# EMPIRICAL EVALUATION OF AUTOMATICALLY EXTRACTED ROAD AXES

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

Olivier Jamet
I.G.N., Laboratoire MATIS
2, avenue Pasteur, 94160 Saint Mandé, France
E-mail: Olivier.Jamet@ign.fr

**Abstract**

*Internal self-diagnosis and external evaluation of the obtained results are of major importance for the relevance of any automatic system for practical applications. Obviously, this statement is also true for automatic image analysis in photogrammetry and remote sensing. However, so far only relatively little work has been carried out in this area. This is mostly due to the moderate results achieved. Only recently automatic systems reached a state in which a systematic evaluation of the results seems to be meaningful.*

*This paper deals with the external evaluation of automatic road extraction algorithms by comparison to manually plotted linear road axes 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 step 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, matching 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 of major importance for the relevance of automatic systems for practical applications. Obviously, this statement is also true for automatic image analysis in photogrammetry and remote sensing. However, so far only relatively little work has been carried out in this area. This is mostly due to the moderate results achieved. Only recently automatic systems reached a state in which a systematic evaluation of the results seems to be meaningful.

Both, internal self-diagnosis and external evaluation should yield quantitative results which are independent of a human observer. A good description for the result of internal self-diagnosis is the traffic light paradigm [8]: 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 some kind of reference data 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.

Only few approaches on evaluation of image analysis results are found in the literature. In [15] and [11] the 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 degree of overlap between the results of the automated extraction and a manually generated reference is determined by matching of the corresponding pixels or voxels, respectively. Subsequently, measures for quantifying exhaustivity and correctness of the extraction result are calculated. Road data from maps are analyzed with regard to distortions which are induced by the map production process in [10]. A data set of the French Topographic Database (BDTopo) is used as reference. The comparison is performed manually. 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 [2] an evaluation methodology is proposed which is supposed to quantify the benefits of automatic and semi-automatic road extraction algorithms compared to manual data capture. The measures comprise geometric accuracy, success rate and in particular the time needed for data capture. [17] presents the evaluation of a multi-phase automatic road extraction. It points out the benefits of the different phases and quantifies the quality of the overall results. The reference data used is a data set of the BDTopo. Measures are geometric accuracy as well as exhaustivity of the extracted data. In [7] the evaluation is directed towards measuring the quality of (semi-)automatic road extraction with different levels of manual intervention. The reference data is generated by a procedure starting at manually selected positions, followed by automatic road tracking and manual editing. Roads are extracted as regions, and matching of the extracted data with the reference data is carried out using an intersection operation. Only the exhaustivity of the extracted data is further considered. [9] evaluates the effectiveness of different methods for the initialization of ribbon snakes as well as the geometric accuracy of the extracted road data. Manually generated road data serve as reference data. The evaluation focuses on the amount of effort needed by an operator which is measured by the number of necessary mouse actions. Measures for the geometric accuracy of the extracted road data are average and maximum deviation from the reference data.

This paper proposes and investigates a scheme for the evaluation of automatic road extraction. In this scheme various quality measures proposed in the literature are fused in a consistent manner. In the next Section three different road extraction schemes, for which the evaluation is carried out, are shortly reviewed. Section 3 is the main part of the paper. It presents the evaluation methodology in detail and discusses some implementation issues. In section 4 the results of the three different algorithms are presented and analyzed. The paper concludes with some final remarks and an outlook.

## 2: Road extraction

Basically, approaches for automatic road extraction use one or both of the following two properties: in low resolution imagery with pixel size of a few meters roads are modeled as lines, whereas

in high resolution imagery (pixel size in the dm range) they are considered as homogeneous, elongated areas with parallel roadsides. In this paper three different approaches for the extraction of roads from digital imagery are evaluated: The first two combine line extraction in low resolution with a high resolution module based on grouping (TUM-G), and a snake-based technique, respectively (TUM-S). Both approaches were developed at Technische Universität München (TUM) and make use of the scale-space behavior of roads [14]. The third approach relies on homogeneity tracking in high resolution imagery and was developed at Institut Géographique National (IGN).

The **TUM-G** approach [4] is based on lines in an image of reduced resolution using the approach of [19] and edges in a high resolution image. By combining both resolutions and introducing 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 [13] is based on lines [19], too. In the high resolution imagery it uses so-called "snakes" [12] in the form of ribbon-snakes to extract 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 or partial occlusions. For the bridging, ziplock-snakes [16], i.e., snakes which are optimized starting at their end points, are employed.

The **IGN** road extractor is based on semi-automatic road plotting [1]. The core process is a road tracker that follows homogeneous elongated areas from a given seed point. This tracking is performed by searching for the continuation of the road through the construction of a local tree of possible paths. The best path is selected according to a quality function depending on the curvature of the path and its average homogeneity calculated from grey value variances of the branches. The stop criterion for the search relies on the computation of local parallel borders and of the dispersion of the search tree. For this experiment, the algorithm was seeded by a road seed detector based on parallel borders in a region-based segmentation of the image [18]. In order to perform road extraction in a fully automatic manner, seed detection was performed only within a buffer around the (approximately) known road position.

Global context and contextual information which describes so-called "outer characteristics" of roads, like "land cover area" and "bordered by" strongly influence the performance of road extraction [3, 5]. All three approaches cannot automatically extract roads in highly textured areas such as forests or urban areas due to their simplified models for roads. Therefore, the extraction is restricted to open areas in all cases. The delineation of the open areas is carried out automatically by texture classification.

## 3: Evaluation scheme

The evaluation of the extracted road data is carried out by comparing the automatically extracted road centerlines with manually plotted road axes used as reference data. Both data sets are given in vector representation. The evaluation is processed in two steps: (1) Matching of the extracted road primitives to the reference network and (2) Calculation of quality measures.

The proposed quality measures aim at assessing exhaustivity as well as geometrical accuracy. Each step 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, matching is an important part of the evaluation process.

The quality measures address two questions: (1) How complete is the extracted road network, and (2) How correct is it. The completeness corresponds to the user's demands: ("how much

is missing in the network"). The correctness, on the other hand is related to the probability of an extracted linear piece to be indeed a road. Thus, it is of high importance for a self-diagnosis module.

### 3.1: Matching procedure

The purpose of the matching is twofold: Firstly, it yields those parts of the extracted data which are supposed to be roads, i.e., which correspond to the reference road data. Secondly, it shows which parts of the reference data are explained by the extracted data, i.e., which correspond to the extracted road data.

There are various ways to perform the actual matching of two networks. Especially if the geometric distortions are large and not known beforehand, relational matching was used successfully [20, 6]. 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. The so called "buffer method", is a simple matching procedure in which every portion of one network within a given distance from the other network is considered as matched. The matching is not affected by different network topologies. The drawback of this procedure is that a highly redundant extraction will not be detected and that direction differences between parts of the two networks are not taken into account. Yet another method for matching consists in searching for a unique, i.e., bijective correspondence between the two networks. Such attempts have been made [21], however, it is not clear how to define such a correspondence for topologically different networks on a general basis.

In our case, there are very good approximations for position and orientation of the road data to be matched. As a consequence, matching is performed according to the buffer method and additional attention is paid to the problem of redundancy and direction differences.

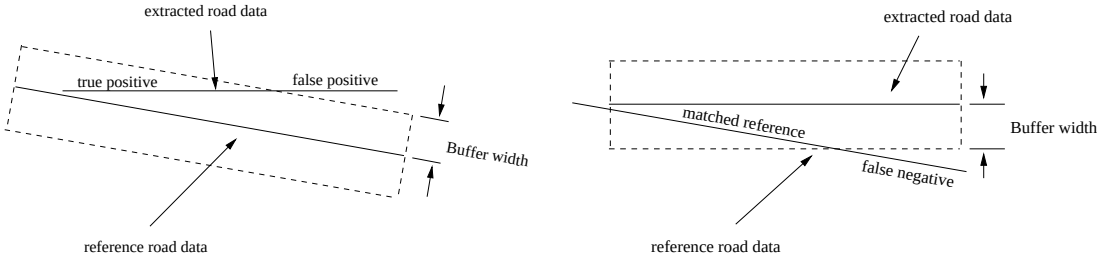
In the first step both networks are split into short pieces of equal length. Then, a buffer of constant predefined width (buffer width) is constructed around the reference road data (see Fig. 1a). The parts of the extracted data within the buffer are considered as matched if the direction difference between the reference road data and the part to be matched does not exceed a given threshold. The direction difference can be derived directly from the vector representations of both networks. Following the notation of [15] and [7] 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 the other way round. The buffer is now constructed around the extracted road data (see Fig. 1b), and the parts of the reference data lying in the buffer and fulfilling the direction constraint are considered as matched. In case of low redundancy their length can be approximated with TP. The unmatched reference data are denoted as *false negative* with length FN.

### 3.2: Quality measures

The quality measures for road extraction are intended to compare the results of different algorithms, rather than to evaluate the extraction and the matching results in an absolute way. Because these results are additionally quite different and still relatively far away from a perfect solution, a simplified set of measures is justified.

The definitions of the quality measures are presented in the following.




(a) Matched extraction (b) Matched reference

**Figure 1. Matching principle**

- 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 lie 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

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

  The quality is a more general measure of the final result combining completeness and correctness into a single measure:

$$quality = \frac{completeness * correctness}{completeness - completeness * correctness + correctness}$$

  The optimum value for the quality is 1.

- Redundancy

$$redundancy = \frac{length\ of\ matched\ extraction - length\ of\ matched\ reference}{length\ of\ matched\ extraction}$$
$$redundancy \in\ ]-\infty; +\infty[$$

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

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

The RMS difference expresses the average distance between the matched extracted and the matched reference network, and thus the geometrical accuracy potential 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

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

    The number of gaps in the reference data, i.e., the number of connected *false negative* parts, is an indicator for the fragmentation of the extraction results.
    The optimum value for the number of gaps per kilometer is 0.
  - Mean gap length

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

    The optimum value for the mean gap length is 0.

Note that the completeness can be calculated from the number of gaps per kilometer and the mean gap length as follows:

$$completeness = 1 - (No.\ of\ gaps\ per\ km * \mu_{gap\ length}/1000)$$

### 3.3: 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. Firstly, the two parameters which have to be chosen for the evaluation process as such, namely the buffer width and the maximum direction difference, are explained. Secondly, a property of extracted road data which may influence the result of the evaluation is discussed.

Buffer width: 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 approaches described in Section 2, a buffer width of approximately half of the road width was chosen, i.e., it is assumed that a road is extracted correctly if the road axis lies between the roadsides.

Maximum direction difference: Some errors will not be detected, if the direction of the road axes is not taken into account during matching. Extracted road data can, e.g., result in a scenario similar to the one displayed in Figure 2. Without considering the direction constraint, the extracted road data within the buffer would be 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. This problem is solved by investigation of the direction difference of the two pieces to be matched. If it is larger than a threshold (maximum direction difference) no match between these pieces is established. Nevertheless, it is possible that matches exist to other pieces further away but with smaller direction differences.

The maximum direction difference should not be chosen too restrictive because of the direction uncertainty of short line segments, especially if they are an approximation of highly curved lines.

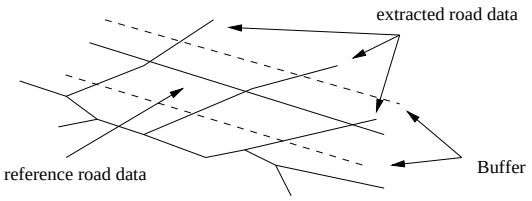



**Figure 2. Incorrectly extracted road data**

Shape: In some cases the results can wiggle around the reference 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 three approaches. Consequently, no such effects have been observed in the extracted data. It should be pointed out that a shape measure becomes important as soon as generalized road axes are used as reference data.

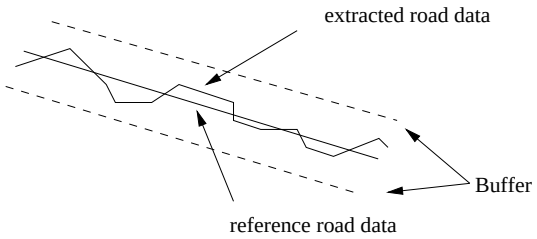



**Figure 3. Extracted road data wiggling around the reference**

### 3.4: Implementation issues

Both, the extracted data and the reference data are introduced into the evaluation procedure in vector representation, each as a set of polygons. Both networks are split into pieces of equal length (split length). Then, the shortest distance between each piece of one network and the other network data is calculated as explained below. The resulting distances are assigned to the respective pieces. Each piece is labeled as *matched* or *unmatched* based upon a check if its distance value is below the buffer width and if the direction constraint is fulfilled.

Because of the known and equal length of the pieces (split length) the length of the matched and unmatched network data can be easily computed by multiplication of the number of *matched*/*unmatched* pieces with the split length. This yields the length of the unmatched reference data (*FN*), the length of the matched extracted data (*TP*), and the length of the unmatched extracted data (*FP*). From these values the quality measures completeness, correctness, quality, and redundancy are computed using the formulas given in Section 3.2. The RMS difference is calculated from the distances assigned to the matched pieces of the extracted data. For the determination of the number of gaps per km and the mean gap length, the connectivity of the unmatched pieces of the reference data is analyzed.

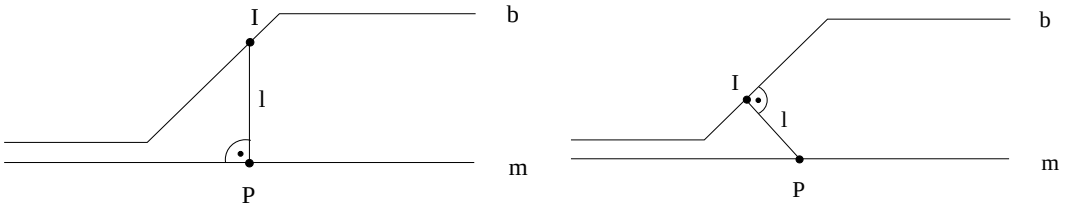
There are three aspects to be discussed in more detail:

1. The definition of the shortest distance
2. The accuracy of the whole matching and the computed measures.

**Ad 1:** The definition of the shortest distance defines the outline of the buffer used for the matching. In the following, two possible definitions are described in detail (there are, of course, other options, they are not further pursued here).

Assume network **m** to be the one which will be labeled as *matched* or *unmatched*, depending on the overlap with the buffer around network **b**. For a particular point **P** on **m** the distance to **b** can be defined as the Euclidean distance between **P** and the intersection **I** between **b** and a straight line **l** through **P**, where **l** can be chosen either perpendicular to **m** in **P** (case 1, see Fig. 4a), or perpendicular to **b** in **I** (case 2, see Fig. 4b).

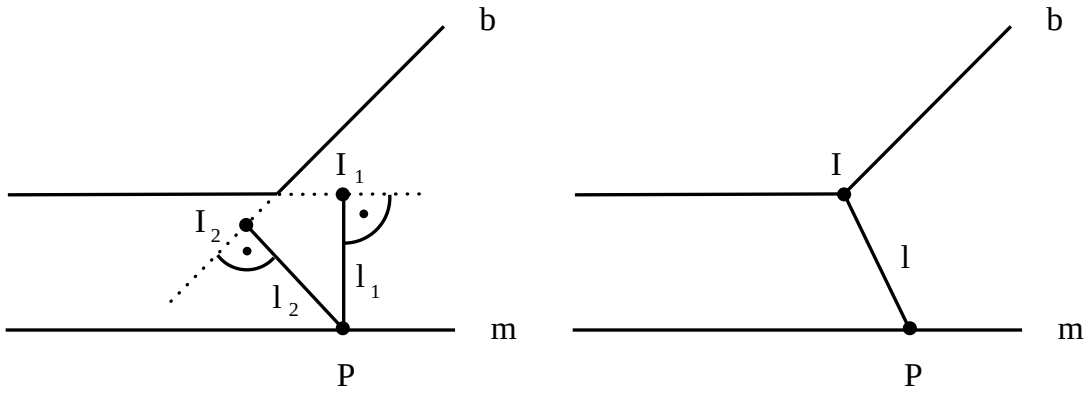
As **P** travels on **m** the shortest distance to **b** is calculated for every piece. In case 1 the calculation is self-evident, whereas it needs some further explanations in case 2: First, the correct **b**-piece has to be chosen. This is done by computing **l** to all **b**-pieces in a predefined vicinity of **P**, and subsequently selecting the piece with the smallest **l** fulfilling the direction constraint. There are pieces of **m** for which **I** lies outside any piece of **b** (see Fig. 5a). This problem is solved by allowing a non-perpendicular intersection of **l** with **b** if **I** is placed on a node of **b** which is connected to more than one other node (see Fig. 5b). In those cases, the direction of **b** is defined as the direction of the




(a) case 1 (b) case 2

**Figure 4. Distance definition**

adjacent piece which yields the smallest direction difference.




(a) **I** outside any piece of **b** (b) **I** on a node of **b**

**Figure 5. Special case of distance calculation in case 2**

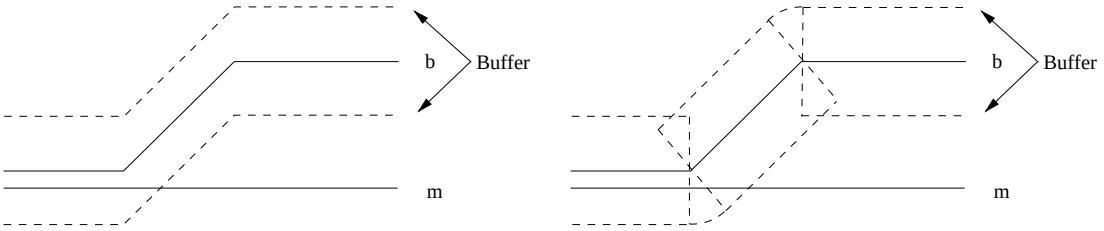
The value of the label (*matched*/*unmatched*) is determined for each piece of **m** based on a check if the distance value is below the buffer width, i.e., the parts of **m** which lie within the buffer around **b** are labeled as matched.

Case 1 of the distance definition yields a buffer width which depends on the direction difference between the two parts to be matched (see Fig. 6a). Problems which can be solved by an analysis of the direction difference (see Section 3.3) are treated implicitly by using a smaller buffer in case of larger direction difference.

In case 2 the resulting buffer width is constant and therefore independent of any direction difference (see Fig. 6b).

The evaluation results described in this paper are based on an implementation of case 2, because of the strict separation between distance and direction.

**Ad 2:** The accuracy of the matching and the quality measures is directly influenced by the split length used for splitting the networks. Obviously, a higher accuracy is achieved if a smaller split length is chosen. The split length 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 are kept to an acceptable level.




(a) case 1 (b) case 2

**Figure 6. Resulting buffer**

## 4: Experiments and results

The proposed scheme has been used to evaluate the results of the three described approaches on three different black and white images. A description of the test images and the evaluation procedure, the evaluation results and a detailed analysis thereof 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 models of the approaches are violated. It should be noted that all three images are rather large, their size amounts to approximately 1/16 of a photogrammetric aerial image.

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

**Table 1. Description of test images**

For the test all three algorithms were run in a totally automatic fashion. As described in Section 2 road extraction was only carried out in the open areas. It should be mentioned that each algorithm requires a number of threshold parameters to be set prior to the computations. E.g., the line algorithm for the TUM approaches is based on low resolution images. They were generated by subsampling the test images to a pixel size of 3.6 m. As mentioned earlier, the aim of the study is to investigate the evaluation scheme. Information concerning the general applicability of each algo-

rithm across different images is of secondary importance only. Therefore, it was thought acceptable to change threshold parameters from one image to the next according to visual inspection.

Also the matching procedure needs some parameters to be set, namely the buffer width, the maximum direction difference, and the split length. These parameters were chosen equal for all evaluations, and were set to 3 m for the buffer width (assuming a maximum road width of 6 m), 20° for the maximum direction difference, and 0.1 m for the split length.

The results are depicted in the Figures 7 through 9. For each test image the image superimposed with the mask for open area, 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

First, the results of each algorithm are discussed separately. Then some comments valid for the whole investigation are given.

TUM-G algorithm: Due to the strong road model and the combination of low and high resolution image information, the TUM-G approach is able to deliver relatively stable road hypotheses which are also used to bridge small gaps. Consequently, the extracted road parts are well connected and thus quite long, the number of gaps is moderate, and the average gap length is not too high. Most important, the correctness is relatively high. The geometrical accuracy lies between 1 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 some problems with the more difficult scenes Erquy and especially Montserrat in which roads have a greatly varying appearance.

TUM-S algorithm: The TUM-S approach is especially aiming at bridging gaps due to shadows and occlusions. As desired TUM-S is bridging some of the gaps in the result of TUM-G. Therefore, the overall number of gaps is smaller than that of TUM-G which indicates that the connectivity of the extracted road parts is better. The geometrical accuracy again lies approximately between 1 and 4 pixels.

For Marchetsreut and Erquy the performance is a little worse than that of TUM-G (except the correctness), whereas better results were achieved on the Montserrat image.

IGN algorithm: The road model of the IGN approach depends mainly on the homogeneity of the road instead of the parallelity of the roadsides. Therefore, a performance different from the two TUM approaches can be expected. This is correctly expressed by the quality measures: the IGN approach performs good for the Marchetsreut image, rather poor for the Erquy image, but best for the Montserrat image. The geometrical accuracy lies between 2 and 4 pixels. The connectivity of the extracted road parts is relatively good.

The high correctness of the extracted data especially for the Montserrat image is partly due to the limited search space used for the road seed detection (see Section 2).

Additional comments: Looking at all results it becomes clear that the proposed quality measures adequately capture the impression obtained when visually inspecting the extracted road data. Thus, they can serve as a basis for the comparison of different automatic road extraction algorithms. Besides, the results are an indication for the applicability of the road extraction algorithms: they can serve as automatic extraction tools for easy scenes. It should be noted, however, that due to the effects mentioned in Section 3.3 and the discretization (see Section 3.4), 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 data into different local categories such as clearly visible road parts, roads in shadow, occluded roads etc. or regional categories like open area, urban area, forest, 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 as results are approaching a point where they become useful for practice. In this paper a methodology for the evaluation of automatic road extraction algorithms based on the comparison to manually plotted reference data is presented. This methodology was tested using the results of three approaches 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. E.g., improvements of the TUM-G approach could be achieved firstly by analyzing the gaps in the extraction results, and finding and subsequently modeling the underlying reasons for these gaps such as shadows, occlusions etc. Secondly, other criteria like the homogeneity of the path employed in the IGN approach could complement the approach.

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. an analysis of topological differences between extracted and reference data, especially in the vicinity of crossings, or 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 gratitude to Albert Baumgartner, Ivan Laptev, and Cecile Huet for contributing the software and the results of the three road extraction algorithms to this study.

# References


[1] S. Airault and O. Jamet. Détection et réstitution automatique du réseau routier sur des images. *Traitement du signal*, 12(2):189–200, 1995.

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

[3] A. Baumgartner, W. Eckstein, H. Mayer, C. Heipke, and H. Ebner. Context-Supported Road Extraction. In *Automatic Extraction of Man-Made Objects from Aerial and Space Images (II)*, pages 299–308, Basel, Switzerland, 1997. Birkhäuser Verlag.

[4] A. Baumgartner, C. Steger, H. Mayer, and W. Eckstein. Multi-Resolution, Semantic Objects, and Context for Road Extraction. In *Workshop on Semantic Modeling for the Acquisition of Topographic Information from Images and Maps*, pages 140–156, Basel, Switzerland, 1997. Birkhäuser Verlag.

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

[6] W.J. Christmas, J. Kittler, and M. Petrou. Structural Matching in Computer Vision Using Probabilistic Relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):749–764, 1995.

[7] CMU. 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), 1997.

[8] W. Förstner. 10 Pros and Cons Against Performance Characterization of Vision Algorithms. In *European Conference on Computer Vision, Workshop "Performance Characteristics of Vision Algorithms"*, pages 13–29, 1996.

[9] P. Fua. *RADIUS: Image Understanding for Intelligence Imagery*, chapter Model-Based Optimization: An Approach to Fast, Accurate, and Consistent Site Modeling from Imagery. Morgan Kaufmann, 1997. O. Firschein and T.M. Strat, Eds. Available as Tech Note 570, Artificial Intelligence Center, SRI International.

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

[11] Y. Hsieh. Design and Evaluation of a Semi-Automated Site Modeling System. Technical Report CMU-CS-95-195, Computer Science Department, Carnegie Mellone University, 1995.

[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *International Journal of Computer Vision*, 1(4):321–331, 1987.

[13] H. Mayer, I. Laptev, A. Baumgartner, and C. Steger. Automatic Road Extraction Based on Multiscale Modeling, Context, and Snakes. In *International Archives of Photogrammetry and Remote Sensing*, volume **32**(3-2W3), pages 47–56, 1997.

[14] H. Mayer and C. Steger. 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*, pages 331–348, Vienna, Austria, 1996. Oldenbourg Verlag.

[15] C. McGlone and J. Shufelt. Projective and Object Space Geometry for Monocular Building Extraction. In *Computer Vision and Pattern Recognition*, pages 54–61, 1994.

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

[17] R. Ruskoné and S. Airault. Toward an Automatic Extraction of the Road Network by Local Interpretation of the Scene. In *Photogrammetric Week '97*, pages 147–157, 1997.

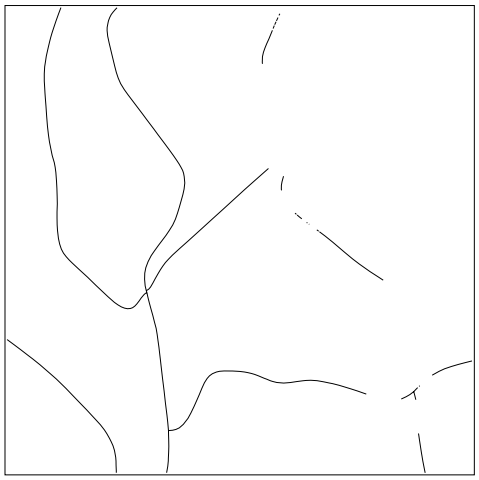
[18] R. Ruskoné, S. Airault, and O. Jamet. Road Network Interpretation: A Topological Hypothesis Driven System. In *International Archives of Photogrammetry and Remote Sensing*, volume **30** (3/2), pages 711–717, 1994.

[19] C. Steger. Removing the Bias from Line Detection. In *Computer Vision and Pattern Recognition*, pages 116–122, 1997.

[20] G. Vosselman and N. Haala. Erkennung topographischer Paßpunkte durch relationale Zuordnung. *Zeitschrift für Photogrammetrie und Fernerkundung*, 6/92:170–176, 1992.

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

a) Grey value image superimposed with mask for open area



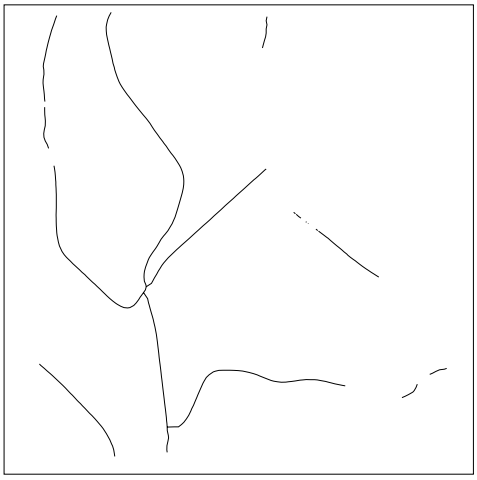
b) Manually plotted reference

| | TUM-G | TUM-S | IGN |
|---|---|---|---|
| Completeness | 0.90 | 0.85 | 0.80 |
| Correctness | 0.99 | 0.98 | 0.94 |
| Quality | 0.89 | 0.84 | 0.76 |
| Redundancy | 0.00 | -0.01 | 0.00 |
| RMS [m] | 0.27 | 0.38 | 0.52 |
| No. of gaps per km | 7.06 | 5.97 | 7.06 |
| $\mu_{gap\,length}$ [m] | 13.58 | 24.80 | 27.84 |

c) Quality measures

d) Results of TUM-G algorithm

e) Results of TUM-S algorithm

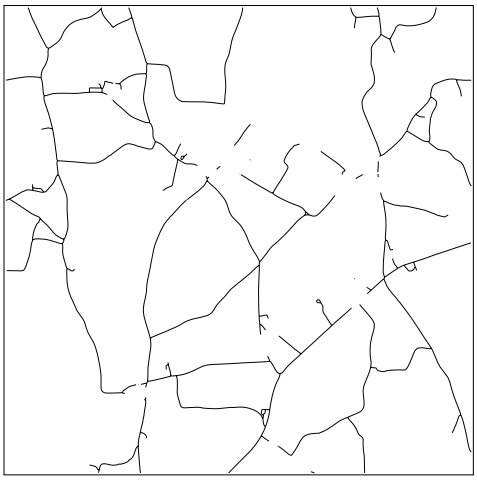


f) Results of IGN algorithm

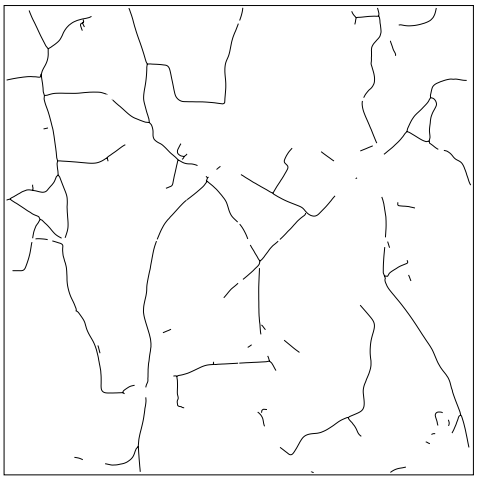
**Figure 7. Test image Marchetsreut**

a) Grey value image superimposed with mask for open area

d) Results of TUM-G extraction



b) Manually plotted reference



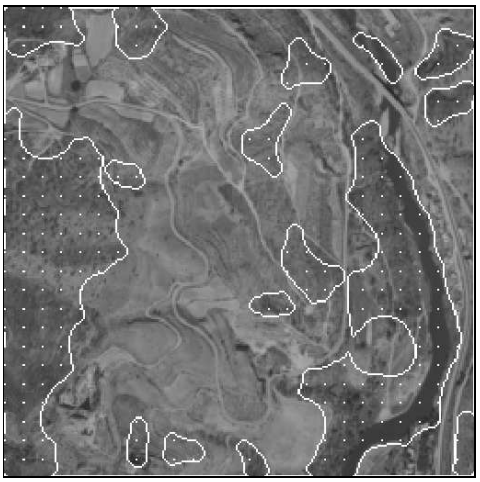
e) Results of TUM-S algorithm

| | TUM-G | TUM-S | IGN |
|---|---|---|---|
| Completeness | 0.77 | 0.72 | 0.43 |
| Correctness | 0.93 | 0.95 | 0.60 |
| Quality | 0.73 | 0.69 | 0.34 |
| Redundancy | 0.02 | -0.01 | 0.00 |
| RMS [m] | 0.53 | 0.47 | 0.92 |
| No. of gaps per km | 5.11 | 4.85 | 4.94 |
| $\mu_{gap\,length}$ [m] | 44.35 | 58.00 | 115.05 |

c) Quality measures

f) Results of IGN algorithm

**Figure 8. Test image Erquy**



a) Grey value image superimposed with mask for open area

d) Results of TUM-G extraction

b) Manually plotted reference

e) Results of TUM-S algorithm

| | TUM-G | TUM-S | IGN |
|---|---|---|---|
| Completeness | 0.31 | 0.45 | 0.46 |
| Correctness | 0.61 | 0.66 | 0.90 |
| Quality | 0.25 | 0.36 | 0.44 |
| Redundancy | -0.01 | -0.03 | 0.00 |
| RMS [m] | 0.91 | 0.66 | 1.05 |
| No. of gaps per km | 5.05 | 4.56 | 4.92 |
| $\mu_{gap\,length}$ [m] | 137.54 | 120.56 | 108.92 |

c) Quality measures

f) Results of IGN algorithm

**Figure 9. Test image Montserrat**