User-Centric Evaluation of Semi-Automated Road Network Extraction

Wilson Harvey, J. Chris McGlone, David M. McKeown and John M. Irvine

Abstract

This paper describes an in-depth usability evaluation of our RoadMAP™ road network extraction system. Six operators used RoadMAP in both manual and semi-automated modes, as well as BAE Systems' SOCET SET®, to extract roads in four image test areas.

An in-depth statistical analysis was performed on the timing results, with the main conclusion being that the performance differences among the three systems were not statistically significant. However, the evaluation highlighted a number of factors other than intrinsic feature extraction competence that affected the results and which point to the potential for significant engineering improvements.

Introduction

Automatic cartographic feature extraction (AFE) has been the subject of significant research activity in both academe and industry. AFE is the application of imagery analysis and interpretation software to detect, delineate, and attribute man-made and natural features such as buildings, roads, forests, rivers and streams. AFE systems can be classified in terms of the degree of user interaction along a continuum ranging from manual to semi-automatic to automatic.

The operational goal for AFE systems has always been human level performance, measured in terms of time and accuracy. However, system evaluation to quantify performance levels and highlight areas for improvement has been problematic. Evaluation methodology has ranged from the display of extraction results on a couple of favorite images to rigorous mathematical analysis of AFE output compared to manually-generated reference datasets. While automated comparison to reference data sets provides a significant improvement over subjective visual evaluation, such evaluations do not address the key issue of usability.

This evaluation was specifically designed to approach performance evaluation in terms of usability. An in-depth statistical analysis was performed on the AFE timing results, with the main conclusion being that the performance of all three systems were comparable. However the evaluation highlighted a number of factors, other than intrinsic AFE competence, that affected the results and point to the potential for significant engineering improvements.

The RoadMAP system and its underlying technology, designed for automated and semi-automated road network extraction, has been the subject of research and development over a period of years. As part of the systems development, we have conducted ongoing evaluations in-house and also in conjunction with our research sponsors. This article describes the most recent and most extensive evaluation of RoadMAP, conducted in cooperation with NGA at the Digital Mapping Laboratory in February 2002.

The main goal of this evaluation was the rigorous comparison of RoadMAP's semi-automated performance against manual road network extraction systems. In this case, we compared the semi-automated RoadMAP system against RoadMAP run in a fully-manual mode, without any assistance from automated processes, and against BAE Systems' SOCET SET, which is used extensively within DoD agencies and commercial mapping organizations.

Six operators used RoadMAP in manual and semi-automated modes along with SOCET SET to extract roads in four image test areas. A number of factors besides pure algorithmic performance can affect overall system performance so the evaluation was designed to isolate and quantify these factors as much as possible.

Previous Evaluation of Road Network Extraction Systems

The Digital Mapping Laboratory has always considered rigorous performance evaluation an integral part of the system development process [McKeown *et al.*, 2000]. Our research in cartographic feature extraction has therefore been paralleled by research into evaluation methods and an ongoing investment in the generation and maintenance of the varied data sets required. We have evaluated both automated and semi-automated systems for the extraction of roads, buildings, elevation, and surface materials.

Until this point we have focused on evaluating RoadMAP in its automated operating mode, as part of its development and refinement [Harvey, 1999; Harvey, 1997]. Extensive testing has studied the effects of different types of trackers and control strategies. However, the emphasis in the evaluation of an automated system is somewhat different from that for a semi-automated system. For automated systems, we are mainly concerned with the accuracy and completeness of the results and the time required for their generation. We have developed a comprehensive set of metrics for this, described in [Harvey, 1999; Harvey, 1997]. For semi-automated systems, we assume that the final results will be correct since an operator has controlled the extraction, and only the time required relative to a manual system is of interest. The design of the experiment, however, must consider human factors such as image familiarity and learning effects.

Previous and the Evolution of RoadMAP

Automated extraction of road networks has been a long-term topic of research. Most early work focused on linear feature extraction approaches [Fischler *et al.*, 1978; Quam, 1978; Fischler *et al.*, 1981], with restrictive assumptions on image resolution and road appearance. Research continues at a number of institutions on various approaches

Wilson Harvey, J. Chris McGlone, and David M. McKeown are with the Digital Mapping Laboratory, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 (wah@cs.cmu.edu, jcm@cs.cmu.edu, dmm@cs.cmu.edu).

John M. Irvine is with SAIC, Inc., 20 Burlington Mall Road, Suite 30, Burlington, MA 01803 (jirvine@bos.saic.com).

Photogrammetric Engineering & Remote Sensing
Vol. 70, No. 12, December 2004, pp. 1353–1364.

0099-1112/04/7012–1353$3.00/0

to the problem, including multiscale approaches [Airault *et al.*, 1994], detailed semantic models [Wiedemann and Hinz, 1999], knowledge-based extraction [Trinder and Wang, 1998] and context cues [Hinz and Baumgartner, 2000]. Research at the Digital Mapping Laboratory began with the use of cooperative tracking methods [McKeown and Denlinger, 1986; McKeown and Denlinger, 1988], and has proceeded through several stages of development.

RoadMAP is a semi-automated road feature extraction system developed by the Digital Mapping Laboratory. In RoadMAP, we explore the utility of integrating automated extraction tools in a graphical user interface to aid an operator in extracting road features from aerial imagery. The extraction tools included are the products of past research in the MAPSLab, operating within a "cooperative methods" framework, where features are extracted by the use of complementary methods instead of reliance on a single approach.

The RoadMAP interface allows the user to direct automated road detection and delineation operations. A manual feature editor is integrated in the interface so that errors in the automatically extracted features can be edited, and topology and semantic information can be added.

Most of the work on RoadMAP in the past two years has been on improving its semi-automated operation, in support of its evaluation at various U.S government facilities. This has mostly involved system engineering, such as improving the user interface, hardening the processing code to improve its robustness, and the production of extensive user documentation. While such engineering work is not research, *per se*, it has a major impact on the performance of semi-automated systems, as shown in this evaluation.

RoadMAP has been part of an ongoing multi-stage AFE Test and Evaluation program conducted by the National Geospatial-Intelligence Agency (NGA). While this process motivated developers to improve system robustness, usability, and documentation, the logistics of installing and supporting systems within NGA proved to be complicated. This was particularly evident in the time and effort required to install and support RoadMAP away from the university and to import and process imagery we could not see. Initial evaluations focused primarily on the graphical user interface (GUI) and adding interactive behaviors to what was a highly automated road extraction system. This work produced the semi-automated system that was evaluated along with the fully automated version of RoadMAP. Given the breadth of performance tests described in this paper, NGA decided that further testing could be accomplished in a shorter period of time if it was conducted at the developers' sites. RoadMAP was the first system to be evaluated in this manner.

Evaluation Methodology

The evaluation was conducted at CMU with a mix of NGA and CMU participants. The basic approach is a controlled comparison of road extraction using RoadMAP in the assisted mode to manual extraction with RoadMAP and manual extraction using SOCET SET. The primary measure of performance was the total time required to extract the roads from each test scene. The philosophy underlying the experiment is that road features must be extracted to a certain level of accuracy to satisfy the product specifications for commercial production. If an initial extraction fails to satisfy this criterion, additional editing must be performed to bring the extraction "up to spec." Experienced production analysts reviewed all extractions to insure that these standards were met. Consequently, measuring extraction accuracy, while useful for understanding algorithm performance, is less relevant to this user-oriented evaluation.

TABLE 1. THE MAKE-UP OF THE ANALYST TEAMS.

Team	Analyst	Organization	Years of Experience with Geospatial Production	RoadMAP Experience	SOCET SET Experience
1	1	NGA	13	Yes	Extensive
1	2	NGA	14	No	No
2	1	NGA	8	No	On-site training
2	2	CMU	NA	Yes	No
3	1	CMU	NA	Developer	No
3	2	NGA	10	Yes	Yes

TABLE 2. TEST IMAGE CHARACTERISTICS.

Name	Type	GSD	Test
Marchetsreut	Aerial	0.25 m	training1
Rancho Bernardo	Aerial	0.60 m	training2, rb2, rb4
Camp LeJeune	IKONOS	1.00 m	cl1
Ft Hood North	IKONOS	1.00 m	nfh2

Analyst Teams

The analysts conducting the evaluation came with varying types and levels of experience, as summarized in Table 1. Each analyst was rotated through each test system and test dataset on a pre-determined schedule.

Analyst Training

Given the analysts' varying levels of experience with RoadMAP and SOCET SET, it was important to ensure that the analysts were consistently trained on each system. Two training images were selected (see Table 2) and a standard set of operating instructions were written for each system and given to each analyst. For each training image and extraction system, the analyst extracted the roads twice consecutively. Each trial was timed and the times compared to ensure that the operator was on the flat part of the learning curve.

Test Images

In addition to the training images described above, three different images were selected for use in the extraction tests, with two or three separate test areas on each. The images were selected to span a variety of scene content, including road types and background clutter (Table 2).

The Rancho Bernardo image (Figure 1) is a standard aerial mapping photograph of a suburban/commercial area in southern California, scanned at 30 micrometers to give a 0.6m ground sample distance (GSD). One test area (`rb4`) consists of the larger roads in the scene, including an interstate highway, while the other (`rb2`) contains mostly local streets. The Camp LeJeune (`cl1`, Figure 2a) and North Ft. Hood (`nfh2`, Figure 2b) images are both from the SpaceImaging IKONOS satellite, with a nominal one meter GSD, and were supplied by NGA. Both show rural areas, with the Camp LeJeune test area having heavy tree cover. The contrast on the Camp LeJeune image was poor, making precise identification of road boundaries significantly more difficult than on the other test images.

The reference data shown superimposed on the test images was manually collected using RoadMAP in the weeks prior to the evaluation. Table 3 gives some summary road statistics for each test area providing an indication of overall road properties.

(a)

(b)

Figure 1. Test areas from the Rancho Bernardo image: (a) Test area rb2, and (b) Test area rb4.

RoadMAP

The RoadMAP user interface (Figure 3) enables the operator to control the manual, semi-automated, and automated extraction of road networks and to edit the extracted networks. In all modes, tracking begins with a road marker which defines a point in the center of the road, the road direction, and the road width. In manual and semi-automated modes, the marker is placed and adjusted by the operator, as shown in Figure 4. In manual tracking, the operator starts from the marker and clicks along the centerline of the road. When the extracted road crosses a previously-extracted road, the operator places an intersection. In semi-automated mode, the system begins tracking the road from the

(a)

(b)

Figure 2. Test areas from IKONOS images, Copyright© 2000, SpaceImaging, Inc.: (a) Test area cl1, and (b) Test area nfh2.

TABLE 3. TEST DATASET ROAD STATISTICS.

Dataset	Figure	Number of Roads	Average Length, km	Standard Deviation, km	Minimum Length, km	Maximum Length, km	Total Length, km
rb2	1a	55	0.29	0.28	0.04	1.28	15.76
rb4	1b	21	2.89	2.64	0.66	10.49	60.74
cl1	2a	31	0.80	0.93	0.06	4.51	24.77
nfh2	2b	30	0.96	1.15	0.04	4.19	28.86

Figure 3. RoadMAP user interface.

marker and continues until it is unable to track any farther or it reaches the edge of the image. When it tracks across a previously-extracted road, the system automatically forms an intersection. If the system loses the track, the operator can edit any erroneous portions of the track and either have the system resume tracking or just complete the road manually. A detailed description of the operation of RoadMAP is found in [Harvey, 1998].

SOCET SET as the Baseline Manual System

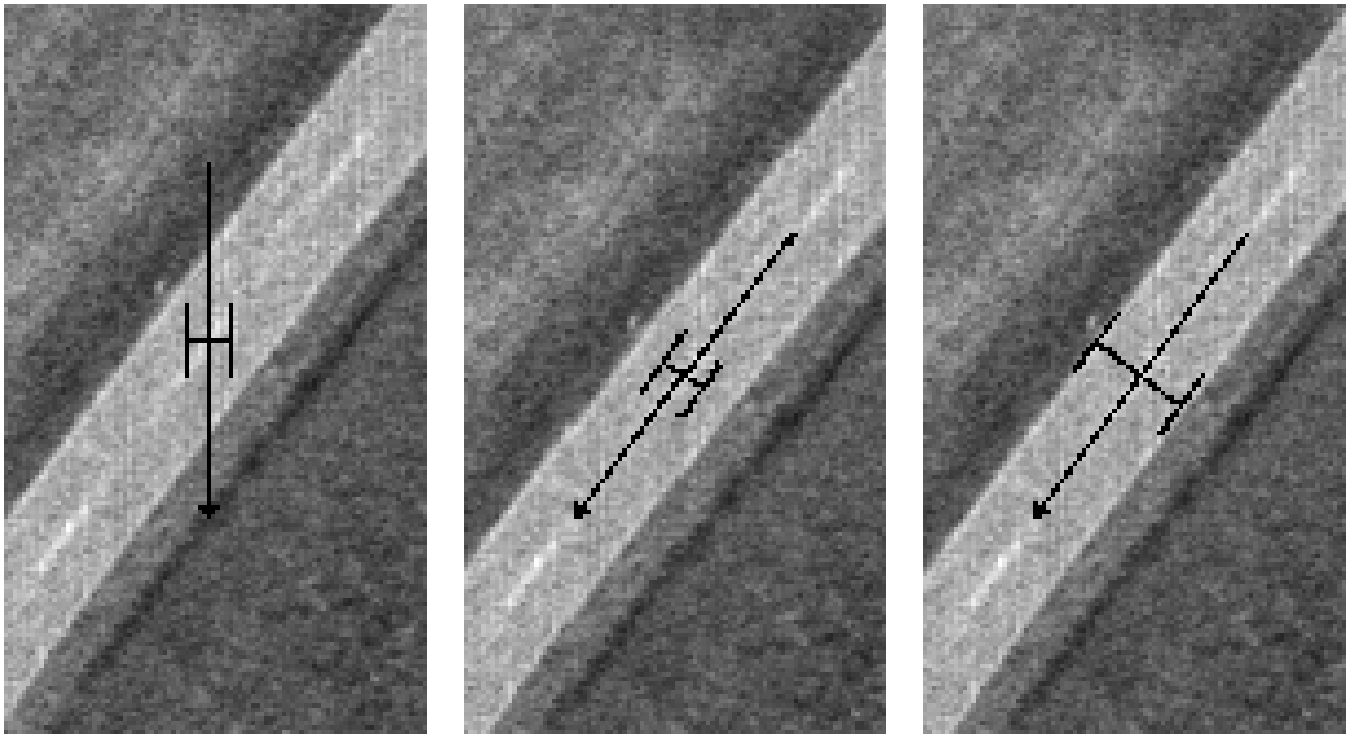
An important issue in the evaluation of semi-automated systems is the baseline manual system to use for the comparisons. To quantify the benefits of the automated processes, the comparison should be made against the same system run without the automated processes, while to quantify performance differences against an existing production process, the manual system should be the current production system. Both questions were of interest in this evaluation, so manual extractions were performed using both RoadMAP without the automated processes and SOCET SET, as representative of the commercial production environment.

SOCET SET tests were performed using Version 4.3.1 running on Windows NT ©. Several configuration changes were made to the SOCET SET key bindings by experienced NGA personnel before the tests, to make the operation of SOCET SET more efficient and closer to their normal working environment.

Data Collection

The data collection phase of the evaluation took place at Carnegie Mellon University during the week of February 4–8, 2002. The analysts on each team worked in pairs, with one extracting data while the other timed the extraction using a hand-held stop watch. Times were manually recorded on prepared data sheets.

The original timing plan was to divide the process into three phases—extraction, editing, and topology collection—and time each phase separately. This was motivated by the assumption that the col-



(a) (b) (c)

Figure 4. RoadMAP marker adjustment: (a) Initial marker placement, (b) Marker aligned with road, and (c) Marker width adjusted.

TABLE 4. COMPUTERS USED FOR EACH SYSTEM.

Application	Hardware	OS	Processor	Memory
SOCET SET	Dell PC	Windows 2000	Two 866MHz Intel P3	512 MB 512 MB
RoadMAP Manual	SGI O2	IRIX 6.5	180 MHZ R5000	256 MB 256 MB
RoadMAP Assisted	SGI Octane	IRIX 6.5	Two 195 MHz MIPS R10000	196 MB 196 MB

lected data would require editing and that topology collection would be a separate step with SOCET SET. As it turned out, neither assumption was true so only one time was actually collected per test area.

Ideally, SOCET SET and RoadMAP would have been run on comparable computers, with the same hardware and operating systems. However RoadMAP currently runs only on Silicon Graphics® (SGI) Workstations while our copy of SOCET SET ran under Windows. Consequently, the evaluations were performed on different computers as described in Table 4.

Operator Discussion and Questionnaires

Not all of the useful information from an evaluation is numerical. Especially for semi-automated systems, operator comments on system issues, data set characteristics, and the evaluation procedures can provide useful insights that are not necessarily reflected in the timings. To capture this information, we scheduled daily wrap-up discussions where system problems or questions, issues with the day's data sets, and the next day's plans were discussed. This made sure that relevant questions were raised and addressed and kept all the analysts informed.

At the end of the week each analyst was given a questionnaire which asked for the operator's opinions on the adequacy of RoadMAP and SOCET SET training and the effectiveness of RoadMAP, along with questions on past experience. The ease and intuitive nature of the RoadMAP interface received high marks. Even so, analyst's opinions were less positive about whether or not having RoadMAP available in their work environment would make their jobs easier (4 of 6 respondents, 2 "easier" and 2 "no effect").

Regarding the adequacy of training, before the timed extractions all the analysts felt that they had adequate training to perform the extractions. However, in the wrap-up discussions, most felt that they would have liked to have had more training using the automated road extraction tool in RoadMAP. This is consistent with the results of the learning curve tests where, upon review, most of the analysts showed improvements of 20% or more in the sequential extractions. In retrospect, we should have paid more attention to these measurements and taken the time to allow for more extensive experimentation and training in RoadMAP.

Evaluation Results and Analysis

The main purpose of this evaluation was to determine if road network extraction using the semi-automated RoadMAP system is faster than current manual extraction methods. The collected timings have been analyzed to answer this question, as discussed in Sections *Timing Analysis* and *Statistical Analysis*. In the discussions, the test areas will be abbreviated as shown in Table 2, while the team and analyst for each test will be abbreviated as `t1a1`, for team 1, analyst 1, etc.

In any experiment, answering one question gives rise to new questions of why and how. We performed additional experiments and anal-

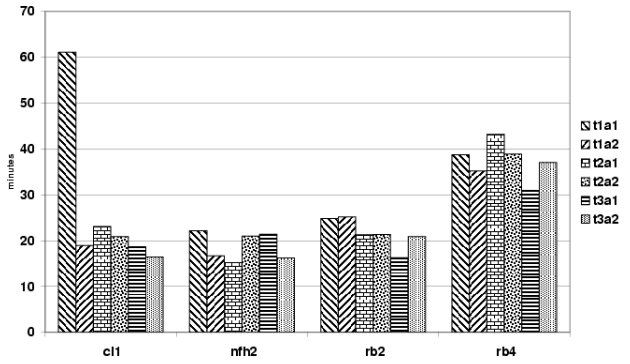


Figure 5. SOCET SET times by operator and test area.

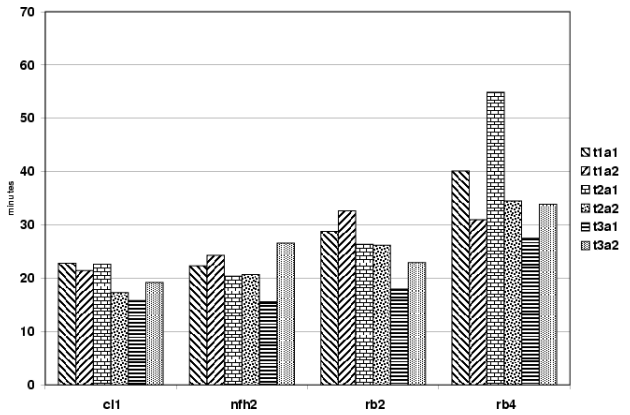


Figure 6. RoadMAP Manual times by operator and test area.

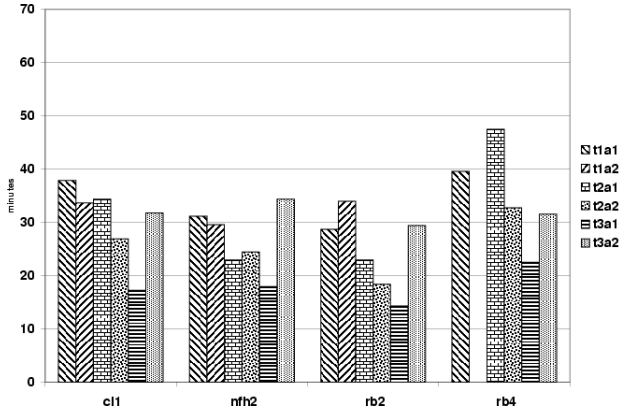


Figure 7. RoadMAP Assisted times by operator and test area.

yses to better understand the causes of the performance differences and to determine how best to improve the performance of RoadMAP. These analyses are discussed in Section *Factors Affecting Extraction Speed*.

Timing Analysis

The first-order timing numbers suggest that an operator using SOCET SET is about as fast as an operator using RoadMAP Manual

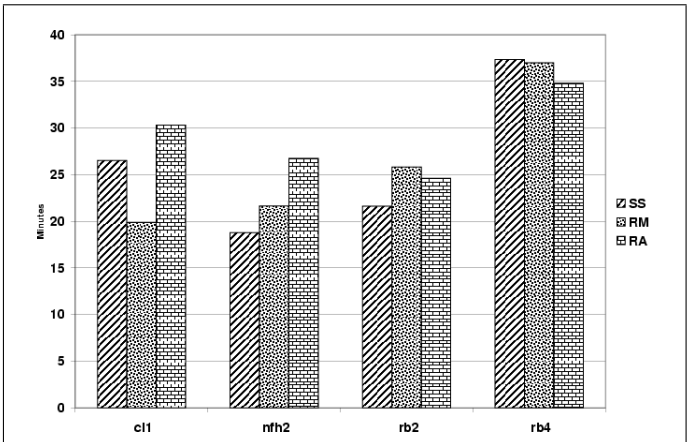


Figure 8. Average system times for each test area.

(average times: 26.07 minutes per data set using SOCET SET versus 26.08 minutes per data set using RoadMAP manually). An operator using RoadMAP's automated road extraction tool is about 12% slower than when using SOCET SET (average time: 29.12 minutes per data set).

Figures 5, 6 and 7 are plots of the raw times by operator and test area for each of the three systems tested. Due to some technical issues, operator `t1a2` was unable to complete the extraction on test area `rb4` using RoadMAP Assisted. The data in Figures 5–7 are summarized in Figure 8 which provides the averages of the operator times per system per test area.

The SOCET SET extraction of the `cl1` test area by operator `t1a1` appears to be an anomaly: the extraction time is almost triple the time recorded for any of the other operators, and operator `t1a1` was our SOCET SET expert. The expert RoadMAP user, operator `t3a1`, can be observed to have the fastest extraction times for all operators when using either RoadMAP Manual or RoadMAP Assisted.

More informative, perhaps, are Figures 9a, 9b, and 9c, which present group statistics for the extraction times per test area for each system. These plots show the average extraction time with upper and lower standard deviations, as well as minimum and maximum timings.

Inspecting Figures 9a–9c shows that (except for the SOCET SET extraction over the `cl1` test area) the standard deviation of the RoadMAP Assisted extraction times is consistently larger than for either of the other two systems. This suggests a wider range of expertise among the users. If we throw out the anomalous data point for the SOCET SET extraction over `cl1`, in contrast, the standard deviation for SOCET SET is consistently lower than that for either of the other systems. In explanation, one of the operators commented that there are more extraction methods available in RoadMAP than in SOCET SET. This is not strictly true, but it does suggest that the SOCET SET extraction interface provides users fewer options, thus making them feel more "focused" to perform the necessary extraction operations faster.

Comparing RoadMAP to a mature commercial GIS product like SOCET SET has highlighted the effects that engineering can have on efficiency. Analysis and discussions after the evaluation lead us to believe that several engineering inefficiencies could be responsible for degrading RoadMAP's performance.

Statistical Analysis

The design of the experiment supports a controlled comparison of the three systems: RoadMAP Manual, RoadMAP Assisted, and SOCET SET. An Analysis of Variance (ANOVA) was performed to compare the timing data across systems, while controlling for effects

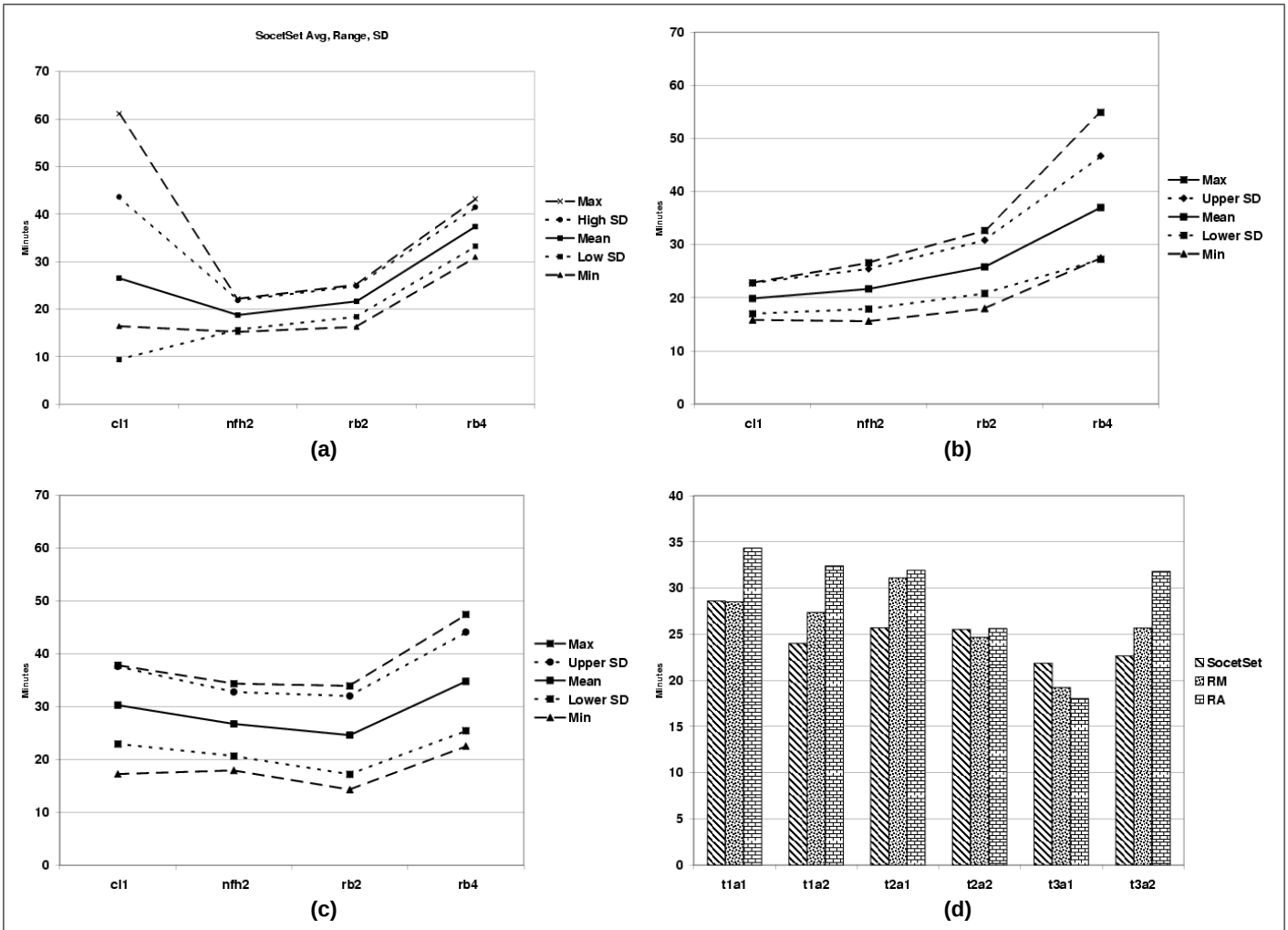


Figure 9. Time statistics by test area: (a) SOCET SET statistics by test area, (b) RoadMAP Manual statistics by test area, (c) RoadMAP Assisted statistics by test area, and (d) Average operator times across test areas for each system.

due to analyst and scene. Performance differences, as measured by the timing data, are to be expected across analysts, since individuals differ in their levels of skill and experience. Similarly, we expect differences among the scenes; some images are more challenging than others. The ANOVA fits a linear model to the data with terms for Scene and Analyst, as well as the terms for the three Systems. The F-tests and associated significance levels for each term indicate whether differences are real or could be attributed to chance. The analysis shows that real differences in performance are attributable to the Scene and Analyst, but not System (Table 5). The significance level of 0.160 for system indicates that the differences in mean extraction time across the three systems are consistent with the chance variability, when in fact no difference exists. The highly significant effects due to Scene and Analyst show that there are real differences here and controlling for these factors (as we have done) is appropriate.

TABLE 5. ANALYSIS OF VARIANCE FOR TIMING DATA.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic	Significance Level
System	165.7	2	82.86	1.887	0.160
Scene	2161.6	3	720.54	16.408	0.0001
Analyst	1241.6	5	248.32	5.655	0.0001
Error	2634.8	60	43.91		

Factors Affecting Extraction Speed

Many of the observations made while conducting this experiment have helped us to formulate hypotheses which may explain the differences in extraction time between RoadMAP and manual extraction. We have data to provide evidence for some but not all of these hypotheses. We intend to conduct future experiments to confirm or refute those hypotheses where we currently lack supporting data.

Some of these hypotheses can be explored with data already collected or by instrumenting RoadMAP to collect some representative data after the fact. Others, namely the cognitive issues, will be more difficult to quantify and, thus, substantiate or refute. In the remainder of this section, we discuss each of these issues and include reasonable estimates for achievable performance improvements.

Width Extraction Overhead

As discussed in Section *Data Collection*, the timed extractions in this evaluation attempted to measure width extraction as a separate process. Though somewhat artificial, this is fine when using SOCET SET and

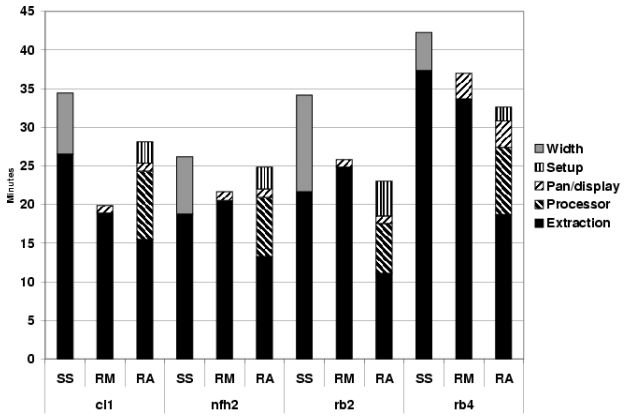


Figure 10. Adjusted collection times by test area and system.

RoadMAP Manual because width extraction can be performed as a separate step. For RoadMAP Assisted, however, measuring the road width is part of setting the initial marker for the automated tracker and is not a separate process that can be easily timed. We therefore have no data to attempt an estimate of the amount of time spent in supplying the width information when using RoadMAP Assisted.

We can, however, estimate the amount of time it would take to collect comparable width information using SOCET SET. On the final day of the evaluation, two analysts (one of them a SOCET SET expert) were timed extracting width data using SOCET SET. About 20 width measurements were collected, and we found that it took about 15 seconds per width measurement. Applying this time to the number of roads in each test area, we found that collecting width data in SOCET SET would have added an average of 8 minutes to each collection, or almost 50% to the total extraction time. This is shown in Figure 10.

Automated Process Setup Time

While the automated road tracker is typically faster than manual tracking, RoadMAP currently requires the operator to select an initial point on the road and specify the road's direction and width (Figure 4). While relatively short, this setup time can be significant in overall data collection time for short roads.

For instance, assume that the automated system tracks at speed $s_a \frac{pixels}{second}$ while the operator manually tracks at $s_m \frac{pixels}{second}$. If the setup time is t_s and the length of the road is l, then for the automated system to be faster than manual tracking,

$$\frac{t_s s_m}{l} < 1 - \frac{s_m}{s_a}$$

This shows that as the length of the "headstart," $t_s s_m$ increases relative to the length of the road, the automated system must be correspondingly faster. Alternatively, as the setup time or the manual tracking speed increase, the tracked roads must be longer for the automated system to be faster.

RoadMAP is not currently set up to separately collect the setup time or to document the time spent manually extracting roads vs. automated extraction time. RoadMAP does, however, maintain counts of the number of roads extracted manually and automatically. This is an underestimate of the number of times the user must set up the automated tracker, since a single road may require multiple starts of the automated tracker to delineate the entire road, but does gives us a reasonable basis to estimate the number of setups.

We assumed that at least five seconds can be saved each time the tracker is started by using an automated process to set the initial parameters. Multiplying this time by the number of automated tracker starts gives an estimate for the total amount of time saved on each data set. This was then averaged across operators to arrive at the final estimate per test area, shown in Figure 10.

These estimates show that setup time is a significant performance issue. Thus, we are studying methods, based on our work in automated road start point generation, to automatically set the direction and width of the road once the starting point is specified.

Differences in Display Refresh Speed

The RoadMAP user interface is built on generic calls to the X11/Motif libraries, so the implementation does not take advantage of hardware acceleration for image panning and zooming. This deficiency is easily observed while running the automated tracker on a long road. The image display takes a noticeable amount of time to redraw the image data, and this refresh time slows the tracker significantly. In comparison, image panning in SOCET SET on the PC with hardware display support is smooth, with no detectable update time. We intend to fix this inefficiency when we re-implement RoadMAP on the Windows platform.

To estimate the amount of time wasted redrawing the image data, we need estimates of the time per panning operation and the number of panning operations required per test area.

To obtain an estimate of the time per panning operation, we measured the time to pan the entire image while counting the number of pan motions. We performed this experiment 12 times on different images and panning in different directions and obtained a time of 1.15 seconds per pan operation.

To estimate the number of pan operations per test area, we computed the diagonal size of the minimum bounding rectangle of each road extracted from the test area and divided that by the size of the RoadMAP extraction window. This provided an estimate of the number of pan operations for each road, and summing these yielded the number of pan operations for that test area. We averaged this value for each operator to arrive at the final number used in our estimates. With this method, we estimate 55 pan operations for test area `cl1`, 58 pan operations for `nfh2`, 51 pan operations for `rb2`, and 179 pan operations for `rb4`. This calculation underestimates the total number of pan operations per test area because it does not include the pan operations necessary to move from one road to the next.

Using these numbers, we find that approximately one minute is spent refreshing the image data for all the test areas except for `rb4`, where we are spending almost 3.5 minutes. This estimate accounts for as much as 10% of the total time for these data sets, demonstrating that a single engineering issue can significantly impact system performance.

Differences in Processor Speed

There were significant effects on the collection times due to the differences in machine speed, since the SOCET SET tests were run on a dual-processor 866MHz PIII, while the RoadMAP Assisted tests were run on a dual-processor 195MHz R10000 SGI Octane. We assumed that these effects would only adversely impact the automated extraction time and have little or no impact on manual extraction time.

To compute an estimate for the processor factor, a command-line version of the road tracker was run on both the SGI and NT machines on three data sets. The time ratios were computed for the individual datasets, then averaged. Our estimate is that the SGI is about 80% slower than the NT machine when running RoadMAP's automated

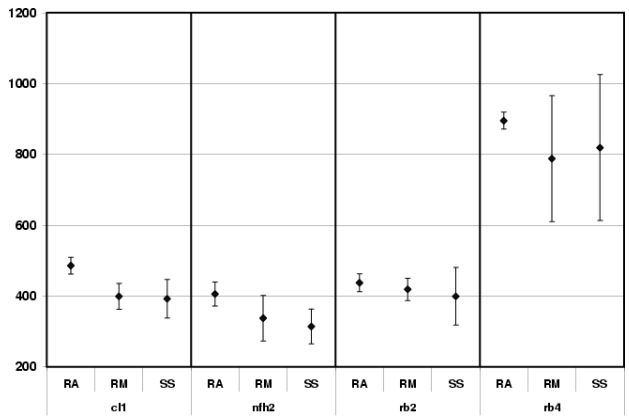


Figure 11. Average number of extracted points and standard deviation for each test area and system.

extraction tools. This estimated processor speed adjustment, applied only to the time spent in automated extraction, is shown in Figure 10.

Operator Experience

The evaluation included two "power users," one very experienced with SOCET SET collection (`t1a1`) while the other (`t3a1`) is the developer of RoadMAP. Not surprisingly, they had better times using the system on which they had the most expertise. The user with previous acquaintance with RoadMAP (`t2a2`) had roughly equal average times on SOCET SET and RoadMAP, while the other three users, with some previous SOCET SET experience but no RoadMAP experience, had better average times on SOCET SET.

Consistency of Collected Vectors

It appears that the data collected by RoadMAP Assisted is more consistent than that collected by users of either SOCET SET or RoadMAP Manual, in terms of positioning and the number of points collected. This has important implications for a cartographic production system, since a large amount of effort is expended in training operators to collect data in accordance with extraction specifications and in editing the collected data to ensure that it meets the specifications. It may very well be that one of the largest benefits of automated systems is the collection of data sets with well-defined characteristics, reducing operator training and monitoring requirements and also post-editing effort. This topic was not explicitly studied in this evaluation, but should be considered in future work.

Figure 11 shows the average number of points extracted using each system for each test area, with the error bars showing the standard deviation of the number of data points. Except for the `rb4` dataset, all the datasets show the same trend: analysts using RoadMAP Assisted collect the most points, and analysts using SOCET SET collect the fewest points. There was a single data point in `rb4` where the analyst using SOCET SET collected almost twice the average number of points. If this data point is thrown out, the `rb4` dataset shows a similar trend. The standard deviation and the range of the number of points collected is smaller for users of RoadMAP Assisted than for the other two systems, implying that the operators collected a more consistent number of points for each test area.

This appears to be mostly due to two factors. First, SOCET SET was typically run at a lower image resolution, either a two or four times reduction, while RoadMAP was run at full resolution. Second, RoadMAP is designed to collect points at a nearly-uniform spacing, without taking into account straight sections of roads where a human operator would capture a point only at the beginning and end of the section.

Figure 12 shows a portion of the vectors extracted by all six analysts over test area `rb2`. The vectors extracted using SOCET SET are shown in Figure 12a and those extracted using RoadMAP Assisted are shown in Figure 12b. A visual inspection of the vector data suggests that operators using RoadMAP Assisted appear to generate more consistent and accurate centerline positions than when using SOCET SET, especially around curves.

Excessive Use of Automated Processes

One behavior observed during the evaluation was that operators would often try to use the automated tracking tool on areas where it would have a poor chance of success, such as roads with poor contrast against the background, or roads which were occluded or shadowed. Many times, the operators appeared to be testing the automated tool in order to get a better "feel" for where it would and would not work. This implies that these operators required more training with the automated tools, and that a future graphical interface should provide feedback to assist operators in determining when the tool is likely to succeed or fail. Inspecting the data in Figures 9a–9c shows that RoadMAP Assisted has a consistently higher standard deviation than the other two systems, which would be consistent with this hypothesis.

Placement of Road Centerline Points

It is possible that the presence of the width polygon in RoadMAP unconsciously slowed the placement of centerline points by making users feel they had to place the road center point more accurately. No quantitative data was collected to support this hypothesis, although this comment was made by several of the analysts in the post-evaluation discussion. This subjective evidence suggests that we should consider designing future experiments to study this hypothesis.

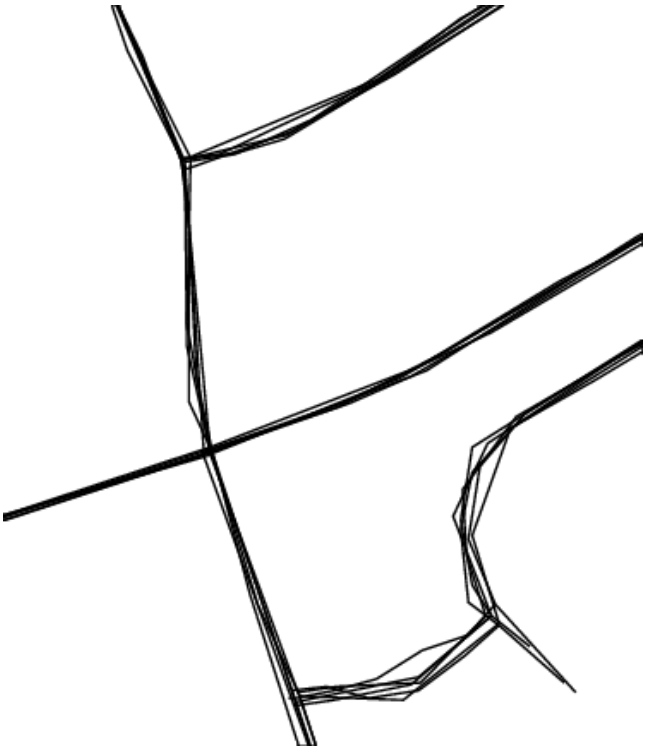
Reflections and Lessons Learned

There were several lessons learned as a result of this rigorous user-centric evaluation of RoadMAP, which can be loosely organized into the areas of overall system performance, evaluation procedure and operator effects. Running this large-scale evaluation highlighted some procedural and data collection changes we will implement for the next evaluation. It also showed the importance of user-centric evaluation for complex AFE systems with a large computer vision algorithm component, in addition to the standard tests in isolation against imagery and extracted feature reference datasets. In this section we discuss some of the lessons learned and how we will apply them to future RoadMAP development and evaluation.

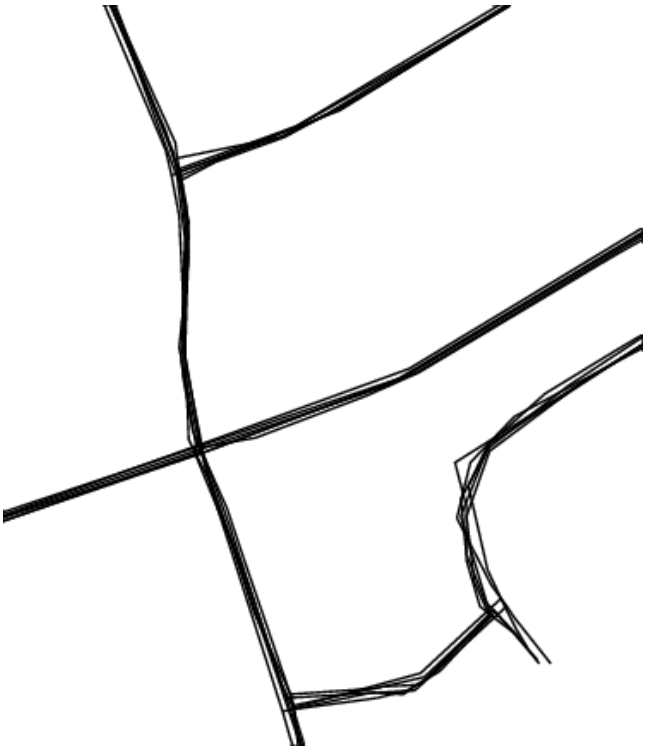
Relative System Performance

While the raw timings showed that RoadMAP Assisted was competitive with SOCET SET, we were surprised by the small actual differences, particularly since SOCET SET provides a largely manual compilation environment. This was partially due to the fact that our internal evaluations have usually compared RoadMAP Assisted against RoadMAP Manual and not against SOCET SET, which was considered a manual extraction system with minimal AFE capability for road network extraction. Our lack of SOCET SET experience meant that we also underestimated some of the subtle user interaction capabilities that allowed a SOCET SET user to be more efficient in their AFE processing than one might have expected. Finally, given that we were the developers of the AFE technology within RoadMAP we had become accustomed to many of its operational shortcomings.

Another important factor, brought out by the comparison of two completely different systems, were the engineering and user interface



(a)



(b)

Figure 12. A comparison of the consistency of the road vectors extracted by SOCET SET and RoadMAP Assisted over a portion of test area `rb2`: (a) Road vectors extracted using SOCET SET, and (b) Road vectors extracted using RoadMAP Assisted.

issues in RoadMAP. Many of these, such as limitations in the display update rate due to multiple layers of X11 display libraries, were known within the research environment but had not been fully appreciated in terms of their impact on system performance.

This evaluation highlighted one of the most important aspects of semi-automated system design, that of operator interaction with the automated processes. Not only must the operator's invocation of the automated process be fast and simple, he must be able to judge when it should be invoked and when it would be simpler to complete the task manually. For instance, a number of roads in the `cl1` test area had very low contrast against the background, compromising the tracker performance. Less-experienced RoadMAP operators would try repeatedly to get the system to track these roads, then finally do them manually, while more experienced operators would not try the automated system at all. We will be looking at ways to have the system predict when and how well it will work on a particular road or area, to guide the operator whether it is useful to invest time in using the automated system or he should just track the roads manually.

Evaluation Procedure

A large amount of effort went into planning and coordinating this evaluation, by both government and MAPSLab personnel, and the procedure served its purpose by providing a basis for system comparison. It also serves as a good lesson learned for other research groups who might want to use similar evaluation methodology, by highlighting the issues for which data were not collected.

In particular, the user-centric evaluations should automatically collect finer-grained timing information for manual and assisted AFE systems such as RoadMAP. Instead of having to estimate the time required for operations such as setup and image panning, it would have been much better to have directly measured timings. A comprehensive set of internal timings also allows the study of factors that may not have been anticipated before the evaluation. While experimental AFE systems can be instrumented to provide this information it is not practical to expect such timing hooks in commercial AFE systems such as SOCET SET, so those comparisons will most likely remain manually timed at a coarser granularity.

Our evaluation also pointed out the importance of well-defined extraction specifications. RoadMAP works at full image resolution, collecting a large number of points with a corresponding high level of detail. Operators doing manual extraction on SOCET SET typically collected points roughly suitable for a 1:50000 product, with widely-separated centerline points. Thus, in many cases, RoadMAP was producing a "more accurate" product, but this was mostly irrelevant for the side by side evaluation.

Finally, there is an issue as to the basis of comparison. When SOCET SET is used within the NGA production environment, it is configured much differently than our commercial version and runs on Sun workstations instead of the dual-processor Windows machines we were using. Establishing the "production" baseline against which RoadMAP is to be evaluated is a subtle question.

Operator Effects

It is hard to deny the performance effects of operator experience on a particular system and their overall level of expertise in data extraction. This motivated our extensive efforts to ensure that operators were well-trained on all systems. While we must be careful not to draw too many conclusions, it does appear that the RoadMAP Assisted becomes more efficient as the operator has more experience with it. This makes intuitive sense; in a semi-automated system, the operator must understand the performance characteristics of the automated processes, in order to decide when to invoke the processes.

It may well be that the extra training or experience required to master a semi-automated system will be offset by the lesser training required in data collection to meet product specifications and in the subsequent editing required to enforce the specification. Product specifications can be embedded in automated system processes or in later automated editing, instead of requiring both the operator and a subsequent editor to judge whether enough or too many points have been collected along a roadway.

Ramifications for Future Production Systems

In evaluating an automated or semi-automated system against a production system, we cannot think of the automated system as a replacement for the current manual system without considering the other aspects of production systems:

- Product specifications, or the database specification if the collected data is not for a specific product.
- Editing impacts. A significant amount of time is required for editing of collected data, whether it is manually or automatically collected. Introduction of an automated process will change the amount and type of editing required. Automated processes produce output data that is more consistent overall, in terms of geometry and points, with possible glitches where the automated process was fooled. Manual editing, on the other hand, is more likely to be concerned with lapses of operator attention which lead to accuracy degradation and non-compliance with specifications for number of points, accuracy, etc. Errors by automated systems are therefore likely to be more easily detected.
- Overall production flow. The adoption of automated processes will likely require a re-engineering of the production process to fully realize the inherent savings, in the same way that current production flows were designed and optimized for the manual collection of data for specific hardcopy products.

Future Improvements to RoadMAP

It is apparent that setup time is a dominant factor (Section *Automated Process Setup Time*), especially for short roads, since the operator must place and align the marker, then set its width. We will investigate automated alignment and width determination given only an initial point from the operator, which should significantly reduce the setup time.

An important element in the speed of SOCET SET was its use at half and quarter resolutions. We will investigate the use of lower-resolution imagery and multiple image resolutions within RoadMAP. It is possible that in many cases, reduced-resolution imagery will give adequate tracking performance. Alternatively, the use of simultaneous trackers on full- and reduced-resolution imagery would allow us to track at high speed where possible, then drop down to the more detailed imagery in problem areas.

Instead of requiring the operator to predict whether a road is likely to be successfully tracked from a particular starting point, we will indicate to the operator when he places a point whether the image contrast, edge strength, and other characteristics are likely to support a successful track. If the system tells the operator that tracking from a specific point is unlikely to be successful, he can try another starting point or choose to track the road manually. In either case, he is not spending time correcting faulty road tracks.

Conclusion

This evaluation provided valuable insight into the performance of RoadMAP, highlighting both strengths and areas for improvement. The user-oriented nature of the evaluation, with participants who were not part of the RoadMAP development team, provided an independent view of the system that cannot be gained from typical engineering evaluations. The operational user brings a different perspective and level of experience. Understanding and addressing the needs of such users is critical to the future success of systems such as RoadMAP.

The major findings of the evaluation are:

- The user-oriented evaluation methodology developed for this study can be successfully applied to understand the strengths and weaknesses of tools for automated or assisted extraction of geospatial information.
- RoadMAP is comparable to current commercial systems, such as SOCET SET.
- Detailed analysis of the road extraction process has revealed several areas where performance gains can be realized. If the projected levels of improvement are accurate, an enhanced RoadMAP system could offer more than a factor of two reduction in road extraction time.

Acknowledgments

We gratefully acknowledge the help of Ted Bulwinkle (CMU/MAPSLab), and James Crawford, Greg Glewwe, Robert Van Winkle, and Mike O'Brien (NGA), who performed the tedious extraction tasks for five days without a single OSHA violation. Doug Hugo, Jerry Lenczowski, and Scott Loomer of NGA provided instrumental support and encouragement. Scott Miller and LH Systems provided our lab with the SOCET SET software.

The research reported in this paper was supported by the National Geospatial-Intelligence Agency (NGA) under contract NMA201–01–C–0024. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Geospatial-Intelligence Agency (NGA), or of the United States Government.

References

[Airault *et al.*, 1994] S. Airault, R. Ruskone, and O. Jamet. Road detection from aerial images: a cooperation between local and global methods. In *Image and Signal Processing for Remote Sensing*, volume Volume 2315, pages 508–518, 1994.

[Fischler *et al.*, 1978] M.A. Fischler, G.J. Agin, H.G. Barrow, R.C. Bolles, L.H. Quam, J.M. Tenenbaum, and H.C. Wolf. The SRI road expert: An overview. *Proceedings of the DARPA Image Understanding Workshop*, pages 13–19, November 1978.

[Fischler *et al.*, 1981] M.A. Fischler, J.M. Tenenbaum, and H.C. Wolf. Detection of roads and linear structures in low resolution aerial images using multi-source knowledge integration techniques. *CGIP*, 15(3):201–223, March 1981.

[Harvey, 1997] Wilson Harvey. CMU Road Extraction Test Results. (Slides presented at Terrain Week '97 in San Antonio, Texas, 13 January 1997.) URL:`www.maps.cs.cmu.edu/RCVW/present/terwek97/roads/r7root.htm`. (last datea accessed: 23 September 2004).

[Harvey, 1998] W. Harvey. *A User's Guide to RoadMap*. Digital Mapping Laboratory, Carnegie Mellon University, December 1998.

[Harvey, 1999] W.A. Harvey. Performance evaluation for road extraction. *Bulletin de la Société Française de Photogrammétrie et Télédétection*, n. 153(1999–1):79–87, 1999. Colloque « Production

de Données Géographiques 3D : vers le Respect des Contraintes Applicatives ».

[Hinz and Baumgartner, 2000] S. Hinz and A. Baumgartner. Road extraction in urban areas supported by context objects. In *International Archives of Photogrammetry and Remote Sensing*, volume 33(B3), 2000.

[McKeown and Denlinger, 1986] D. M. McKeown and J. L. Denlinger. Cooperative methods for road tracking in aerial imagery. Technical Report CMU–CS–86–175, Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, 1986.

[McKeown and Denlinger, 1988] D. M. McKeown and J. L. Denlinger. Cooperative methods for road tracking in aerial imagery. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 662–672, Ann Arbor, Michigan, June 1988.

[McKeown *et al.*, 2000] David M. McKeown, Ted Bulwindle, Steven Cochran, Wilson Harvey, Chris McGlone, and Jefferey A. Shufelt. Performance evaluation for automatic feature extraction. In *International Archives of Photogrammetry and Remote Sensing*, volume XXXIII, B2, pages 379–394, 2000.

[Quam, 1978] L. Quam. Road tracking and anomaly detection. *Proceedings of the DARPA Image Understanding Workshop*, 1:51–55, May 1978.

[Trinder and Wang, 1998] J.C. Trinder and Y. Wang. Knowledge-based road interpretation in aerial images. In *International Archives of Photogrammetry and Remote Sensing*, volume 32(4), pages 635–640, 1998.

[Wiedemann and Hinz, 1999] C. Wiedemann and S. Hinz. Automatic extraction and evaluation of road networks from satellite imagery. In *International Archives of Photogrammetry and Remote Sensing*, volume 32(3-2W5), 1999.