan, 2004.
11. M. Persson, T. Duckett, and A. Lilienthal. Virtual sensor for building detection by an outdoor mobile robot. In *Proceedings of the IROS 2006 workshop: From Sensors to Human Spatial Concepts*, pages 21–26, Beijing, China, Oct 2006.
12. M. Persson, T. Duckett, C. Valgren, and A. Lilienthal. Probabilistic semantic mapping with a virtual sensor for building/nature detection. In *The 7th IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA2007*, June 21-24 2007.
13. C. Scrapper, A. Takeuchi, T. Chang, T. H. Hong, and M. Shneier. Using a priori data for prediction and object recognition in an autonomous mobile vehicle. In G. R. Gerhart, C. M. Shoemaker, and D. W. Gage, editors, *Unmanned Ground Vehicle Technology V. Edited by Gerhart, Grant R.; Shoemaker, Charles M.; Gage, Douglas W. Proceedings of the SPIE, Volume 5083*, pages 414–418, Sept. 2003.
14. D. Silver, B. Sofman, N. Vandapel, J. A. Bagnell, and A. Stentz. Experimental analysis of overhead data processing to support long range navigation. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2443–2450, Beijing, China, Oct 9-15 2006.
15. F. Tupin and M. Roux. Detection of building outlines based on the fusion of SAR and optical features. *ISPRS Journal of Photogrammetry & Remote Sensing*, 58:71–82, 2003.