EVALUATION AND TESTING OF SEVERAL DIGITAL PHOTOGRAMETRIC SYSTEMS FOR SYSTEM ACQUISITION BY A NATIONAL MAPPING AGENCY

Christoph Käser [1], Christoph Eidenbenz [1], Emmanuel Baltsavias [2]

[1] Swiss Federal Office of Topography, Seftigenstr. 264, CH-3084 Wabern, Switzerland

Tel./Fax +41-31-9632111/9632459, {Christoph.Kaeser, Christoph.Eidenbenz}@lt.admin.ch

[2] Institute of Geodesy and Photogrammetry, Swiss Federal Institute of Technology, ETH-Hoenggerberg, CH-8093 Zurich, Switzerland, Tel./Fax +41-1-6333042/6331101, manos@geod.ethz.ch

Commission II, Working Group 8

KEYWORDS: Digital Photogrammetric Systems, Film scanners, Aerial Triangulation, DTM, Orthoimage, Evaluation, Benchmark Tests

ABSTRACT

This paper reports on the evaluation of four digital photogrammetric systems by the Swiss Federal Office of Topography. Three of the systems were complete ones, including scanner; the fourth one provided aerial triangulation and DTM generation only. The whole evaluation process took one year. It was thoroughly planned and executed based on a long list of different evaluation criteria with varying weight, preliminary discussions and demos with the related companies and extensive benchmark tests performed at the company offices. In all benchmark tests common input data of medium to high complexity were used. The test results were quantitatively analysed using accurate reference data. The evaluation was divided in several components, of which the most important were scanner, aerial triangulation, DTM, orthoimage and mosaicking. Details on the above procedures, advantages and pitfalls to be avoided are presented.

1. INTRODUCTION

Users of photogrammetric technology are faced with the problem of evaluation when buying new systems. Evaluation of complete digital photogrammetric systems with various components is more difficult due to the complexity of the systems and critical due to their high cost. A thorough and successful evaluation is particularly important for organisations, public and private, involved in production, which often requires high product quality, fast generation, system reliability and low costs. Although certain organisations, especially public ones, have already performed evaluations of digital photogrammetric systems, almost nothing is published on the followed procedures, their advantages and pitfalls. Other users, e.g. from small private firms but also universities, sometimes buy without a thorough evaluation but rather based on some limited tests, short demos, opinions of friends and colleagues etc. It is our belief, that in all cases a carefully planned and executed evaluation has higher probability to lead to a correct decision, which may spare a lot of subsequent problems and costs. In addition, through the evaluation process the personnel of the involved organisation is gaining knowledge and experience and the transition to new, digital procedures becomes easier. Based on this motivation, we report in this paper on the experiences gained with the evaluation of four digital photogrammetric systems (DPS) at the Swiss Federal Office of Topography (L+T). Due to lack of space we will mainly report on the procedure and the different processes. Detailed results on some components (scanner, DTM and orthoimage generation) can be found in Baltsavias and Kaeser, 1998a, 1998b, while some results on the aerial triangulation (AT) will be presented here. No mention of concrete company names will be made, since the presented evaluation procedure is independent of specific companies.

2. THE EVALUATION AND ACQUISITION PROCEDURE

In 1996 the L+T evaluated and acquired a DPS consisting of a film scanner, a data server with aerial triangulation software and a digital photogrammetric workstation in order to implement a transition from analytical to digital processing techniques. Since the system cost was expected to exceed the limit of 263'000 SFr., the whole evaluation and acquisition had to be performed under the Gatt/WTO conditions.

Taking into account these conditions, the evaluation and acquisition procedure included the following steps:

- On 26th February 1996, the *announcement* of the *selective approach* was published in the Official Swiss Commercial Newspaper (SHAB). Within the next 25 days, companies could express their intent for participation.
- Till 22nd March 10 candidates announced their *request for participation* but only 6 fulfilled the conditions of an appropriate company. The rest were unknown companies without any proof of experience in digital photogrammetric systems.
- On 4th April the candidate companies received the required system specifications (*invitation for tender*) and then had 40 days time to make an *offer*.
- On 19th April the „information day" took place where the companies were informed about the system requirements and had the possibility to get answers to their questions.
- Till 24th May 5 offers were received. In one case, the offer consisted of products of more than one company. One of them offered only parts of the required system (AT and DTM) and

was considered only later in accordance with the other competitors (called system 4 in the sequel).
- In June and July, the offers were checked, and technical discussions with the companies took place, as well as the first system demonstrations during the ISPRS Congress in Vienna.
- Intensive system tests (*benchmarks*) took place in August and September. In one case, the system development was so recent that the benchmark could take place only in the USA, a condition that could not be accepted.
- During September and October the test data was analysed and evaluated in cooperation with ETH Zurich.
- From middle of October to middle of November contract negotiations took place with the company with the relatively best offer in price and performance. The so-called nomination of the final offer was approved by the board of directors of L+T.
- At the end of November the *award of contract* took place, the contract was signed and the system was delivered in December 1996.
- On 24th February 1997, the *bid winner* was made publicly known in the Official Swiss Commercial Newspaper (SHAB).

The whole procedure lasted one year, although the most relevant evaluation part (benchmark tests and result analysis) had a duration of only three months.

3. BENCHMARK TESTS

3.1 Some Principles

The focus of the whole evaluation, from a technical point of view, lied on the benchmarks. Firstly, a detailed list was set up on what should be tested and how. Thereby, the test was divided in several components, with the main ones being scanner, digital aerial triangulation, automatic DTM generation, orthoimage generation and mosaicking. Secondly, for each of these components the necessary test data was carefully thought and selected in advance. Thereby, the following considerations were made:
- due to time limitations both during the benchmarks and for the data analysis, the test data, e.g. for DTM generation or AT, did not cover all possible land cover types and configurations. Instead, characteristic cases for the L+T of medium to high degree of complexity were chosen.
- ground truth data were acquired to permit a quantitative analysis of the results.
- methods for processing and analysis of the results were considered a priori to ensure that the data could be timely processed.
- the same input data and data analysis methods were used for all four systems, whereby efforts were made to use the same starting conditions, e.g. orthoimages were generated starting from the same DTM and sensor orientation to allow an objective comparison of the planimetric accuracy, and for the DTM generation the same exterior orientation was used. Of course the output specifications, e.g. test area and grid spacing in DTM and orthoimage generation were the same for all systems.

To be sure that the test data could be read, a data set (digital images and map, vector data, GPS data, etc.) was delivered to the manufacturers one month before the benchmark.

The test data itself was delivered only at the beginning of the benchmark. It consisted of camera calibration protocols, ground control point (GCP) coordinates, GPS camera stations, the DTM for orthoimage generation, a 1:25'000 scanned topographic map, and DXF vectors, acquired at an analytical plotter, in the region used for DTM and orthoimage generation. Only the 30 digital images for the AT were sent to the companies some days in advance because it takes some time to read them and, whenever necessary, convert them into the system's internal format.

It must be noted that in certain tests we did not specify how some parameters should be selected, although the quality of the results greatly depended on this selection. Thus, we let the companies decide on and perform the scanner calibration, determination of minimum and maximum density in scanning, choice of number and position of tie point regions in AT, selection of matching strategy in AT and DTM etc. It was assumed that the companies had sufficient expertise for the optimal choice of these parameters, although, as it was proved later, this was not always the case (which naturally resulted in additional minus evaluation points).

3.2 Test Data

The test data included the following. For the scanner test a glass grid plate with 25 x 25 crosses and 1 cm spacing, a calibrated grey scale wedge (0.05D – 3.1D), a resolution pattern, the empty scanner stage glass plate, as well as B/W and colour films were scanned with high resolution. Some test patterns were scanned with different resolutions (to check the influence of pixel size), and in colour even if they were B/W (to check differences between spectral channels with respect to geometry and radiometry).

For AT a block of 3 x 10 (in N and E direction respectively, with E/W being the flight line direction) B/W images over hilly and quite rough terrain with 40% - 45% forests, villages and small towns was used. The forward and side overlap was 60% and 30% respectively. The images cover the 1:25,000 topographic map sheet 1095 „Gais" with an area of 12 x 17.5 km^2. The height range of this region is 400 – 1500 m, but the majority of the area has 700 – 1100 m height. 25 well-defined GCPs evenly distributed over the whole region were measured with GPS and an estimated accuracy of ca. 10 cm. 5 GCPs were used as control and the rest as check points.

For DTM generation, one B/W model over hilly terrain (height range 340 m) including forests, rivers and creeks as well as urban regions was used. In this model, the reference data included a DTM measured at an analytical plotter excluding points on or close to trees, buildings and other nonterrain objects (16'400 points), as well as a separate file only with breaklines (1100 points), in order to check separately the quality of the systems with respect to these important terrain features.

For orthoimage generation, a colour stereopair was used (same region used for DTM), whereby the DTM and the sensor orientation were delivered from L+T to all companies. The orthoimages, each from left and right image, were subsequently mosaicked.

All used images were scanned at a photogrammetric scanner with high geometric accuracy and a pixel size of 15 µm and had a scale of 1:30'000 (typical scale that the L+T is using). The camera was a Leica Wild RC30 with a 152 mm lens, and was

connected to a GPS for measurement of the camera station positions.

3.3 Tests and Evaluation

The benchmarks were performed at the companies in presence of 2-4 persons from the L+T and ETHZ and had a duration of 2 days each. During these two days, the system data were generated according to our scenario. At the same time, a prepared questionnaire (incl. notes and remarks) was filled out. For the evaluation, a range from 0 to 3 points per item was used.

The questionnaire included for each test component general remarks like functionality, workflow, handling, processing and computing times, but also remarks on the overall impression. Some generated data was coarsely checked on-site. E.g., the mosaic was checked by overlaying the scanned map and the DXF vectors, while the generated DTM was checked using stereo display. Latter was also used to check the functionality for data acquisition and processing, as well as the whole ergonomy and ease of use.

Aspects like software development environment, data management and archival, as well as feature extraction were examined with small system demonstrations and critical questions.

4. TEST CRITERIA, PROCEDURES AND PARTIAL RESULTS

After the benchmarks, the questionnaires of the 2-4 participants were combined and unified. The data analysis and quantitative evaluation took place at ETH Zurich and required ca. four manmonths. All the results were evaluated per manufacturer according to a common key and a detailed report was prepared. The following aspects were examined.

4.1 Scanner

Details, especially technical specifications, of the scanners were collected before the benchmarks. Remaining questions were clarified later. Apart from scanning the test patterns, the software was examined in detail, and especially the calibrations (geometric, radiometric, spectral), and the degree of automation and the ease of setting of the scan parameters (particularly the film density determination). The quantitative analysis included the following.

Geometric accuracy. The grid plate crosses were measured with Least Squares Template Matching. They were transformed to the reference coordinates by an affine transformation using as control all, 8, and 4 points. For each transformation, the RMS, average with sign (bias) and maximum absolute values of the errors were computed. All residuals were plotted to detect systematic errors, especially between the tiles or swaths (for area and linear CCDs respectively) of the scans. This was performed for the R, G, B channels (colour scan). The pairwise differences of the pixel coordinates between the 3 colour channels were computed to estimate colour misregistration errors and similar statistics and plots as above were generated. In addition, edges vertical to the seam lines between neighbouring tiles or swaths were visually inspected. Broken edges clearly indicated geometric misregistration between tiles or swaths.

Radiometric accuracy. The grey scale film was scanned in colour (in one case with two pixel sizes, in another one with two integration times). In all cases, a linear LookUp Table was used. Subimages at the centre of each film density were used to calculate the mean and standard deviation of grey values. Based on these values, the noise level, dynamic range and system linearity (latter by plotting the calibrated densities versus the logarithm of the mean grey values) were estimated. This procedure was performed for each colour channel, allowing thus estimation of the spectral variation and goodness of colour balance. Comparison of the results with different scan pixel sizes and integration times showed the influence of these two parameters on the radiometric performance. In the system where two pixel sizes were used, the comparison even correctly indicated that the subsampling for the coarser pixel size was wrong, i.e. the scanned film area for each pixel was smaller than it should be. Multiple scans of neighbouring tiles or swaths using the scanner glass stage were used to estimate the temporal and spatial variation of noise. Finally, the histograms of the B/W aerial films were computed. They were very different among the three systems and in two of them in the bright grey values every fourth value occurred much more frequently than its neighbours were.

Geometric resolution. A USAF resolution pattern on glass was used for visual estimation of the resolution in horizontal and vertical direction.

Artefacts and radiometric problems. The scanned patterns including the B/W and colour aerial films were contrast enhanced and visually checked. Thus, several noise patterns like vertical and horizontal stripes, electronic dust, radiometric differences between neighbouring tiles or swaths, repetition of the signal of one spectral channel in the remaining ones etc. could be detected and then the errors were quantified using the original images. In one case, even data losses in the transfer from scanner to host were observed.

4.2 Aerial Triangulation

In AT, apart from accuracy investigations, functionality and execution time were examined in detail. Thereby, the following observations were made.

The project and camera set-up were usually graphically aided, manually input and easy-to-use. In the block definition, there were quite some differences. With most of the systems the GPS camera stations could be imported interactively or in batch mode from an ASCII-file and the flight lines had to be defined by just using the first and last image of each strip. A block overview with image thumbnails, image status (pyramid, interior orientation) and image orientation was in all but one case standard. In addition, one system displayed image overviews as wireframes.

The execution time for the image pyramid generation laid between 1 - 3 minutes per image and it could be started both from the GUI and in batch mode. For the interior orientation, it was standard to measure 2 fiducials manually and the rest was measured automatically. Only one system with fully automated fiducial measurement did not support semi-automated one. Of course, every system allowed manual fiducial measurement. The execution time was between 15 - 80 seconds per image and the result was stored in a protocol file with all other measurement parameters.

The semi-automated way to measure ground control points was generally allowed but every system had a different implementation/interface, and in some systems the procedure did not work properly. In all systems, the control point had to be measured in one image first. One system allowed a fairly precise positioning in a zoomed image overview and then the point was matched automatically in all considered images. On the other hand, one system selected the approximate positions in all images automatically, and then matching was used for the fine positioning, but the approximations had to be fairly accurate. The interactive point editing with zoom, pan, image overview and edit measurement was satisfactory in all systems.

The tie point regions had in some systems a fixed number, in others were freely defined. Their position and size were either defined interactively, using e.g. a model image, or read from a strategy file. One system used a subdivision of the regions. To find the homologue regions some systems needed a block initialisation with/without generation/use of a DTM, others used an average terrain height. The quality of the approximations of the tie points, which depend on the possibility/need of using GPS camera stations and a DTM, is very critical for successful matching. Cases critical for matching, like steep slopes, no texture, forests, large water bodies, large scale urban regions, moving shadows, could not all be covered by the test image material and were not examined in detail. In case of no/few match-points in the tie regions, one system tried to automatically redefine these regions, while all others required manual intervention. Some systems could treat arbitrary scanned image rotations, others required a manually performed prerotation. A big difference was noticed in the way of automated tie point measurement (model by model - all involved images simultaneously), the amount of matched tie points per image (15 - 1500), the execution time for the whole block (16 - 200 min) and the achieved accuracy (see Tables 1 and 2).

All but one system, which used external third-party products, offered an integrated bundle block adjustment. Some major differences between these programs existed with regard to use of GPS camera station observations, elimination of blunders, use of additional parameters, and inversion of normal matrix. The calculation times were very fast (2 minutes for 180 points - 10 min for 32'000 points) but the achieved accuracies very different.

A big lack was a tool to find easily and fast weak connections and/or inaccurate observations. Only in one case, the quality control and point measurement was possible in stereo mode. This is good for point interpretation but in practice, it is not needed very often. When the matching algorithm has a good internal quality control, the only remaining problem is weak connection of the images. In this case, additional tie points have to be measured, which is faster in stereo.

The last function tested was the export of the orientation elements into an ASCII-file, which was no problem and to an analytical plotter, which was in all systems a question of interface implementation.

Table 1. Accuracy measures for the AT [m]

System		Residuals' RMS		
	Points *	DY	DX	DZ
1	5 control	0.03	0.03	0.02
	16 check	0.33	0.52	1.52
2	5 control	1.38	0.50	0.86
	20 check	0.39	0.41	0.65
3	5 control	0.22	0.13	0.52
	19 check	0.47	0.69	0.63
4	5 control	0.05	0.03	0.02
	19 check	0.22	0.39	0.42

* Although the use of all 20 check points was requested, this information was received only for system 3.

Table 2. Comparison of camera stations for a stereo model [m] *

	YO	XO	ZO	DY	DX	DZ
Photo 1331 **	751759.13	244014.80	4463.75	-	-	-
system 1	751760.88	244018.16	4464.20	1.74	3.36	0.44
system 3	751759.76	244013.81	4463.91	0.62	0.99	0.16
system 4	751759.30	244013.37	4463.52	0.16	1.43	0.23
Photo 1332 **	750066.51	244029.85	4467.48	-	-	-
system 1	750067.28	244033.31	4468.42	0.77	3.45	0.94
system 3	750067.52	244029.36	4468.18	1.00	0.49	0.70
system 4	750067.21	244029.09	4467.70	0.70	0.77	0.22

* No results received for system 2.

** Reference values.

The accuracy aspects were examined by use of control/check points and, to a lesser extent, the orientation of a stereo model in the block centre measured at an analytical plotter. The block was fixed using 5 control points (4 corners and centre of the block). Certain check points were not ideal for accuracy analysis, since they were close to the control points or lying at the weaker block border. With some systems many AT trials had to be performed before delivering after the benchmark the results. Some systems did not provide standard deviations for the control points or even σ_o, others were listing standard deviations of the control points, although no inversion of the normal matrix took place, and no explanation was given on how these values were computed. Thus, for the accuracy analysis we used the differences between adjusted and GPS coordinates to calculate the residuals' RMS in Table 1. In the table, Y represents Easting (flight line direction) and X Northing. The values for the control points with systems 2 and 3 are very high. For system 3, it was necessary to reduce the weights of the control points, otherwise the block could not be adjusted. For system 2, no explanation was provided by the company. In Table 2 the camera positions estimated by the systems are compared to the reference values. From the tables it can be seen, that only system 4 provided reasonable and understandable results.

4.3 Orthoimage and Mosaic

The 0.5 m pixel size orthoimages were generated using the Swiss National DTM (DHM25) and a bilinear interpolation. 11 GPS points existed in the overlap region of the orthoimages. An interpolation of these points in the DHM25 revealed differences up to 5 m. At positions of large differences, the DHM25, and thus also the planimetric position in the orthoimages, were erroneous, so an accuracy analysis using these points was not reasonable. Thus, only 4 GPS points with differences less than 2 m were kept.

The radiometric quality of the orthoimages was visually controlled. Surprisingly, one system generated orthoimages that were totally saturated in the bright regions, leading to loss of many details, and a second one images with a coarse resolution (although the pixel size was 0.5 m according to the specifications), as if the orthoimage were generated from a higher pyramid level, or by a nearest neighbour interpolation (although bilinear was specified).

With respect to geometry, first the relative accuracy was checked. For each system, about 50 well-defined and well-distributed points lying on the ground in the overlap region were selected in one orthoimage and were transferred by semi-automatic Least Squares Matching (LSM) in the second one. The pixel coordinates of the two orthoimages should ideally differ by a constant known offset (difference of origins of two orthoimages), while the standard deviation of the differences should ideally be zero. The actual standard deviation showed relative errors between the two orthoimages, while the offset error showed a systematic shift of both orthoimages. The relative error was 0.4 - 1 pixel, while the shift error was 0 – 1 pixel. The differences between the systems were not very big. The errors in the Y (base) direction were larger than in X.

For the absolute accuracy check the 4 GPS and ca. 40 additional control points were used. These 40 points were well-defined and well-distributed points lying on the ground in the overlap region. They were selected in one orthoimage and transferred in the remaining five by LSM. Using these pixel (planimetric) coordinates for each orthorectified image pair, the heights interpolated in the DHM25, the known interior and exterior orientation and the procedure described in Baltsavias, 1996, correct object coordinates for these points could be derived, even if the DHM25 was locally erroneous. Then, statistical measures of the differences between known and measured coordinates were estimated. Again, the differences between the systems were not big, with RMS in the range of 0.6 – 1.4 pixel and the errors were larger in the base direction.

The geometric accuracy of the two mosaicked images was checked by using the above mentioned relative errors and visual control of high contrast straight edges crossing the seam line (broken edge in case of misregistration). The radiometric balance was checked visually in the region along the seam line.

It was interesting to note that it proved quite difficult to get from two companies the planimetric coordinates of the origin of the orthoimages. In one case, even a wrong answer was given. In addition, the orthoimage generation protocols either did not provide this information or it was hidden among a pile of numbers without any explanation.

4.4 Digital Terrain Model

With one system two software versions were used (old and new Beta version). With another one (the one expected to be among the best) 15 and 30 µm images were used to check the effect of pixel size on DTM accuracy. The product of all systems was a regular 10 m grid DTM. Two of the systems had an identical DTM module but the user interface and the match-parameter settings differed.

The two sets of reference values (DTM, breaklines) were interpolated in the automatically generated DTMs. Statistical values of the differences were computed, as well as error histograms using predefined classes. For the DTM the RMS was 0.6 – 1.6 m, the average with sign 0.1 to 0.8 m, and the maximum absolute error 6 – 65 m. For the breaklines the accuracy was 1.5 – 2 times worse and the respective values were: RMS 1 – 2.4 m, average 0.3 – 1.7 m, maximum absolute 3.7 – 8.8 m. The errors were sorted and the larger ones were overlaid on an orthoimage to check the position of these points and try to explain the failure reasons. As expected, large errors occurred at or close to surface discontinuities, perspective differences, low texture, and edges parallel to the epipolar lines. With systems trying to filter out buildings etc. errors also occurred at ground points that were erroneously corrected in order to fit them to the majority of the neighbouring points which lied on nonterrain objects, e.g. points in small forest openings. Two systems showed very large to huge errors at the border of the overlap region.

Additional methods to visualise the errors included: contours (detection of gross errors, quality of geomorphologic details and breaklines, noisiness), their comparison to the map contours and their overlay on orthoimages, 3-D wireframe models from different views, representation of the automatically generated DTMs as grey level images (detection of gross errors and quality of geomorphologic details), generation of error contours and their overlay on the orthoimage.

The results with the 30 µm images as compared to the 15 µm ones were 15-20% worse in RMS, very similar in the average

with sign, and very similar or even better in the maximum absolute error. Methods that used very dense measurements and then a thin-out with parallel blunder detection performed better. Cheaper modules, encountered also in remote sensing packages, performed better than some more expensive ones. The two systems with identical DTM modules differed a lot (by a factor 3 in RMS) proving how sensitive the results are to the selection of the match-parameters, and the difficulty of appropriately setting them even for expected experts.
Other important evaluation criteria included execution time (varied a lot) and tools to visualise and edit the results (sufficiently good tools only with one system).

4.5 Stereo Display

It is not easy to judge the ergonomy of the stereo display of a softcopy station. During the benchmarks the 3-D image quality was evaluated only within some minutes which is not comparable to the conditions met by an operator who is working 6 - 8 hours every day on the system. Concerning the image quality, clarity, phantom images, smoothness of movement and brightness/contrast adaptation were closely examined. One system used a passive stereo display, the rest active glasses.

Further items of the questionnaire dealt more with the ease of use and functionality in stereo mode and were the following:
- preparation: image selection, calculation of epipolar images, calculation time
- model overview and speed of movements (image reload)
- editing: points, lines, surfaces, groups, segments, snap, pan/scroll, add, delete, replace
- superimposition: grid model, contour model, DXF vectors,
- automatic positioning of the cursor on the surface

There were no significant differences between the systems.

4.6 Feature Extraction

This component is a big research topic. Only one system could demonstrate some results in the case of a semi-automated building extraction. Other feature like roads, forests, creeks or lakes are not yet solved. Our wish to use existing 2-D vector data and image matching in order to determine them in 3-D could not be fulfilled by any system. The alternative proposed was to first generate a DTM and then use it to interpolate the heights of the 2-D vector points.

4.7 Data Management and Archival

During the demonstration, the data management tools were closely examined. Parameters for project and image management, selective data storage for one block with integrated management of image files and metadata (e.g. orientation elements) were subject of the qualitative aspects considered. The systems had different solutions. One put all information in an ASCII-file, another one used a database to store this information, and the third one had all this information in binary files. During the benchmarks all these solutions worked without any problems.

Another point was the data archival. All companies considered archival as a very important part of a digital photogrammetric system, but no one could show us a solution. The usual answer was that this is a development topic for the next year.

4.8 Development Environment

During the demonstration of the development environment the focus was on items like
- handling: structure of a software application, integration of own software
- development environment: debugger, libraries, fullscreen text editor
- program languages: C, C++, macro languages.

All systems used the development environment of the operating system (Unix from SGI or SUN) and offered the required program languages, fullscreen text editor and debugger. Only one system had a special development environment for the photogrammetric applications and consequently not accessible libraries, thus not allowing integration of own software modules.

4.9 Overall Benchmark Impression

After the benchmark, personal impressions of the system configuration and demonstrations, as well as general impression of the software were summarised.

The system configurations used in the benchmark tests were good and the system computer performance was similar. There was a big difference in the quality of the demonstrations. Although serious problems during every benchmark occurred, it was very interesting to see how the different companies handled these problems.

As a general impression, GUI, good image quality on screen and input command tools (mouse, keyboard) as well as execution of modules in batch mode are standard. The workflow in every system is generally well thought but there are some differences in user guidance, online help (none to very good) and the number of opened windows. In some cases there were too many open windows, leading to an easy loss of the overview and complex handling.

5. COMBINATION OF EVALUATION CRITERIA AND FINAL RESULTS

Firstly, the evaluation of the benchmark was checked with a so-called sensitivity analysis. Thereby, the influence of changing weights per evaluation criterion and their impact on the ranking list was observed. Thus, a more reliable result for the whole evaluation was achieved. Based on the L+T needs, the most important components (scanner, AT) were weighted the most, followed by orthoimage and benchmark impression, and then the remaining components. This was a subjective, L+T – specific decision guided by certain facts. E.g. the L+T has recently completed a nation-wide coverage with a 25 m grid DTM, thus this component was not of primary importance. Four different weighting schemes were used: uniform (i.e. all criteria had same weight), optimal (weights adjusted to the L+T needs and aims), minimal (less important components and criteria had zero weight or were down-weighted). Two slightly varying optimal weightings were used, but in the tables, only one is listed for readability purposes. The three different cases of weighting are shown in Table 3. The weights of the first scheme are not directly comparable to the ones of the other two, since the weight sum (155, showing the large number of evaluation criteria used just for the benchmark test) was different. The maximum number of points was computed by multiplying the weights with the maximum score of 3.

In Table 4, the evaluation of the benchmark with the sensitivity analysis is shown. In all cases, system 3 was the best. The big difference between the total scores of Table 4 and the maximum possible scores of Table 3 shows that even for the best system there is a lot of space for improvement. The total score of system 4 in Tables 4 and 5 can not be directly compared to the ones of the rest, since the system offered only some components.

Finally, for the overall evaluation the following components were used: technical (focus on accuracy), system demonstration (benchmark), company impression (general/support), and costs. The system comparison was made again with a sensitivity analysis. For all weighting schemes, the best system was a fictitious system 5, which consisted of the AT from system 4 and the remaining components from system 3 (the DTM modules of these two systems were almost identical). The results for the optimal weighting scheme are shown in Table 5.

Table 3. Benchmark test: weighting schemes and maximum number of points (Pmax) for various evaluation components

Version	Uniform		Optimal		Minimal	
	weight	Pmax	weight	Pmax	weight	Pmax
Scanner	25	75	40	120	50	150
Aerotriangulation	40	120	40	120	50	150
Orthoimage & Mosaic	23	69	30	90	30	90
DTM	12	36	15	45	10	30
Stereo Display	13	39	10	30	10	30
Feature Extraction	7	21	5	15	0	0
Data Management & Archival	7	21	20	60	20	60
Development Environment	8	24	10	30	0	0
Benchmark Impression	20	60	30	90	30	90
TOTAL	**155**	**465**	**200**	**600**	**200**	**600**

Table 4. Benchmark test: results for various evaluation components and weighting schemes

Version	Uniform				Optimal				Minimal			
System	1	2	3	4	1	2	3	4	1	2	3	4
Scanner	39	36	42	-	65	62	68	-	88	65	103	-
Aerotriangulation	71	55	80	88	71	51	80	90	97	68	105	116
Orthoimage & Mosaic	43	36	40	-	50	39	57	-	44	36	51	-
DTM	27	18	20	22	29	27	18	32	21	22	16	24
Stereo Display	23	19	21	-	18	12	20	-	21	14	24	-
Feature Extraction	4	0	0	-	7	0	0	-	0	0	0	-
Data Management & Archival	4	9	11	8	19	37	46	35	16	33	41	32
Development Environment	19	13	13	-	25	15	17	-	0	0	0	-
Benchmark Impression	35	35	41	27	49	53	67	40	50	54	69	38
TOTAL	**263**	**219**	**267**	**144**	**332**	**294**	**372**	**197**	**337**	**291**	**409**	**210**

Table 5. Overall evaluation: weighting schemes and maximum number of points (Pmax) for various evaluation components, and results for the optimal weighting scheme

Version	Uniform		Minimum		Optimal						
Weight / Pmax / System	weight	Pmax	weight	Pmax	weight	Pmax	1	2	3	4	5 *
Technical (accuracy)	**9**	**27**	**20**	**60**	**20**	**60**	**40**	**18**	**46**	**18**	**52**
Scanner	1	3	4	12	4	12	8	4	12	-	12
Aerotriangulation	1	3	4	12	4	12	4	0	8	12	12
Orthoimage & Mosaic	1	3	3	9	3	9	9	3	6	-	6
DTM	1	3	2	6	2	5	3	0	3	5	5
Stereo Display	1	3	2	6	1	3	1	2	3	-	3
Feature Extraction	1	3	0	0	1	2	2	0	0	-	0
Data Management & Archival	1	3	2	6	2	6	4	4	4	-	4
Development Environment	1	3	0	0	1	3	3	2	1	1	1
General	1	3	3	9	3	9	6	3	9	-	9
Benchmark	**11**	**33**	**20**	**60**	**20**	**60**	**31**	**20**	**49**	**22**	**58**
Scanner	1	3	4	12	4	12	8	4	12	-	12
Aerotriangulation	1	3	4	12	4	12	4	0	8	12	12
Orthoimage & Mosaic	1	3	3	9	3	9	6	3	9	-	9
DTM	1	3	3	9	2	5	3	2	0	5	5
Stereo Display	1	3	1	3	1	3	2	1	3	-	3
Feature Extraction	1	3	0	0	1	2	2	0	0	-	0
Data Management & Archival	1	3	0	0	2	6	0	4	6	2	6
Development Environment	1	3	0	0	1	3	3	1	2	-	2
General	1	3	2	6	2	6	2	4	6	-	6
Demonstration	2	6	3	9	1	3	1	1	3	3	3
Company	**4**	**12**	**10**	**30**	**8**	**24**	**8**	**8**	**24**	**24**	**24**
Support	2	6	5	15	4	12	4	4	12	12	12
General	2	6	5	15	4	12	4	4	12	12	12
Costs	**10**	**30**	**10**	**30**	**12**	**36**	**23**	**13**	**32**	**10**	**33**
System	1	3	2	6	2	6	4	2	6	-	6
Scanner	1	3	1	3	1	3	1	3	2	-	2
Aerotriangulation	1	3	1	3	1	3	3	0	1	2	2
Orthoimage & Mosaic	1	3	1	3	1	3	2	1	3	-	3
DTM	1	3	1	3	1	3	1	0	3	2	3
Stereo Display	1	3	1	3	1	3	3	1	2	-	2
Maintenance	1	3	1	3	1	3	3	3	3	3	3
2nd photogrammetric system	1	3	2	6	2	6	4	2	6	-	6
Licenses	1	3	0	0	1	3	2	0	3	1	3
Education / Training	1	3	0	0	1	3	0	1	3	2	3
TOTAL	**34**	**102**	**60**	**180**	**60**	**180**	**101**	**59**	**151**	**73**	**166**

* Fictitious system consisting of a combination of systems 3 and 4.

6. DISCUSSION AND CONCLUSIONS

Good knowledge of the theories involved in digital photogrammetric processes and of the algorithms used in the different system modules is essential. It allows a better design of the evaluation process, definition of appropriate evaluation criteria, formulation of proper questions and tasks for the benchmarks, and supports a better result analysis (search for and explanation of errors etc.). In the cases where important algorithmic details had not been published, e.g. in some DTM matching procedures, the companies were requested to provide additional information and explanations. Concerning this point, the cooperation between L+T and ETH Zurich proved to be optimal. ETH provided its theoretic/algorithmic know-how and experience from previous evaluations of digital photogrammetric components, while L+T brought its practical experience, good definition of theirs needs and preparation of the test data, had the overall guidance, decided on the evaluation criteria, their weights, each company's score, and of course the bid winner.

The division of each benchmark in the main test components scanner, AT, DTM and orthoimage, each with own independent input data, was proven very appropriate. On one hand, this allowed a test and evaluation of each component, independently of previous results. On the other, if during a benchmark a grave problem occurred in one component, work could continue on another one without time delay.

The 2-day duration for each benchmark was rather short for the extent of the tests. Thus, some results had to be delivered later, partly also due to software errors, poor performance due to wrong parameter settings and in one case even insufficient computer RAM. Three days would be more suitable. The presence of at least two persons in the benchmarks, demos and all critical meetings is essential to enable higher objectivity and more complete understanding and following of the complex processes. The selection of test data and quality of reference values proved to be very sufficient, with the exception of AT. A larger image block and more GCPs in the nonborder images would allow a more reliable evaluation of the achieved accuracy. During the benchmarks, a few new tests were recognised to be useful and added to the list (e.g. scan of the grey scale film in colour). This, however, was rare and by no means influenced the evaluation outcome.

The analysis of the results was made more difficult due to the poor quality of protocol files and output listings (relevant information missing, numbers listed with no or unclear indication of what they represented, huge listings in the AT with much useless information, while important one was missing). Some algorithms provided some precision indicators, e.g. in DTM generation and AT. Often these indicators were optimistic and far away from the empirical accuracy values, and thus are not to be trusted. The evaluation was made more complex and time consuming by testing new software versions (sometimes in addition to the older ones). This can be a greater problem with rapid updates when the evaluation is lengthy. Naturally, the newest available software/hardware should be tested, but this should be of a rather stable instead of Beta** condition.

As shown in chapter 5, an evaluation partially involves a personal view on the systems and companies. This is clearly shown by the introduced weights per criterion or some subjectively evaluated criteria like the appraisal of the companies. On the other hand, a more objective evaluation by use of the sensitivity analysis could be achieved. The outcome and decision to use components from two companies might seem strange and a disadvantage on first thoughts. However, it is clear that no company provides the best product in all system components. As long as the quality difference between the products is significant, and a communication between the different systems, at least through data exchange modules, is secured, then a combination of different systems, in spite of certain disadvantages, should not be excluded.

All companies were very helpful and responsive. However, certain weaknesses beyond the classical „demo effects", some of which are serious, should be noted. None of the companies was proven ready and well prepared for a test of this extent, although all were well informed in advance about the test data and test scenario. In all cases, algorithmic details are known only by 1-2 development experts who usually work at the headquarters (luckily for us they were still employed by the companies). The personnel giving the demos and performing benchmarks should thus improve its knowledge of the underlying algorithms, the effect of parameters on the results, and important implementation details (some persons did not even know some aspects mentioned in the publicly available manuals). This weakness also led to many parameter trials and errors until reasonable results could be produced. Sometimes, very new versions of software modules were used without sufficient previous testing or even a single trial (a characteristic example was the generation of a colour orthoimage with a new module; instead an image including nine repeated, small, B/W orthoimages each with a white border was produced). Some companies did not follow the test prescriptions (e.g. DTM generation in the wrong area), leading thus to extra work for both companies and evaluators. In other cases, they delivered incomplete data or did not deliver the parameters of the algorithms (e.g. for DTM generation), making thus the analysis and comparison difficult.

None of the systems fulfilled to a large extent the specified needs. Although one and half years have passed since the tests and some improvements have been performed in between, today the DPS have still many weaknesses, instabilities, poor performance or outdated algorithms, missing or complex functionality etc. They have matured since their introduction at the beginning of the 90s, but significant improvements could and should be achieved.

The extent of the evaluation was partly guided by time and cost constraints. Other organisations can further extend or downsize such evaluations depending on their conditions. In spite of small pitfalls, the evaluation procedure was very successful and contributed in the correct system choice. Knowledge of the system performance also enabled a definition of proper acceptance conditions in the contract with the company that won the bid and their check with a second evaluation of certain components after system installation. In addition, it helped gain knowledge and experience, which made the transition to digital processing techniques and use of the installed system easier. The here presented procedures are general enough and can be used by various organisations in similar evaluations. L+T – specific aspects included the criteria weights, image scale, and terrain slope and cover, factors, which should be adapted to the needs of each organisation.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous companies for permitting the publication of the evaluation results.

REFERENCES

Baltsavias, E.P., 1996. Digital Ortho-Images – A Powerful Tool for the Extraction of Spatial- and Geo-Information. ISPRS Journal of Photogrammetry & Remote Sensing, Vol. 51, pp. 63–77.

Baltsavias, E.P., Kaeser, Chr., 1998a. Evaluation and Testing of the Zeiss SCAI Roll Film Scanner. In: Int'l Archives of Photogrammetry and Rem. Sensing, Vol. 32, Part 1, Bangalore, India, pp. 67-74.

Baltsavias, E.P., Kaeser, Chr., 1998b. DTM and Orthoimage Generation – A Thorough Analysis and Comparison of Four Digital Photogrammetric Systems. In: Int'l Archives of Photogrammetry and Rem. Sensing, Vol. 32, Part 4, Stuttgart, Germany (to be published).