A model driven approach to extract buildings from multi-view aerial imagery

Luuk Spreeuwers, Klamer Schutte* and Zweitze Houkes

Laboratory for Measurement and Instrumentation, Dept. of Electrical Engineering
University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands
*TNO Physics and Electronics Laboratory, Netherlands

Abstract

This paper describes a system for analysis of aerial images of urban areas using multiple images from different viewpoints and its evaluation. The proposed approach combines bottom-up and top-down processing. In this paper the emphasis is on the discussion of the experimental evaluation. To evaluate statistically the performance of the system, a set of 100 realisations of 5 images from different viewpoints was used, which was generated by combining real and ray-traced images. The experiments show a significant improvement of reliability and accuracy if multi-view imagery is used instead of single-view.

1 Introduction

The goal of this research is to design and evaluate a system capable of analysing aerial photographs of urban arias. The output of this process is a 3-D scene description which can be used to update a GIS. Basically, the process involves the recognition of objects present in the scene and estimation of the parameters describing the objects: position, size and orientation. If the camera model and parameters are known, in most cases, the 3-D parameters of objects can be estimated from a single image. The obtainable accuracy is, however, highly dependent on the viewpoint. Furthermore, from certain viewpoints objects may be difficult to recognise, because parts of them are invisible or blend with the background. Also if objects occlude each other, it may be impossible to reliably recognise the imaged objects or to obtain reliable estimates of object parameters. Stereo vision provides a more robust estimation for the object parameters but does for the general case not solve the recognition or occlusion problem. In the presented work, therefore, multiple images are used recorded from different viewpoints.

2 A model based approach

The proposed method combines top-down and bottom-up techniques. Figure 1 depicts the basic setup of the system.

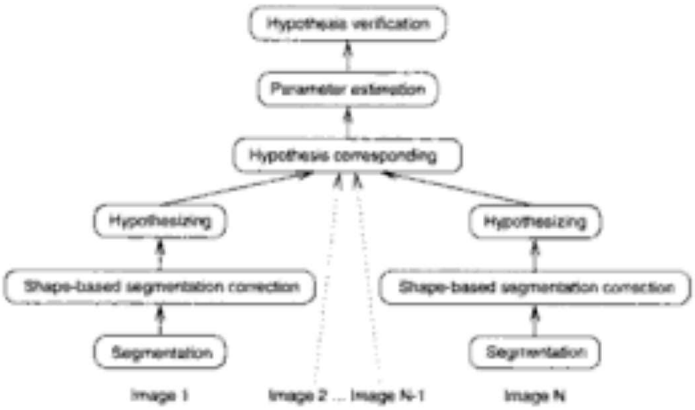


Fig. 1: Setup for the proposed system

The following six steps are distinguished:

Segmentation: region based segmentation of the images.

Shape-based segmentation correction: using knowledge about the expected shape of segments of man-made objects the segmentation is improved.

Hypothesising: using local evidence candidate scene descriptions are generated using a single image. For all images those scene descriptions are generated which have a sufficiently high likelihood.

Hypothesis corresponding: find out which hypotheses in the set of images correspond and thus refer to the same objects.

Parameter estimation: find the best set of parameters for all candidate scene descriptions using all the images, by predicting the segment shapes and selecting those parameters that result in the highest compatibility with the segmented images.

Hypothesis verification: Select from all candidate scene descriptions those that are most compatible with the measured images and do not contradict each other.

2.1 Segmentation

The segmentation process (Schutte 1994) consists of a region growing process (Schutte 1993) and a segmentation improvement step in which a priori shape knowledge is used. In the shape based segmentation correction we used a set of procedures for incorporating such knowledge into the segmentation process, similar to the rule bases proposed by Nazif and Levine (Nazif and Levine 1984). The shape knowledge used is based on the use of polyhedra to describe manmade objects. The projections of the polyhedra on the image plane are polygons. Also we use the fact that the polygons tend to have few corners and a certain minimum area.

2.2 Hypothesising: from regions to parametric object models

The input to the hypothesis generation is a description of the regions found in the image. Such a description is noisy, due to the nature of the images and the segmentation process. This means that some regions are found which do not correspond to visible object faces, and vice versa. The method should recognise the object, even if not all of the faces of the object correspond to a region. The hypotheses to be found consist of parametric object models. The models used are volumetric objects, such as a block, representing an office building, house etc. The output of the hypothesising method should include initial estimates needed by the parameter estimation procedure. Erroneous hypotheses generated by the hypothesis generator will be discarded by either the hypothesis corresponding or the final hypotheses verification process.

The hypothesising method consists of 4 steps. The first step (detection) comprises the extraction relational graphs from the segmentation. The second step is a relaxation process to find the best match with precalculated graphs of object models (*aspects*), stored in a database. Bipartite matching ensures unambiguity. In the last step the graph descriptions are transformed into parametric object models. A full description of the hypothesising method can be found in (Schutte and Boersema 1993).

The model data base, shown in figure 2 consists of the various objects which are of interest and can be expected in the scene. The objects currently defined are BlockShapedBuilding and House. For each object a set of aspects exists in the database.

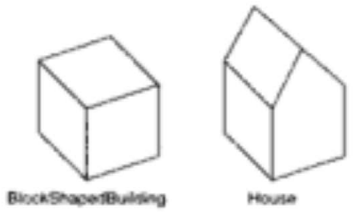


Fig. 2: The objects in the model database

2.3 Hypothesis correspondence in multiple view imagery

After the hypothesis generation stage on single images, for each image there is a list of hypotheses, containing for each hypothesis the object class and initial estimation of position, orientation and size parameters.

The objective of the hypothesis corresponding stage is firstly to find correspondences between hypotheses for the different images and reduce the total number of hypotheses by creating hypothesis groups with corresponding hypotheses. Secondly, unreliable hypotheses (e.g. that occur only in a single view) are discarded. Thirdly, not corresponding hypotheses that occupy the same space are marked *mutually exclusive*, since they cannot be valid simultaneously. Finally hypotheses that are *close* and do not correspond are marked, because they may cause occlusion. In order to determine whether two hypotheses i and j correspond, are close or mutually exclusive, three distance measures are defined:

$D(i,j)$ geometrical distance between the centres of gravity of the two hypotheses i and j

$O(i,j)$ measure of *overlap*, i.e. how much space is shared by the ground planes of the hypotheses i and j

$M(i,j)$ feature match quality, i.e. how well the hypothesised objects i and j resemble, taking into account: object class, size, orientation

Correspondence is defined as:

$$(O(i,j) \geq O_{min}) \textbf{ and } (M(i,j) \geq M_{min}) \tag{1}$$

so for correspondence there must be a certain minimum of overlap between the hypotheses and the hypotheses must resemble sufficiently. Two hypotheses are marked mutually exclusive if:

$$(O(i,j) \geq O_{min}) \textbf{ and } (M(i,j) < M_{min}) \tag{2}$$

i.e. the hypotheses occupy the same space, but do not resemble each other, hence it is impossible that both are correct. Finally two hypotheses are *close* if:

$$(D(i,j) < D_{max}) \tag{3}$$

In the above formulas 1-3, the constants O_{min}, M_{min} and D_{max} depend on (among others) the size of the buildings, the flight height and viewing angles.

Note that two hypotheses can at most have *one* of the above described relations: they *either* correspond *or* are exclusive *or* are close *or* have none of the relations.

2.4 Parameter estimation

The scene descriptions resulting from the hypothesis corresponding stage, consist of a list of hypothesis groups each with corresponding initial estimates of the parameters (position, size and orientation). For each hypothesis group the estimation process predicts the segments in all the images and adjusts the parameters for maximum compatibility with the segmented images. The setup of the estimation process is shown in fig.3.

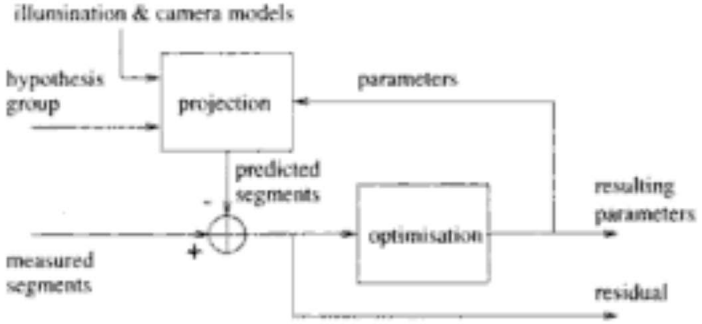
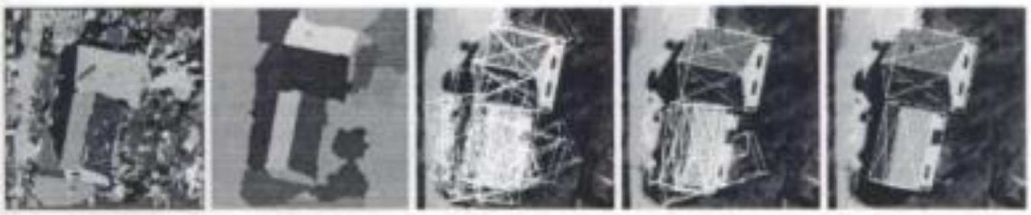


Fig. 3: Setup of the iterative parameter estimation process

stage for one realisation for all the 5 images, a total of 19 hypotheses, are projected on one image. Fig.5d shows the hypotheses groups that were formed by the hypothesis corresponding process. The minimum number of views an object must be observed by was set to 2. Six hypotheses groups remain after this stage: only a single hypothesis group for the top house and five competing hypotheses groups for the bottom house. Figure 5e shows the final result after estimation and verification. Since all five hypotheses groups of the bottom house were mutual exclusive, only one remained after verification.



a b c d e

Fig. 5: a) result after the split & merge step of the segmentation b) result after shape based segmentation correction; c) hypotheses generated by hypothesis generation step; d) hypothesis groups formed in hypothesis correspondence step; e) final result after estimation and verification.

3.3 Robustness and accuracy of single and multiple view approaches

In order to determine the reliability and accuracy of the single and multiple view approaches, the building recognition and estimation process was carried out for both approaches. First the recognition and estimation process was carried out for every realisation of each of the 5 views separately, thus using a single view. This yields 5 sets of 100 recognition and parameter estimation results. Next, the recognition and estimation process was carried out for the multiple view approach, combining the 5 views and again yielding a set of 100 recognition and estimation results.

To evaluate the performance of the multiple and single view approaches, the detections were classified in correct and spurious detections. A spurious detection is a hypothesis of which the parameters deviate too much from the actual parameters of the house. Only the position and orientation parameters were taken into account. A hypothesis is considered spurious if:

$$(x - \hat{x})^2 + (y - \hat{y})^2 + 250(\gamma - \hat{\gamma})^2 \geq 10 \tag{8}$$

This allows a maximum displacement of $\sqrt{10}$ [m] or a maximum orientation error of 0.2 [rad]. Based on occlusion and visibility, a prediction can be made about the expected performance for the different views. E.g. in fig.4a the walls are bright and no occlusion occurs, which should yield good results. The expected performances are summarised in table 1. The detection and spurious rates depend on the choice for the threshold for the fom, used in the verification stage to discard unreliable hypotheses. The detection rates and spurious rates were determined for a range of thresholds, see fig.6.

Clearly the multiple view approach results in significantly higher detection rates and lower spurious rates than any of the single view approaches. From the graphs it can also be seen

viewpoint	view dir.	fig.	comments	hb	ht
(-200,-200)	↖	4.a	no occlusion, walls light	+	+
(-200,200)	↗	4.b	hb occludes ht, walls dark	+/-	-
(200,-200)	↙	4.c	ht occludes hb, walls light	-	+
(200,200)	↘	4.d	no occlusion, walls dark	+/-	+/-
(0,0)	•	4.e	no occlusion, no walls visible	+/-	+/-
multiple	all			+	+

Tab. 1: Expected performance for different viewpoints. ht and hb are the top resp. bottom house.

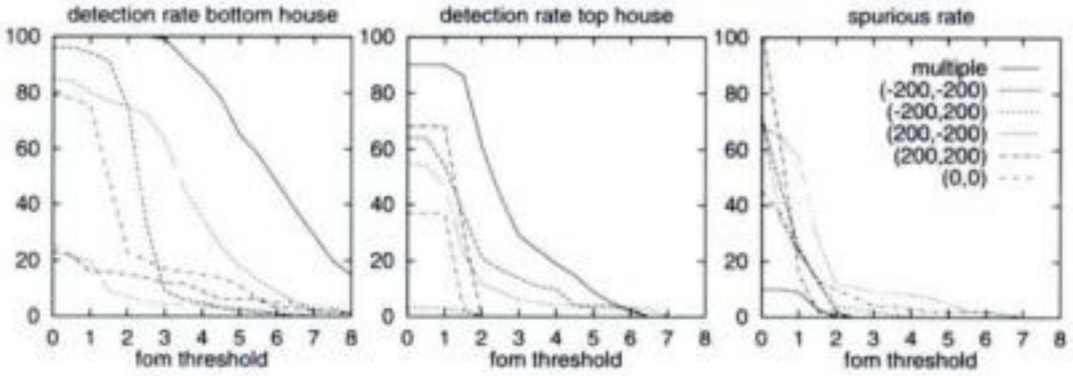


Fig. 6: Detection and spurious rates for a range of thresholds for the fom

that the view (-200,-200) yields the best results for the single views, while e.g. the view (-200,200) has a low detection rate for the bottom house as was expected from table 1. In general the graphs reflect the expectations from the table well. An optimal threshold of 1.8 was found for the multiple view approach. In this case no spurious detections were left and the detections rates for the bottom and top house were 100% resp. 78%. If no threshold is applied, the detection rate of the second house increases to 90%, but 10 spurious detections occur. None of the single views could approach these rates.

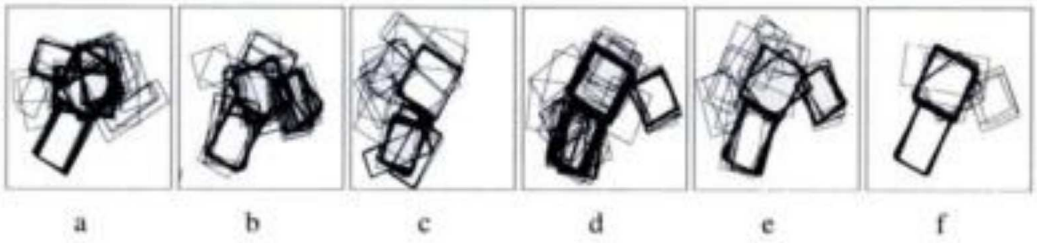


Fig. 7: Estimations of ground-planes for single views approaches: a) (-200,-200), b) (-200,200), c) (200,-200), d) (200,200), e) (0,0) and f) multi view

The difference in performance between multiple and single view approaches is also illustrated in figure 7, where the estimated ground-planes of the two houses for all 100 realisations are drawn in a single figure (no threshold on the fom is used here).

The accuracy of the estimated parameters is shown in tables 2 and 3. Here no threshold was applied to the fom to discard unreliable hypotheses and the spurious detections are not taken into account. The tables show the average and the standard deviations of the parameters of the correctly detected houses. It can be seen that the multiple view approach always yields accurate and often the most accurate results. Note that the multiple view approach also detects the difficult cases, which tend to have somewhat larger errors. For the single views, these are spurious detections and, hence, are not taken into account in the calculation of the averages and standard deviations. Also, for this evaluation the exact parameters are known. This is not the case in the operational situation, where only a threshold for the fom is available to discard unreliable hypotheses. As could be seen from fig.6, the multiple view approach in this case clearly outperforms all single view approaches.

view point	#det	$\overline{x}$	σ_x	$\overline{y}$	σ_y	$\overline{\gamma}$	σ_γ	$\overline{w}$	σ_w	$\overline{l}$	σ_l	$\overline{h}$	σ_h
true values	100	62.0		36.0		0.53		10.5		16.6		8.8	
(-200,-200)	96	62.2	0.3	36.1	0.3	0.51	0.02	10.1	0.5	16.3	0.2	8.3	0.8
(-200,200)	85	62.6	0.3	35.4	0.1	0.54	0.03	10.6	0.3	16.4	0.2	7.4	0.8
(200,-200)	26	61.7	0.3	36.7	0.3	0.47	0.06	10.3	0.1	16.6	0.2	7.1	1.2
(200,200)	23	61.8	0.4	35.7	0.3	0.50	0.06	11.1	0.6	16.5	0.5	7.6	1.5
(0,0)	79	62.0	0.2	35.8	0.2	0.45	0.06	10.4	0.3	16.5	0.3	9.4	1.1
multiple	100	62.1	0.1	35.9	0.1	0.48	0.03	10.3	0.2	16.5	0.2	8.8	0.4

Tab. 2: Average and standard deviations of estimated parameters of the bottom house. All parameters are in [m] except γ, which is in [rad].

view point	#det	$\overline{x}$	σ_x	$\overline{y}$	σ_y	$\overline{\gamma}$	σ_γ	$\overline{w}$	σ_w	$\overline{l}$	σ_l	$\overline{h}$	σ_h
true values	100	69.0		20.9		2.11		14.8		14.3		9.8	
(-200,-200)	64	69.2	0.3	21.1	0.2	2.09	0.01	15.3	0.5	14.0	0.1	9.1	0.9
(-200,200)	3	68.9	0.2	20.7	0.2	2.09	0.02	15.0	0.3	14.1	0.1	9.5	0.7
(200,-200)	55	68.8	0.1	20.9	0.2	2.11	0.01	15.3	0.4	14.1	0.1	9.8	0.6
(200,200)	68	68.8	0.3	20.7	0.5	2.10	0.01	15.1	0.4	14.2	0.2	9.0	2.4
(0,0)	37	68.8	0.2	20.8	0.1	2.09	0.01	15.2	0.5	14.4	0.2	8.1	3.2
multiple	90	68.9	0.1	20.9	0.3	2.06	0.03	15.3	0.4	14.1	0.2	9.3	1.0

Tab. 3: Average and standard deviations of estimated parameters of the top house. All parameters are in [m] except γ, which is in [rad].

4 Conclusions and Suggestions

A complete system for the recognition of buildings from multiple images is described and experimentally evaluated. The performance of the system was evaluated using a set of 100 artificially generated realisations of 5 images, acquired from different viewpoints, of a scene containing 2 houses. The experiments show that using multiple images drastically improves performance. Detection rates are improved considerably compared to a single

view approach. Simultaneously an increase in the accuracy of the estimated parameters is obtained. The improvement in the detection rates is the result of applying a hypothesis corresponding step, which removes incompatible hypotheses, generally caused by segmentation errors. The identifiability of the parameters is increased by using more and more independent data from multiple images, which results in an increased accuracy of the estimated parameters.

Acknowledgements

This work was supported by the Foundation for Computer Science in the Netherlands (SION) and the Dutch Organisation for Scientific Research (NWO). The authors wish to thank the Institute for Geodesy and Photogrammetry, Swiss Federal Institute of Technology (ETH) for making available photogrammetric test data (Mason et al. 1994).

References

Gill P. E., W. Murray, M. H. Wright (1981) *Practical Optimization*, Academic Press, London ISBN 0-12-283950-1.

Mason S., M. Baltsavias, D. Stallmann (1994) *High precision photometric data set for building reconstruction and terrain modelling*, Technical report, Institute for Geodesy and Photogrammetry, Swiss Federal Institute of Technology (ETH) Zurich, Switzerland.

Nazif A., M. Levine (1984) *Low level image segmentation: An expert system*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, No. 5, pp. 555–577.

Schutte K. (1993) *Region growing with planar facets*, Proceedings of The 8th Scandinavian Conference on Image Analysis, Vol. 2, Tromso, pp. 719–725.

Schutte K. (1994) *Knowledge Based Recognition of Man-Made Objects*, PhD thesis, University of Twente ISBN90-9006902-X.

Schutte K., G. Boersema (1993) *Hypothesizing a 3-D scene from a segmented aerial photograph*, Second Conference on Optical 3-D Measurement Techniques, Wichmann, Karlsruhe, Zurich, pp. 452–459.

Schutte K., G. Hilhorst (1993) *Comparison levels for iterative estimators for model-based recognition of man-made objects in remote sensing images*, Proc. IS&T/SPIE 1993 Symposium on Electronic Imaging: Science and Technology, Vol. 1904 of SPIE, San Jose, pp. 222–228.

Spreeuwers L. J., K. Schutte, Z. Houkes (1995) *A solution to the correspondence problem in multi-view imagery*, Proceedings of the Conference on Image and Signal Processing for Remote Sensing II, EUROPTO '95, Vol. 2579 of SPIE, Paris, France, pp. 274–284.