EVALUATION OF AUTOMATIC ROAD EXTRACTION

C. Heipke, H. Mayer, C. Wiedemann
Chair for Photogrammetry and Remote Sensing
Technische Universität München, D–80290 Munich, Germany
Tel: +49-89-2892-2676, Fax: +49-89-2809573
E-mail: {chris}{helmut}{wied}{albert}@photo.verm.tu-muenchen.de

O. Jamet[1]
I.G.N., Laboratoire MATIS
2, avenue Pasteur, 94160 Saint Mandé, France
E-mail: ojamet@teaser.fr

ABSTRACT

Internal self-diagnosis and external evaluation of the obtained results are essential for any automatic system. In the long run these factors are of major importance for the relevance of the system for practical applications. Obviously, this statement is also true for image analysis in photogrammetry and remote sensing. However, so far only relatively little work has been carried out in this area.
This paper deals with the external evaluation of automatic road extraction algorithms by means of comparison to manually plotted linear road axis used as reference data. The comparison is performed in two steps: (1) Matching of the extracted primitives to the reference network; (2) Calculation of quality measures. Each part depends on the other: the less tolerant is matching, the less exhaustive the extraction is considered to be, but the more accurate it looks. Therefore, the matching process is an important part of the evaluation process. The quality measures proposed for the automatically extracted road data comprise completeness, correctness, quality, redundancy, planimetric RMS differences and gap statistics. They aim at evaluating exhaustivity as well as assessing geometrical accuracy. The evaluation methodology is presented and discussed in detail. Results of a comparative evaluation of three different automatic road extraction approaches are presented. They show the overall status of the road extractors, as well as the individual strengths and weaknesses of each individual approach. Thus, the applicability of the evaluation method is proven.

1 INTRODUCTION

Internal self-diagnosis and external evaluation of the obtained results are essential for any automatic system. In the long run these factors are of major importance for the relevance of the system for practical applications. Obviously, this statement is also true for image analysis in photogrammetry and remote sensing. However, so far only relatively little work has been carried out in this area to date.
Both, internal self-diagnosis and external evaluation should yield quantitative results which are independent of a human observer. Internal self-diagnosis can be based upon the traffic light paradigm (Förstner, 1996): a green light stands for a result found to be correct as far as the diagnosis tool is concerned, a red light means an incorrect result, and a yellow light implies that further probing is necessary. External evaluation needs reference data of some sort and compares them to the automatically obtained results. In this paper we deal with the external evaluation of automatic road extraction algorithms by means of comparison to manually plotted linear road axes used as reference data.
A few approaches on evaluation of image analysis results can be found in the literature. In (McGlone and Shufelt, 1994) and (Hsieh, 1995) evaluation of automated building extraction is reported. The results of the extraction are pixels (in image space) or voxels (in object space) which are classified as “building” or “non-building”. The analysis of the degree of overlap between the results of the automated extraction and a manually generated reference is carried out by comparison of corresponding pixels or voxels, respectively. In (Guérin et al., 1995) road data from maps are analyzed with regard to distortions which are induced by the map production process. The evaluation is performed manually. First, the accuracy of the position of crossroads as well as the orientation of the connected roads, and their number and nature are investigated. Evaluation of the roads concentrates on measures for their geometrical accuracy. In (Airault et al., 1996) an evaluation methodology is tackled which is directed towards quantifying the benefits of automatic and semi-automatic road extraction algorithms. The proposed measures comprise geometric accuracy, success rate and particularly the capture time. In (Ruskoné, 1996) evaluation of a multi-phase automatic road extraction is performed to point out the benefits of the different phases as well as to quantify the quality of the overall results. The measures used are geometric accuracy as well as exhaustivity of the extracted data. In (CMU, 1997) evaluation is directed towards measuring the quality of (semi-)automatic road extraction with different levels of manual intervention. Only the exhaustivity of the extracted data is regarded.
In this paper an attempt is undertaken to fuse the quality measures discussed in the references cited above for automatic road extraction. In the next Section three different road extraction schemes, for which the evaluation is carried out, are shortly reviewed, and the proposed evaluation methodology is presented. Then, some implementation issues are discussed and in Section 4 the results of three different algorithms for automatic road extraction are presented and analyzed. The paper concludes with some final remarks and an outlook.

2 METHODOLOGY

This Section describes the generation of the road network data and the evaluation scheme together with a detailed description of the proposed quality measures.

[1] Most of this work was carried out, while Olivier Jamet was a visiting professor at Technische Universität München in autumn 1996

2.1 Road extraction

Basically, approaches for road extraction use one or both of the following two properties of roads: in low resolution images roads are usually modeled as lines, whereas in high resolution imagery they are considered as homogeneous, elongated areas with parallel roadsides. In this paper three different methods are used for the extraction of roads from digital imagery: Line extraction in low resolution by itself, an algorithm combining line extraction with a high resolution module based on grouping, and a third algorithm combining line extraction with a snake-based technique in high resolution. The two latter approaches use the results of the first one as input, and thus make use of the scale-space behavior of roads (Mayer and Steger, 1996). All three algorithms were developed at the Technische Universität München. Their choice for this study is motivated by the interest in the gain, a high resolution module yields in comparison with line extraction in low resolution. Also, the behavior of the two high resolution modules were to be compared in a controlled situation.

Line extraction (Steger, 1996) is based on differential geometry. Line points are characterized by having a local minimum or maximum in the direction perpendicular to the line. This direction is assumed to be the one in which the second derivative of the image function (i.e. the curvature) attains its maximum absolute value. To calculate the derivatives of the image function, the image is convolved with the corresponding derivatives of a Gaussian smoothing kernel with scale σ. The direction of maximum curvature is subsequently computed from the elements of the Hesse matrix, containing the second partial derivatives in row and column direction, and the mixed second derivative. The value of σ can be derived from the maximum line width which is to be extracted. The result of this step are individual line points with sub-pixel location, and directions. In a second step these points are linked into lines using a hysteresis threshold technique.

The TUM-G approach (Baumgartner et al., 1997) is based on the extraction of lines in an image of reduced resolution using the approach of (Steger, 1996) and the extraction of edges in a high resolution image. Using both resolution levels and explicit knowledge about roads, hypotheses for roadsides are generated. The roadsides are used to construct quadrilaterals representing road-parts and polygons representing intersections. Neighboring road-parts are chained into road-segments. Road-links, i.e., the roads between two intersections, are constructed by grouping of road-segments and closing of gaps between road-segments.

The TUM-S approach (Mayer et al., 1997) is based on the line extraction (Steger, 1996), too, and thus also takes advantage of the information of more than one scale resolution. Opposed to the TUM-G approach, it uses so-called "snakes" in the form of ribbon-snakes to verify roads and discriminate them from other line-type objects by means of the constancy of the width. What is more, the approach is able to bridge gaps between the lines, resulting e.g. from shadows cast on the roads. For the bridging ziplock-snakes (Neuenschwander et al., 1995), i.e., snakes which are optimized starting at their end points, proved to be important.

Global context or contextual information which describes so-called "outer characteristics" of roads, like "land cover area" and "bordered by" strongly influences the performance of road extraction , (Baumgartner et al., 1997), (Bordes et al., 1997). Using the sketched road models for all three approaches roads cannot be automatically extracted in highly textured areas such as forests or urban areas. Therefore, the extraction is restricted to open areas. The delineation of the open areas is carried out automatically by texture classification.

2.2 The evaluation scheme: matching procedure and quality measures

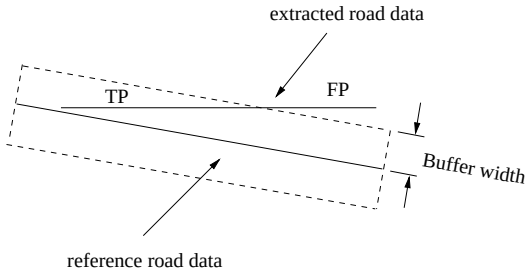
The evaluation of the extracted road data is made by comparison of the automatically extracted road centerlines with manually plotted road axes used as reference data and is processed in two steps: (1) Matching of the extracted road primitives to the reference network and (2) Calculation of quality measures. The proposed measures aim at assessing exhaustivity as well as geometrical accuracy of the results. Each part depends on the other: the less tolerant is matching, the less exhaustive the extraction is considered to be, but the more accurate it looks. Therefore, the matching process is an important part of the evaluation process.

2.2.1 Matching procedure The purpose of the matching is twofold: First, it yields those parts of the extracted data which are roads, i.e., match reference road data. Secondly, it shows which parts of the reference data are explained by the extracted data, i.e., match extracted road data, and which are not.

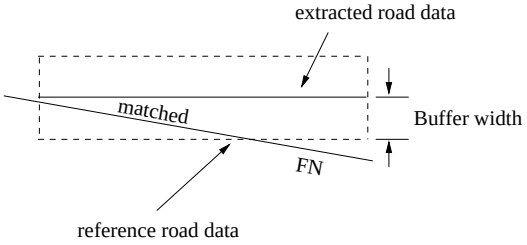
There are various ways to perform the actual matching of two road networks. Special issues arise from the fact that the topologies of the reference and the extracted network can be different, and that the extraction can be redundant, i.e., extracted pieces overlap each other. A simple matching process consists in the so called "buffer method", in which every portion of one network within a given distance from the other is considered as matched. This procedure is not satisfying in itself in the sense that a highly redundant extraction will not be detected. Another method consists in searching for a unique, i.e., bijective correspondence between the two networks. Such attempts have been made (Walter, 1996), however it is not clear how to define such a correspondence for topologically different networks on a general basis. As a consequence matching is performed according to the buffer method and additional attention is paid to the problem of redundancy. In the first step a buffer of constant predefined width (buffer width) is constructed around the reference road data. The parts of the extracted data within the buffer are considered as matched (see Fig. 1a). Following the notation of (McGlone and Shufelt, 1994) and (CMU, 1997) the matched extracted data are denoted as *true positive* with length TP emphasizing the fact that the extraction algorithm has indeed found a road, the unmatched extracted data is denoted as *false positive* with length FP, because the extracted road hypotheses are incorrect.

In the second step matching is performed in the other direction. The buffer is now constructed around the extracted road data, and the parts of the reference data lying in the buffer are considered to be matched (see Fig. 1b). In case of low redundancy their length can be approximated with TP (see above) The unmatched reference data are denoted as *false negative* with length FN.

2.2.2 Quality measures For the evaluation of the road extraction results a number of quality measures is defined. The measures are not meant to evaluate the extraction and the matching results in an absolute way. Rather, they are used to compare the results of different algorithms. This



(a) Matched extraction



(b) Matched reference

Figure 1: Matching principle.

view justifies a simplified set of measures.
Two questions are thought to be answered by means of the quality measures: (1) How complete is the extracted road network, and (2) How correct is the extracted network. The completeness corresponds to the user's demands ("what is missing in the network I want"), whereas the correctness is related to the probability of an extracted linear piece to be indeed a road. Thus, the correctness can be used within a self-diagnosis module.
The definitions of completeness and correctness, as well as the other quality measures are presented in the following.

- Completeness

$$\begin{aligned} completeness &= \frac{length\ of\ matched\ reference}{length\ of\ reference} \\ &\approx \frac{TP}{TP+FN} (for\ low\ redundancy) \\ completeness &\in [0;1] \end{aligned}$$

  The completeness is the percentage of the reference data which is explained by the extracted data, i.e., the percentage of the reference network which lies within the buffer around the extracted data.
  The optimum value for the completeness is 1.

- Correctness

$$\begin{aligned} correctness &= \frac{length\ of\ matched\ extraction}{length\ of\ extraction} \\ &= \frac{TP}{TP+FP} \\ correctness &\in [0;1] \end{aligned}$$

  The correctness represents the percentage of correctly extracted road data, i.e., the percentage of the extracted data which lie within the buffer around the reference network.
  The optimum value for the correctness is 1.

- Quality

$$\begin{aligned} quality &= \frac{length\ of\ matched\ extraction}{qq} \\ &= \frac{TP}{TP+FP+FN} \\ quality &\in [0;1] \\ qq &= length\ of\ extracted\ data \\ &\quad + length\ of\ unmatched\ reference \end{aligned}$$

  The quality is a measure of the "goodness" of the final result. It takes into account the completeness of the extracted data as well as its correctness.
  The optimum value for the quality is 1.

- Redundancy

$$\begin{aligned} redundancy &= \frac{rr}{length\ of\ matched\ extraction} \\ redundancy &\in [0;+\infty[\\ rr &= length\ of\ matched\ extraction \\ &\quad - length\ of\ matched\ reference \end{aligned}$$

  The redundancy represents the percentage to which the correct (matched) extraction is redundant, i.e., it overlaps itself.
  The optimum value for the redundancy is 0.

- RMS difference

$$\begin{aligned} RMS &= \sqrt{\frac{\sum_{i=1}^{l}(d(extr_i;ref)^2)}{l}} \\ RMS &\in [0;buffer\ width] \\ l &= number\ of\ pieces\ of\ matched\ extraction \\ d(extr_i;ref) &= shortest\ distance\ between\ the\ i-th\ piece\ of\ the\ matched\ extraction\ and\ the\ reference\ network \end{aligned}$$

  The RMS difference expresses the average distance between the extracted and the reference network, and thus the geometrical accuracy of the extracted road data. The value depends on the buffer width. If an equal distribution of the extracted road data within the buffer around the reference network is assumed, it can be shown that

$$RMS = \frac{1}{\sqrt{3}} * buffer\ width$$

  The optimum value for RMS is 0.

- Gap statistics
  - Number of gaps per kilometer

$$\begin{aligned} No.\ of\ gaps\ per\ km &= \frac{n}{ref.\ length\ [km]} \\ n &= number\ of\ gaps \end{aligned}$$

    The number of gaps in the reference data, i.e., the number of connected FN parts, is an indicator for the fragmentation of the extraction results. Only gaps larger than the approximate road width are taken into account, since the smaller gaps are usually closed during the extraction process. The optimum value is 0.
  - Mean gap length

$$\begin{aligned} \mu_{gap\ length} &= \frac{\sum_{i=1}^{n}(gl_i)}{n} \\ gl_i &= length\ of\ the\ i-th\ gap \end{aligned}$$

    The optimum value is 0.

2.3 Discussion of the evaluation scheme

There are several issues worth mentioning which may influence the significance of the proposed evaluation scheme and the accuracy of the results. They are mainly related to the employed matching strategy.

2.3.1 Buffer width The only parameter which has to be chosen for the evaluation process is the buffer width used for the matching. A suitable setting of the buffer width has to consider the expected internal accuracy of the road extraction algorithm. If the buffer width is set too large, false extractions close to the actual road will incorrectly be considered as roads. If it is chosen too small, correct road data which are only slightly geometrically inexact will be rejected.
For the evaluation of the results of the extraction algorithms described in Section 2.1 we have chosen a buffer width of approximately half of the road width, i.e., it is assumed that a road is extracted correctly if the road axis lies between the roadsides.

2.3.2 Direction There are extraction errors which cannot be detected during the matching because the direction of the road axes is not taken into account. Extracted road data can e.g. result in a scenario similar to the one displayed in Figure 2. Based on the simple matching criterion employed the extracted road data are matched to the reference data, although this result is obviously incorrect. The directions of the extracted and the reference roads differ significantly which is an indicator for the error. Based on our experience, however, these problems occur mostly in highly textured areas such as forests which are excluded from our investigations (see Section 2.1).

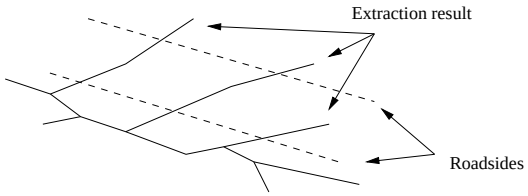


Figure 2: Incorrectly extracted data.

2.3.3 Shape In some cases the extraction results can wiggle around the reference road data as depicted in Figure 3. A shape measure for the extracted data, e.g based on curvature, can detect such problems. In the current implementation, however, no such measure has been included, because roads are implicitly or explicitly modeled as more or less straight segments in all 3 algorithms. Consequently, no wiggling effects have been observed in the extraction results.

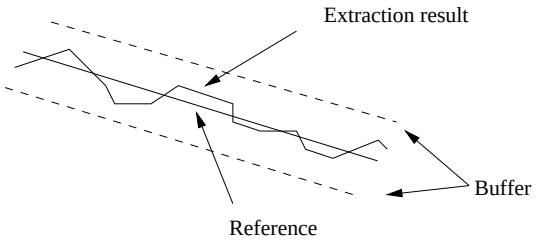


Figure 3: Extracted data wiggling around reference.

2.3.4 Road crossings and junctions In the present implementation of the evaluation approach roads and road crossings/junctions are handled in the same way. This may lead to some inaccuracies in the case of crossings/junctions. One such case is depicted in Figure 4. The reference data show a road junction. The extraction algorithms, however, only delivered the horizontal road, the branching road was not detected. Nevertheless, the part of this branching road lying inside the buffer of the extracted road is incorrectly considered to be matched.

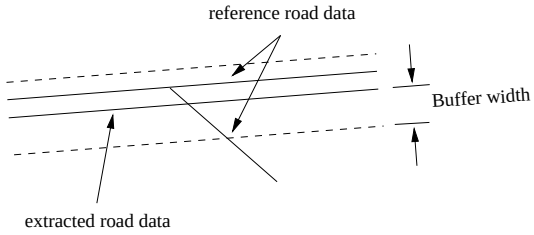


Figure 4: Matching in the vicinity of a junction.

The influence of this error on the accuracy of the final measures is considered to be small because of the small number of crossings and junctions compared to the overall length of the road network.

3 IMPLEMENTATION ISSUES

Due to resulting flexibility, simplicity and speed, matching is implemented in the raster domain using the $\mathcal{HORUS}$ image analysis system and is carried out pixel by pixel. It consists of the following steps:

- Vector/raster conversion of the extracted and the reference data,
- Calculation of the exhaustiveity measures:
  - Dilation of extracted and reference data (buffer generation)
  - Intersection of reference data with dilated extracted data, yielding the matched reference and unmatched reference data (FN)
  - Intersection of extracted data with dilated reference data, resulting in the matched extracted data (TP) and unmatched extracted data (FP)
  - Computation of completeness, correctness, quality and redundancy
- RMS Computation:
  - Distance transformation of the matched parts of the extracted data
  - Intersection of reference data with squared result of distance transformation
  - Calculation of the mean value of the intersection result. The square root of this mean value is the desired RMS difference.
- Gap statistics:
  - Determination of the number and length of connected FN sets and computation of mean gap length

There are two aspects to be discussed in more detail:

1. The dilation of the raster data and
2. the accuracy of the whole matching and the computed measures.

Ad 1: Buffer generation through conventional dilation leads to a result similar to the one shown in Figure 5. If this buffer is used for the matching, particularly for the intersection of reference and dilated extracted data, smaller gaps will be closed by the dilation. This lengthens the matched reference data and affects completeness, quality, redundancy, and the gap statistics. This problem is fixed by constructing the buffers from rectangles and circles (see Fig. 6). This method uses the vector representation as input. Around two nodes of the road network a rectangle is constructed with a width of twice the buffer width. A circle with radius equal to the buffer width is centered on each node which is connected to more than one other node. The resulting buffer is used for the matching.

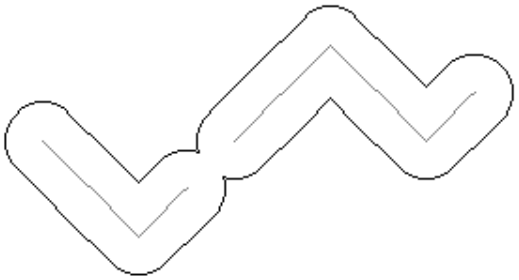


Figure 5: Buffer resulting from conventional dilation using a circle as structuring element.

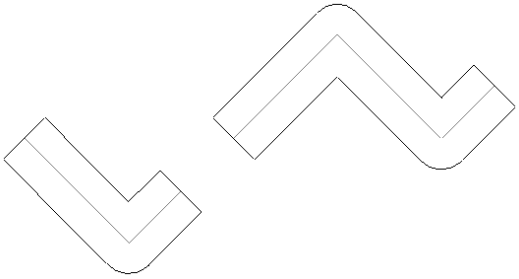


Figure 6: Outline of buffer constructed from rectangles and circles.

Ad 2: The accuracy of the matching and the quality measures is influenced by the pixel spacing used in the vector/raster conversion. Obviously, a higher accuracy is achieved if smaller pixels are chosen. The spacing used in this study was considerably smaller than the pixel size of the original grey value image (see also discussion on threshold parameters below). In this way, problems associated with the discretization such as directional effects in the computation of the lengths TP, FP, and FN, as well as an overly optimistic estimation for the RMS difference are kept to an acceptable minimum.

4 EXPERIMENTS AND RESULTS

The proposed evaluation scheme has been tested for the results of the three described algorithms and three different black and white test images. A description of the test images and the test procedure, the obtained results and a detailed analysis are presented in this Section.

4.1 Test images and test procedure

The test images are described in table 1 and are depicted below along with the extraction results. The image Marchetsreut with a groundel size (pixel size on the ground) of 0.225 m is a rather easy scene, Erquy (groundel size 0.45 m) and Montserrat (groundel size 0.225 m) are more difficult, because in some parts, the road model used for the extraction is violated. It should be noted that all three images are rather large, their size amounts to approximately 1/4 of a photogrammetric aerial image.
As described in Section 2.1 road extraction was only carried out in the open areas. The line algorithm needs low resolution images. They were generated by subsampling the test images to a pixel size 3.6 m. The results of the line algorithm were subsequently fed into the high resolution modules of the two other algorithms.
All three algorithms were run in a totally automatic fashion. It should be mentioned, however, that each algorithm requires a number of threshold parameters to be set prior to the computations. As mentioned earlier, the aim of the study is to evaluate the extraction results based on different test images. Information concerning the general applicability of each algorithm across different images is of secondary importance only. Therefore, in the case of the lines and the TUM-G algorithm it was thought acceptable to change the threshold parameters required for the extraction algorithms from one image to the next according to visual inspection. For the TUM-S algorithm this parameter tuning was not performed.
Also, the matching procedure needs some free parameters, namely the pixel spacing and the buffer width for matching. These parameters were chosen equal for all evaluations, and were set to 0.1 m for the pixel spacing, and 3 m for the buffer width, thus assuming a maximum road width of approximately 6 m.
The results are depicted in the Figures 7 through 9. For each test image the image superimposed with the forest mask, the manually plotted reference road axes, the three results delivered by the extraction algorithms, and a table listing the quality measures are shown.

4.2 Discussion of results

We first discuss the results of each algorithm separately, before giving some comments valid for the whole investigation.

4.2.1 Lines algorithm Judging from visual inspection the lines algorithm delivers the most detections. However, the extracted data are highly fragmented, i.e., there is a large number of small gaps, and many road hypotheses are incorrect. The reason for this behavior is the rather weak road model adopted: roads are assumed to be bright lines on a dark background, and there are hard constraints about the connectivity of line pixels (see Section 2.1). This model indeed fits many road parts, but also a lot of other linear structure in the image, and there is no further information to discriminate between the two groups. Consequently, the completeness is rather high, but the correctness and the quality have comparatively low values. Redundancy is not a problem for the algorithm due to the small buffer width of

Name of test site	description of image content	scale	pixel size in image space [μm]	groundel size in object space [m]	image size [pixel]	length of reference network [km]	buffer width [m]
Marchetsreut	flat, agricultural, easy	1:15,000	15	0.225	4000^2	3.84	3
Erquy	flat, agricultural, difficult	1:30,000	15	0.45	4500^2	24.10	3
Montserrat	hilly, agricultural, very difficult	1:15,000	15	0.225	4000^2	8.42	3

Table 1: Description of test images

3 m compared to the employed pixel size of 3.6 m used for the extraction. The geometrical accuracy is adequate considering this 3.6 m pixel size. These observations are valid for all three test images.
An amelioration of the results can only be obtained by introducing a stronger road model. Since the roads in low resolution are only a few pixels wide, information about the surrounding objects plays a major role in strengthening the model. Such models, however, are very difficult to realize.

4.2.2 TUM-G algorithm Due to the stronger road model and the high resolution image information, the TUM-G approach is able to discriminate much better between "roads" and "non-roads" than the lines algorithm, thus delivering more stable road hypotheses. These hypotheses are also used to bridge small gaps. Consequently, the extracted road parts are better connected and thus longer, the number of gaps is greatly reduced, and the average gap length is larger. Most important, the correctness is significantly larger. Some of the detected lines in the low resolution module of TUM-G are in fact roads but do not comply with the model criteria for the high resolution module (parallel road edges, homogeneous surface). Therefore, they are lost lost during processing. An example is the long road in the lower right of the Erquy image running at a 45 degree angle. Thus, the completeness is somewhat reduced. The resulting quality, however, is better than that of the lines algorithm, especially for Marchetsreut, but to a lesser extent also for Erquy and Montserrat.
The algorithm has a small problem with redundancy due to a known weakness: when extracting parallel roadsides, multiple parallel edges rather than only a pair of anti-parallel edges are allowed. The geometrical accuracy lies in between 2 and 4 pixels. Although this result is acceptable, there is room for improvement.
Whereas for Marchetsreut most parts of the road network have been extracted, the algorithm has problems with the more difficult scenes Erquy and especially Montserrat in which roads have a greatly varying appearance. This result is an indication for the applicability of the algorithm: it can serve as an automatic extraction tool for easy scenes. Improvements can be expected by analyzing the gaps in the extraction results, and finding and subsequently modeling the underlying reasons for these gaps such as shadows, occlusions etc.

4.2.3 TUM-S algorithm The TUM-S algorithm has been designed to overcome some of the problems of the TUM-G algorithm, and is especially aiming at generating very stable road hypotheses. For Marchetsreut the exhaustivity results are very similar, there is no redundant extraction, and sub-pixel accuracy was reached for the RMS difference. The advantages of the TUM-S algorithm are most obvious when inspecting the Erquy results: completeness and correctness are significantly improved due to the ability of the algorithm to analyze and bridge gaps in the extraction results, e.g. for shadowed or partly occluded areas. Especially striking is the fact that the completeness is higher than that of the lines algorithm, even though the long road running at the 45 degree angle was missed (see also TUM-G algorithm). This shows the effectiveness of the gap bridging. The limits of focusing the algorithm on generating stable road hypotheses become clear in the Montserrat scene: the winding roads with partly changing width and non-homogeneous surface are not handled correctly.
As in the case of the TUM-G approach improvements can be expected by advanced modeling of the roads and their surroundings.

4.2.4 Additional comments Looking at all the presented results it becomes clear that the proposed quality measures adequately capture the impression obtained when visually inspecting the extracted roads data. Thus, they can serve as a basis for the comparison of different automatic road extraction algorithms. It should be noted, however, that due to the effects mentioned in Section 2.3 and the discretization (see Section 3) the numerical accuracy of the quality measures is not extremely high. The significance of these measures for a detailed comparison can be further improved by classifying the reference road network into different categories such as clearly visible road parts, roads in shadow, occluded roads etc.
The number and mean length of the gaps needs some further discussion. Obviously the best result consists in having very few and short gaps. However, it is not clear whether a small number of long gaps is to be preferred to a large number of short gaps. The choice depends on the extra work necessary for closing the gaps. More detailed investigations are needed to clarify this issue.

5 SUMMARY AND CONCLUSIONS

Automatic evaluation of the obtained results is an increasingly important topic in image analysis. In this paper a methodology for the evaluation of automatic road extraction algorithms based on the comparison to manually plotted reference data as presented. This methodology was tested using the results of three extraction algorithms across three different test images.
The obtained results are representative for the state-of-the-art of automatic road extraction from aerial images. In easy scenes a completely automatic extraction is possible. As the scenes become more difficult the obtained results start

degrading. Low resolution images with a pixel size of a few meters can (and should) be used as a preprocessing step in the extraction. Reliable extraction only on the basis of these images, however, is not realistic. The key factor for improvement is a more detailed modeling of the roads and their surroundings.

The proposed evaluation scheme adequately captures the characteristics of the individual extraction results and can thus serve as a basis for their comparison and integration. Depending on the application at hand some of the quality measures such as completeness in a semi-automatic environment may be more relevant than others. Additional measures could be thought of, e.g. the ratio between completeness and correctness, which should remain constant over different images, once a suitable ratio has been found. Also, the algorithmic complexity and thus the computational effort needed will become a criteria as automatic road extraction advances further towards practical applications.

The proposed evaluation scheme can also form the basis for automatic updating of geo-data, which is becoming an increasingly important issue. In this case, the reference data are substituted by existing, but out-dated geo-data, and these are compared to extracted data from an up-to-date image. It should be noted that in this scenario some of the proposed measures such as the correctness and the RMS difference loose their significance, and a sound uncertainty management is needed because the assumption that the given vector data be correct and complete is no longer valid. This topic will be further investigated in future research.

6 ACKNOWLEDGMENT

We would like to express our thanks to Carsten Steger, Albert Baumgartner, and Ivan Laptev for contributing the software and the results of the three road extraction algorithms to this study.

REFERENCES

Airault, S., Jamet, O. and Leymarie, F., 1996. From Manual to Automatic Stereoplotting: Evaluation of Different Road Network Capture Processes. In: International Archives of Photogrammetry and Remote Sensing, Vol. **31**(3), International Society for Photogrammetry and Remote Sensing, pp. 14–18.

Baumgartner, A., Steger, C., Mayer, H. and Eckstein, W., 1997. Multi-Resolution, Semantic Objects, and Context for Road Extraction. In: Workshop on Semantic Modeling for the Acquisition of Topographic Information from Images and Maps, pp. 140–156.

Bordes, G., Giraudon, G. and Jamet, O., 1997. Road Modeling Based on a Cartographic Database for Aerial Image Interpretation. In: Semantic Modeling for the Acquisition of Topographic Information from Images and Maps, Birkhäuser Verlag, Basel, Switzerland, pp. 123–139.

CMU, 1997. Performance Evaluation for Feature Extraction. Slides presented at the Terrain Week 1997 (http://www.cs.cmu.edu/afs/cs/usr/maps/www/rcvw/terrainweek97/roads/tw97-roadeval.ROOT.html).

Förstner, W., 1996. 10 Pros and Cons Against Performance Characterization of Vision Algorithms. In: Workshop "Performance Characteristics of Vision Algorithms", Cambridge, p. 22.

Guérin, P., Jamet, O. and Maître, H., 1995. Distortion Model in Road Networks from Topographic Maps: identification and Assessment. In: SPIE: Integrating Photogrammetric Techniques with Scene Analysis and Machine Vision II, Vol. 2486, pp. 232–243.

Hsieh, Y., 1995. Design and Evaluation of a Semi-Automated Site Modeling System. Technical Report CMU-CS-995-195, Computer Science Department, Carnegie Mellone University.

Mayer, H. and Steger, C., 1996. A New Approach for Line Extraction and its Integration in a Multi-Scale, Multi-Abstraction-Level Road Extraction System. In: IAPR TC-7 Workshop: Mapping Buildings, Roads and other Man-Made Structures from Images, Oldenbourg Verlag, Vienna, Austria, pp. 331–348.

Mayer, H., Laptev, I., Baumgartner, A. and Steger, C., 1997. Automatic Road Extraction Based on Multiscale Modeling, Context, and Snakes. In: ISPRS Workshop on Theoretical and Practical Aspects of Surface Reconstruction and 3-D Object Extraction, Haifa, Israel, Sept. 9.–11.

McGlone, C. and Shufelt, J., 1994. Projective and Object Space Geometry for Monocular Building Extraction. In: Computer Vision and Pattern Recognition.

Neuenschwander, W., Fua, P., Székely, G. and Kübler, O., 1995. From Ziplock Snakes to Velcro^{tm} Surfaces. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images, Birkhäuser Verlag, Basel, Schweiz, pp. 105–114.

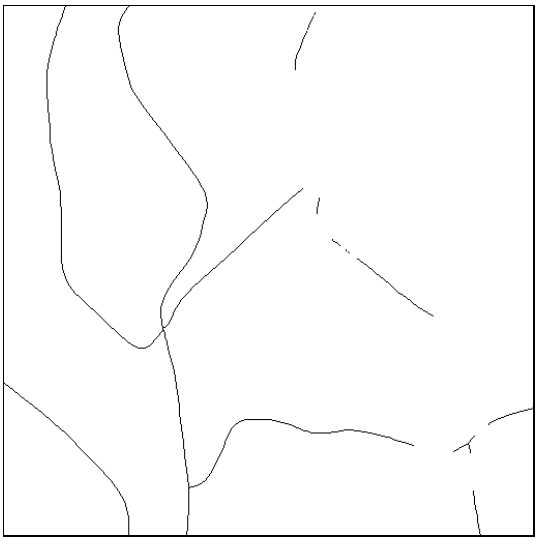
Ruskoné, R., 1996. Road Network Automatic Extraction by Local Context Interpretation: Application to the Production of Cartographic Data. PhD thesis, Université Marne-La-Vallée, France.

Steger, C., 1996. Extracting Curvilinear Structures: A Differential Geometric Approach. In: Fourth European Conference on Computer Vision, pp. 630–641.

Walter, V., 1996. Zuordnung von raumbezogenen Daten - am Beispiel der Datenmodelle ATKIS und GDF. PhD thesis, Fakultät fü Bauingenieur- und Vermessungswesen, Universität Stuttgart.

a) Grey value image superimposed with mask for open area.

d) Results of line extraction.

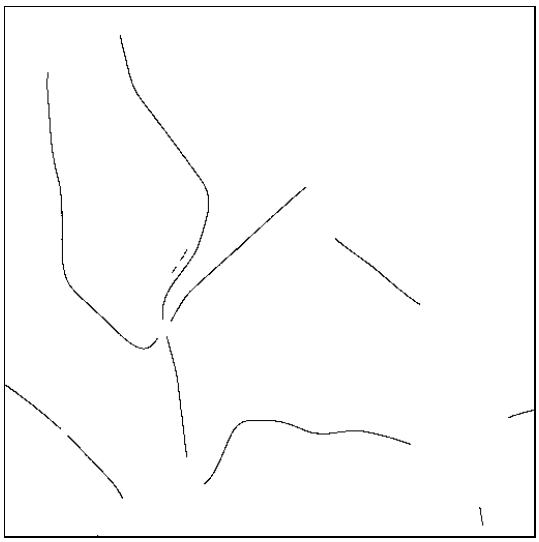


b) Manually plotted reference.

e) Results of TUM-G algorithm.

	Lines	TUM-G	TUM-S
Completeness	0.72	0.77	0.75
Correctness	0.42	0.95	0.97
Quality	0.36	0.76	0.74
Redundancy	0.05	0.11	0.01
RMS [m]	1.74	0.60	0.24
No. of gaps per km	29.2	6.8	10.4
$\mu_{gap\ length}$ [m]	9.7	33.7	23.7

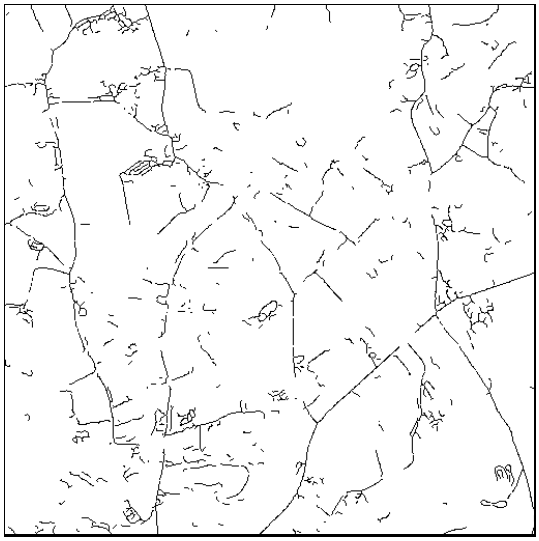
c) Quality measures.



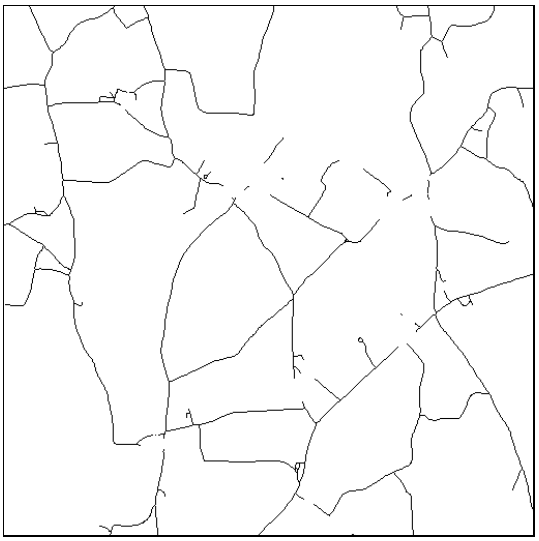
f) Results of TUM-S algorithm.

Figure 7: Test image Marchetsreut

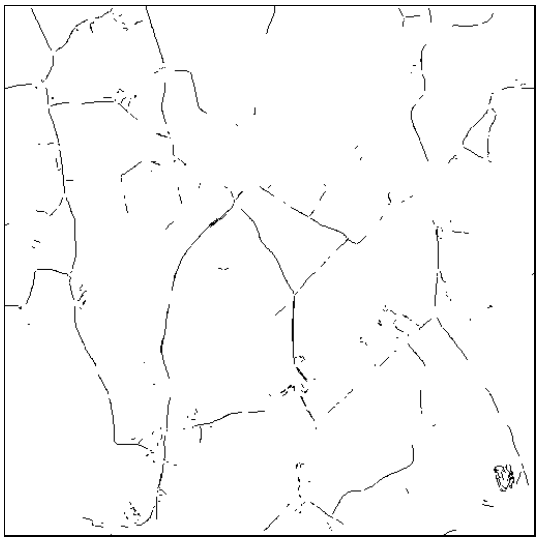
a) Grey value image superimposed with mask for open area.



d) Results of line extraction.



b) Manually plotted reference.



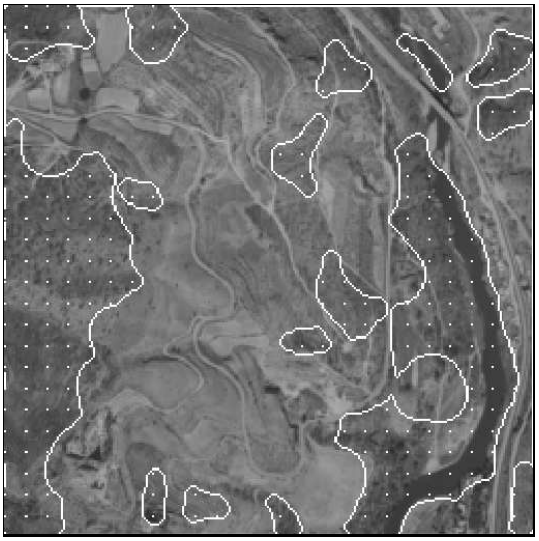
e) Results of TUM-G algorithm.

	Lines	TUM-G	TUM-S
Completeness	0.63	0.47	0.66
Correctness	0.42	0.78	0.87
Quality	0.34	0.42	0.60
Redundancy	0.05	0.04	0.01
RMS [m]	1.59	0.45	0.46
No. of gaps per km	22.6	7.8	9.0
$\mu_{gap\ length}$ [m]	16.3	68.4	37.7

c) Quality measures.

f) Results of TUM-S algorithm.

Figure 8: Test image Erquy

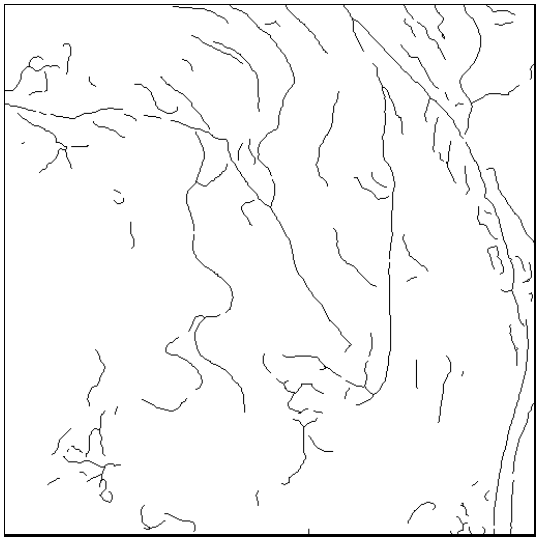


a) Grey value image superimposed with mask for open area.

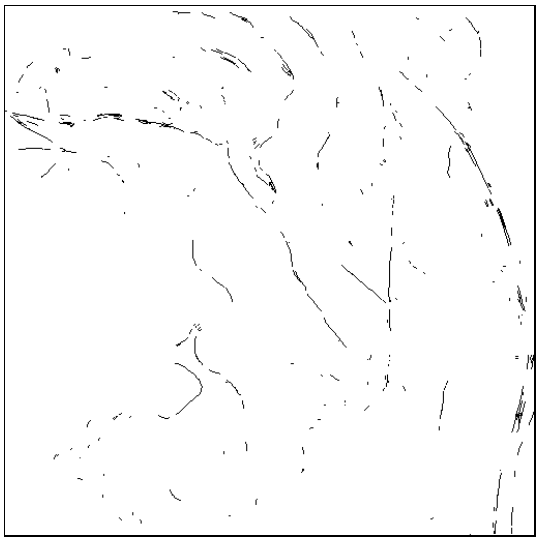
b) Manually plotted reference.

	Lines	TUM-G	TUM-S
Completeness	0.47	0.33	0.15
Correctness	0.36	0.61	0.55
Quality	0.26	0.29	0.13
Redundancy	0.03	0.09	0.01
RMS [m]	1.67	0.99	0.58
No. of gaps per km	21.3	12.0	4.3
$\mu_{gap\ length}$ [m]	25.0	55.9	199.1

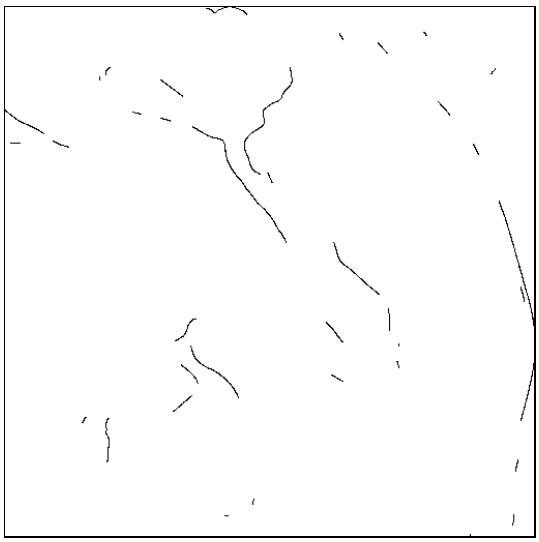
c) Quality measures.



d) Results of line extraction.



e) Results of TUM-G algorithm.



f) Results of TUM-S algorithm.

Figure 9: Test image Montserrat