

## DTM AND ORTHOIMAGE GENERATION – A THOROUGH ANALYSIS AND COMPARISON OF FOUR DIGITAL PHOTOGRAMMETRIC SYSTEMS

Emmanuel Baltsavias<sup>1</sup>, Christoph Käser<sup>2</sup>

<sup>1</sup>Institute of Geodesy and Photogrammetry, Swiss Federal Institute of Technology, ETH-Hoenggerberg, CH-8093 Zurich, Switzerland, Tel./Fax +41-1-633 3042 / 633 1101, manos@geod.ethz.ch

<sup>2</sup>Swiss Federal Office of Topography, Seftigenstr. 264, CH-3084 Wabern, Switzerland  
Tel./Fax +41-31-963 2111 / 963 2459, Christoph.Käser@lt.admin.ch

Commission IV, Working Group 2

**KEYWORDS:** Digital Photogrammetric Systems, DTM, Orthoimage, Evaluation, Benchmark Tests

### ABSTRACT

This paper reports on the evaluation of digital photogrammetric systems by the Swiss Federal Office of Topography with respect to DTM and orthoimage generation. Four systems (LHS DPW 770, Zeiss PHODIS, Autometric Softplotter, INPHO Match-T) were used for DTM, while the first three were used for orthoimage generation. The evaluation process 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 qualitatively and quantitatively analysed using accurate reference data. In DTM generation, the differences between the systems are significant and for each system the results heavily depend on the choice of many and not well explained matching parameters. In all cases, blunders remain in the results and since there is no reliable indication about their position, time consuming manual editing of all points is required. Satisfactory pre- and post-editing tools rarely exist. Even in orthoimage generation, the radiometric differences between the systems were higher than expected.

### 1. INTRODUCTION

Users of photogrammetric technology are faced with the problem of evaluation when buying new systems. 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. In this paper, we will report on the experiences gained with the evaluation of four digital photogrammetric systems (DPS) at the Swiss Federal Office of Topography (L+T).

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. After first system demonstrations during the ISPRS Congress in Vienna, intensive system tests (benchmarks) took place and the data was analysed and evaluated in cooperation with ETH Zurich. The components of the test included, among others, automatic DTM generation, and orthoimage generation and mosaicking. Thereby, products of the companies LHS (DPW 770), Zeiss (PHODIS), Autometric (Softplotter) and INPHO (Match-T for DTM ; orthoimage generation is not supported by INPHO products) were tested. For DTM generation with Softplotter two software packages (old and new Beta version) were tested.

### 2. BENCHMARK TESTS

#### 2.1 Test and Reference Data

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 did not cover all possible land cover types and terrain roughness. 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, another dataset (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 the scanned images, 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.

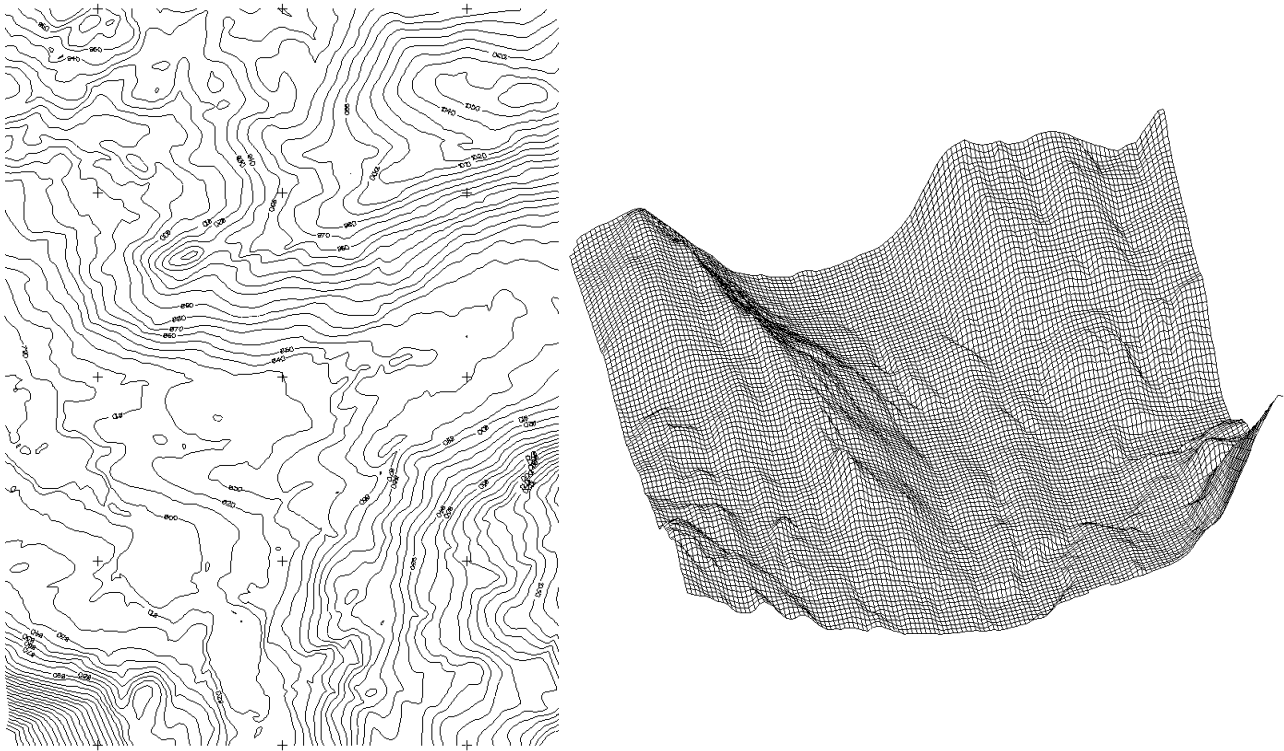


Fig. 1. Reference dataset (mass points). Left: contours with 10 m interval. Right: 3-D wireframe model.

For DTM generation, one B/W model over hilly terrain (3 km x 4 km, 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 (30 m regular grid and additional points = 16'400 mass points, see Fig. 1), 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. To check the planimetric accuracy of the orthoimages, 11 well defined, evenly distributed GCPs measured with GPS and an estimated accuracy of ca. 10 cm were used.

All used images were scanned at a photogrammetric scanner with high geometric accuracy and a pixel size of 15  $\mu\text{m}$  and had a mean scale of ca. 1:24'000. The camera was a Leica RC30 with a 152 mm lens, and was connected to a GPS for measurement of the camera station positions.

## 2.2 Benchmark 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. 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 ergonomics and ease-of-use.

It must be noted that in DTM generation we did not specify how the matching parameters should be selected, although the quality of the results greatly depended on this selection. 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.

The data analysis and quantitative evaluation took place at ETH Zurich and required ca. 2.5 man-months. All the results were evaluated per manufacturer according to a common key and a detailed report was prepared. In the sequel, the LHS DPW770 is called system A, the Autometric Softplotter system B (whereby for DTM generation B1 and B2 denote the old and new software version respectively), the Zeiss PHODIS system C and INPHO Match-T system D (with D1 and D2 denoting two DTM generation versions with 15 and 30  $\mu\text{m}$  scan pixel size respectively). Since Match-T was expected to be among the best systems based on previous tests, 15 and 30  $\mu\text{m}$  images were used to check the effect of pixel size on DTM accuracy. PHODIS and Match-T use identical software but the user interface and the matching parameter settings differed. The DTM and orthoimage modules of Softplotter are also used in the ERDAS Orthomax software. Since Intergraph digital photogrammetric stations also use Match-T, this test included all major photogrammetric systems for DTM generation with the exception of VirtuoZo, on which it has been reported elsewhere (see Baltasvias et al., 1996).



Fig. 2. From left to right: orthoimages generated with systems A (note coarse structure of edges at centre and top right), B (note saturation of bright areas) and C.

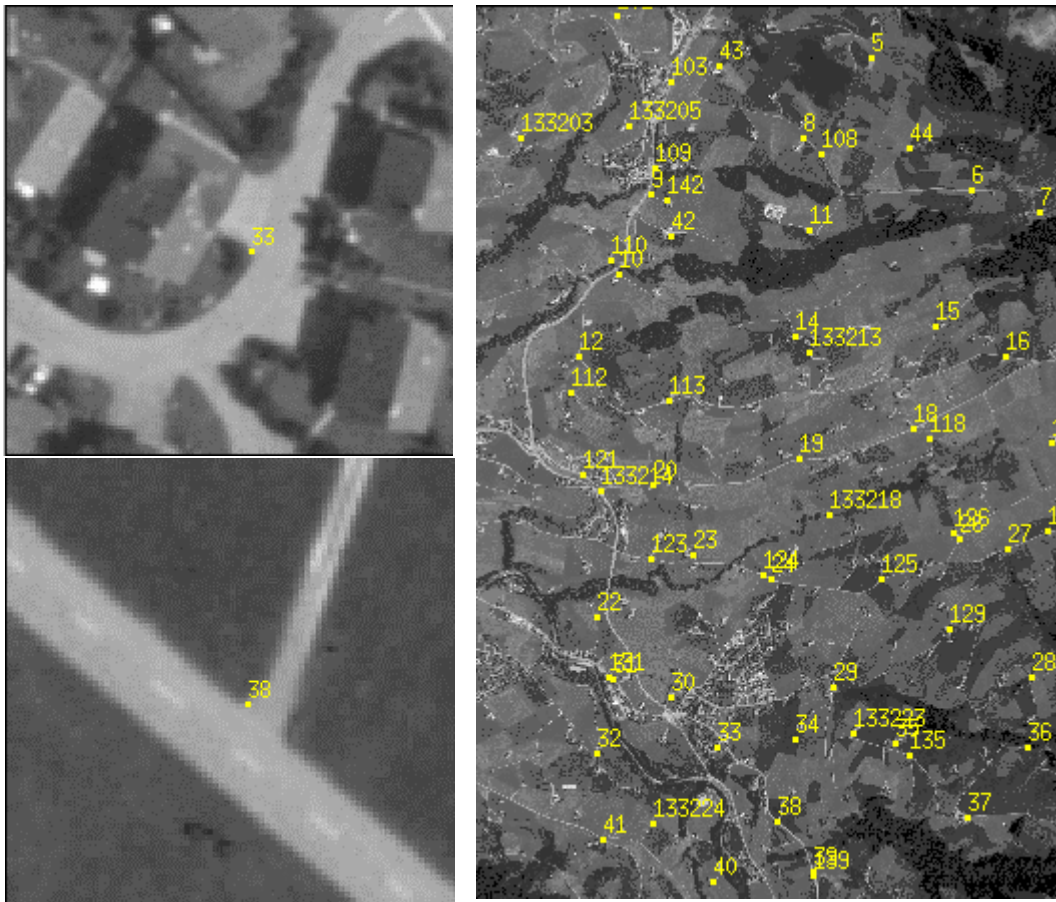


Fig. 3. Left: example of two measured orthoimage tie points. Right: distribution of measurement points in the overlap region.

### 3. TEST CRITERIA, PROCEDURES AND RESULTS

#### 3.1 Orthoimage and Mosaic

The 0.5 m pixel size orthoimages (from left and right image for each of the three systems) 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 four GPS points with differences less than 2 m were kept. The radiometric quality of the orthoimages

was visually controlled (see Fig. 2). 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 (see Fig. 3) were selected in one orthoimage and were transferred by semi-automatic Least Squares Matching (LSM) in the second one. Similar or identical points were used for all three systems. 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 local relative errors between the two orthoimages, while the offset error showed a systematic global shift between the two orthoimages (see Table 1). The local relative error was 0.4 – 1.1 pixel, while the offset error was 0 – 1 pixel. The differences between the systems were not very big, with the exception of the X-offset. The errors in the X (base) direction were larger than in Y.

For the absolute accuracy check, the four GPS and 41 additional control points were used. These 41 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. These coordinates were treated as known reference values. Then, statistical measures of the differences between known and measured coordinates were estimated (see Table 2). Again, the differences between the systems were not big (only the RMS and mean values in X for the left orthoimage of system B were larger), with X- and Y-RMS in the range of 0.6 – 1.5 pixel. The errors were larger in the base direction. The Z-differences represent

the local accuracy of the DHM25 and are very similar for all orthoimages, since the planimetric coordinates were almost identical. In general and taking Tables 1 and 2 into account, the results of system C are slightly better than those of system A and latter slightly better than those of system B. The differences between the systems can only be due to different interior orientation and internal computations for the orthoimage generation. Although small, they are larger than expected.

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 orthoimage 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. Bugs existed even in this relative simple and well established procedure. With one system which was using a new software release, the result, instead of a colour orthoimage, was nine repeated small B/W orthoimages each with a light border and decreasing brightness from left to right.

Table 1. Statistics of relative planimetric accuracy, i.e. differences between left and right orthoimage (in pixel)

System	No. of points	X - Standard Deviation	Y - Standard Deviation	X - Mean (Offset)	Y - Mean (Offset)
A	54	0.94	0.45	-0.53	0.07
B	53	1.09	0.50	-1.03	-0.05
C	41	0.56	0.39	-0.26	-0.40

Table 2. Statistics of absolute planimetric and height accuracy of orthoimages (in pixel for X and Y, in m for Z)

System	Ortho-image	X			Y			Z		
		Mean	Max. Abs.	RMS	Mean	Max. Abs.	RMS	Mean	Max. Abs.	RMS
A	Left	-0.07	4.09	1.05	-0.21	2.25	0.86	0.27	1.67	0.78
	Right	0.07	3.73	0.88	-0.04	1.91	0.70	0.27	1.72	0.80
B	Left	-1.22	2.34	1.52	-0.04	1.72	0.63	0.24	1.75	0.78
	Right	-0.39	3.00	0.94	-0.05	2.25	0.70	0.26	1.73	0.80
C	Left	0.11	3.94	0.97	-0.11	1.71	0.58	0.28	1.67	0.78
	Right	0.36	3.91	0.97	-0.52	2.38	0.89	0.27	1.67	0.78

### 3.2 Digital Terrain Model

System A uses normalised crosscorrelation with epipolar images and an image pyramid. It includes many predefined strategy files, out of which the steep\_plus strategy with 8 pyramid levels and 15 x 15 pixel patch size was used. Steep\_plus is intended for steep terrain, tries to remove

buildings and trees and performs a smoothing at “non critical” DTM points. Other strategies, e.g. IOR strategies, deliver according to our experience slightly better results, but at high computational costs. System B1 is using a similar hierarchical crosscorrelation matching method, but the correlation patches are orthorectified on-the-fly, i.e. the method can work with both raw and epipolar images. In comparison to system B1, system B2 uses epipolar images, is faster and has better image memory

management. The crucial difference, however, is that it generates very dense measurements (in this test 6.5 times denser than the DTM nodes), from which at the end a regular grid is interpolated based on a weighted average interpolation with variable radius and selection of one out of 6 predefined interpolation formulas. This is a similar thick-to-thin philosophy as used in Match-T, but in latter a robust filtering instead of a weighted averaging is used and thus blunders can be better eliminated, given a sufficient number of measurements for each grid node. Both systems B1 and B2 used 6 pyramid levels and an adaptive patch size of 7 x 7 or 9 x 9 pixels. Systems A and B use different predefined matching strategies with many parameters, which are often not explained or at least not well enough. They also use many criteria for detecting poor correlations and interpolate them, and filter out spike errors. Systems C and D use the same algorithm, but the matching parameters were set differently. System C selected points with an interest operator in every fourth epipolar line, used 6 pyramid levels and parallax bound (8 pixels) and terrain smoothing (medium) for rolling terrain. System D selected points in every epipolar line, used 10 pyramid levels, and set parallax bound (15 pixels) and terrain smoothing (low) for mountainous terrain. All systems produced a 10 m regular grid (ca. 120,000 points). System C, for unexplained reasons, produced a grid in an area almost twice larger than the one requested. At the borders of this area, large errors occurred (see contours in Fig. 6b). However, since the reference data was not covering these border regions, the errors occurring there did not influence the statistical results, with the exception of 6 points (see Fig. 7d) with errors between 55 and 65 m. All remaining points had errors less than 10 m. System B1 had huge errors at the borders (see Fig. 4). Thus, the border regions were excluded from any further analysis.

The two sets of reference values (mass points, breaklines) were bilinearly interpolated in the automatically generated DTMs. Statistical values of the differences were computed, as well as error histograms using predefined classes (see Tables 3 and 4). For the mass points the RMS was 0.6 – 1.6 m, the ABS (average with sign) 0.1 to 0.8 m, and the maximum absolute error 6 – 65 m. For the breaklines the respective values were: RMS 1 – 2.4 m, ABS (average with sign) 0.3 – 1.7 m, and maximum absolute error 3.7 – 8.8 m. Compared to the statistics for the mass points these values were 1.4 – 2 times worse in RMS, 1.6-5.2 times worse in the mean, but the maximum absolute errors were smaller, clearly indicating that there are other more serious sources of blunders. Methods with dense measurements (systems B2, D1, D2) lead to lower mean and RMS errors with breaklines. The mean was usually negative, especially when strong smoothing was used (systems A and C), indicating that matching measures higher than the terrain. The maximum absolute errors were 9.4-40.7 RMS for the mass points and 3.0-5.2 RMS for the breaklines, i.e. many errors exceeded the 3 RMS limit.

System B2 as compared to B1 gave better results with the exception of the maximum absolute error for the mass points. The reason is probably the fact that the first method measures 6.5 more points but since their reduction to grid nodes is based on weighted averaging existing blunders in the raw measurements remain in the grid nodes.

The results with the 30  $\mu\text{m}$  images, as compared to the 15  $\mu\text{m}$  ones, were 16% / 20% worse in RMS for the mass points / breaklines, very similar in the average with sign, and very

similar or even better in the maximum absolute error, since large errors often drift away from the correct solution as the pyramid level is decreased. The main differences between the above two versions were, as expected, at "terrain" discontinuities like forests, trees and buildings due to the different image resolution. This small difference between 30 and 15  $\mu\text{m}$  is partly explained by the reduction of many measurements to one grid node, which leads to a smoothing of small local errors due to the resolution difference.

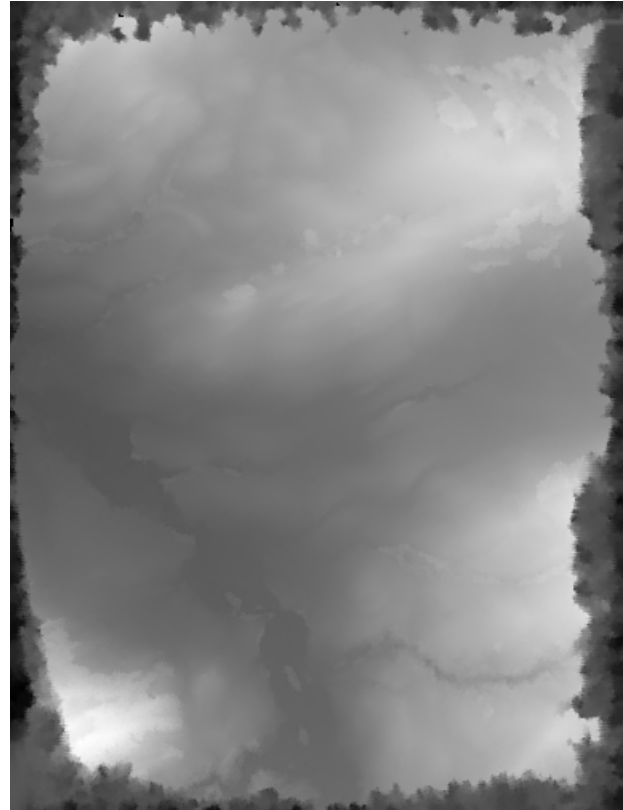


Fig. 4. DTM of system B1 displayed as grey level image. Note the huge errors at the borders. Big errors appear as bright or dark spots (e.g. black spots at top right of the image).

Systems C and D1 with identical DTM modules differed a lot (for mass points / breaklines by a factor 2.8 / 2.3 in RMS and 4.8 / 5.4 in mean) proving how sensitive the results are to the selection of the match parameters, and the difficulty of appropriately setting them even for expected experts. A phenomenon slightly observed with system C, and clearly visible with systems C, D and Intergraph in previous tests, is the formation of tiles in the contours, which may be related to errors in fitting together the overlapping DTM patches that these systems generate. System C also had a bug, which has been corrected in between, i.e. when starting the program from the menu interface, matching parameters manually edited in the project file were changed. In addition, some default parameter values, e.g. selection of 5 pyramid levels independently of the terrain steepness, were unreasonable (also corrected lately).

Table 4 with the error class frequencies is compatible with the above statements. Systems C and especially A perform the worst (frequencies are not monotonically falling with increasing errors, small error classes have low frequencies, system A has too many errors exceeding 3 RMS = 2.3 m), while systems D1,

D2 and B2 perform the best. For these three systems and using the assumption that the expectable error per measurement is 1 pixel (corresponding to 0.76 and 1.52 m height error for 15 and 30  $\mu\text{m}$  images respectively) and thus 3 RMS is 2.3 and 4.6 m for the two image resolutions, then the percentage of mass point errors exceeding 3 RMS is 0.41%, 0.02% and 1.38% for the systems D1, D2 and B2 respectively.

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 (see some examples in Fig. 7). 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. Additional methods to visualise the errors included: generation of contours with 10 m interval (detection of gross errors, quality of geomorphologic details and breaklines, noisiness, see Fig. 5 and 6), their comparison to the map contours and their overlay on orthoimages, 3-D wireframe models from different viewpoints, representation of the automatically generated DTMs as grey level images (detection of gross errors and quality of geomorphologic details) and comparison to an equivalent representation of DHM25, generation of absolute error contours with an interval of 2 m and their overlay on the orthoimage and map. Shaded relief representations show very well DTM errors but due to time limitations were not produced. The best overall methods for visual evaluation are: overlay of contours on orthoimages and reference contours, overlay of error contours and position of largest errors on orthoimages, shaded relief representations. Regarding the contours (see Fig. 5 and 6), the following can be noted. As compared to system B1, the contours of system B2 are less noisy, include more details (trees, buildings etc., also compared to all other systems) due to the higher measurement density, have less errors (see contours due to blunders at top right of Fig. 5b) and give a slightly better representation of terrain details. The contours of system C are too smooth and many details are lost. The contours of system A are also smooth and the performance at terrain discontinuities (see the two horizontal creeks at the lower right part) is the poorest of all systems, due to the large correlation patch size. The contours of systems D1 and D2 are similar to those of system B2.

All systems provide for some accuracy indicators. Unfortunately, the global accuracy indicators are too optimistic and the quality indicators for each individual grid node are not reliable. System A provides for each point a Figure Of Merit (FOM). Points with  $\text{FOM} < 33$  are considered to be poor. Although rejection of all such points leads to a considerable accuracy increase, errors, including large ones, still remain in the dataset (see an example in Baltsavias et al., 1996). Using a \*plus strategy, causes almost all points to have a FOM of 25 (i.e. smoothed), so even the above mentioned deletion of bad points becomes impossible. Systems B provide some global criteria like minimum, maximum, mean and sigma of precision, which however are of little help, when it comes to DTM editing. The points are divided in: good, fair, poor, other (interpolated) and off-image. These quality criteria are colour coded and can be used by the user when editing the measurements. However, as with all systems, these criteria are not reliable, so one has to practically check the whole dataset. Systems C and D provide a theoretical accuracy which mainly depends on the redundancy (i.e. number of measurements for each DTM node) and is very optimistic (e.g. for systems D1 and D2 it was 0.048 ‰ and 0.077 ‰ of the flying height over ground respectively). A better quality indicator, which again does not allow safe detection of errors, is the number of raw measurements per DTM node.

Methods of systems B2 and D that used very dense measurements and then a thin-out with parallel blunder detection performed better. Cheaper modules (B2), encountered also in remote sensing packages, performed better than some more expensive ones (system A). Feature-based matching (systems C and D) generally perform better at discontinuities than area-based ones due to their smaller areal support, but require many measurements per grid node to filter out blunders and fail when blunders are concentrated as with little, repeatable texture like dot patterns in agricultural fields. It should be noted that the good results of the best systems with respect to RMS and mean errors, are partly due to the fact that all reference measurements were not on or close to nonterrain objects, areas that very often exhibit large matching errors. Operational considerations lead to the conclusion that it is faster to pre-exclude problematic areas like dense trees and forest, dense built-up areas and water bodies, and let matching work primarily on bare terrain where it can reach or even exceed the accuracy achieved with analytical plotters.

Table 3. Statistics of differences for mass points / breaklines between reference data and automatically generated DTMs (in m)

	System A	System B1	System B2	System C	System D1	System D2
<b>No. of points</b>	16236 / 1102	15659 / 1065	15085 / 1037	16357 / 1108	16357 / 1108	16357 / 1108
<b>Mean</b>	-0.73 / -1.68	-0.17 / -0.88	0.1 / -0.46	-0.82 / -1.52	0.17 / -0.28	0.13 / -0.37
<b>Max. Abs.</b>	18.46 / 8.18	11.55 / 8.81	18.11 / 6.24	65.17 / 6.52	8.03 / 3.73	6.30 / 4.32
<b>RMS</b>	1.51 / 2.42	0.87 / 1.72	0.78 / 1.21	1.60 / 2.16	0.58 / 0.95	0.67 / 1.14

A measurement error of one pixel corresponds to a height error of 0.76 m and 1.52 m or 0.22 ‰ and 0.43 ‰ of the flying height over ground for 15 and 30  $\mu\text{m}$  images respectively.

Table 4.: Classes of absolute differences for mass points/breaklines between reference data and automatically generated DTMs (in %)

Classes	System A	System B1	System B2	System C	System D1	System D2
<b>Diff. &gt; 10.0 m</b>	0.05 / 0.54	0.01 / 2.25	0.04 / 4.24	0.04 / 0.00	0.00 / 0.00	0.00 / 0.00
<b>6.0 m &lt; Diff. &lt; 10.0 m</b>	0.15 / 0.82	0.10 / 0.56	0.07 / 0.10	0.04 / 0.36	0.01 / 0.00	0.01 / 0.00
<b>4.0 m &lt; Diff. &lt; 6.0 m</b>	0.92 / 7.80	0.36 / 2.26	0.20 / 0.19	0.42 / 5.32	0.01 / 0.00	0.02 / 0.00
<b>3.0 m &lt; Diff. &lt; 4.0 m</b>	3.14 / 15.06	0.80 / 4.70	0.27 / 1.26	1.06 / 11.19	0.07 / 0.27	0.19 / 0.99
<b>2.2 m &lt; Diff. &lt; 3.0 m</b>	9.14 / 15.88	1.59 / 9.02	0.89 / 5.50	2.29 / 14.98	0.35 / 1.71	0.67 / 3.16
<b>1.7 m &lt; Diff. &lt; 2.2 m</b>	10.81 / 12.07	2.33 / 10.98	1.40 / 7.14	5.05 / 13.54	0.78 / 5.51	1.55 / 7.22
<b>1.2 m &lt; Diff. &lt; 1.7 m</b>	14.91 / 12.25	5.46 / 14.09	4.03 / 12.34	16.02 / 15.44	2.78 / 6.77	4.24 / 9.93
<b>0.9 m &lt; Diff. &lt; 1.2 m</b>	12.43 / 9.16	7.15 / 9.67	6.71 / 13.02	18.97 / 8.12	5.60 / 13.90	6.99 / 16.70
<b>0.6 m &lt; Diff. &lt; 0.9 m</b>	14.68 / 8.80	15.25 / 13.90	14.56 / 14.66	22.79 / 9.75	14.55 / 19.77	16.15 / 17.24
<b>0.3 m &lt; Diff. &lt; 0.6 m</b>	16.21 / 9.16	27.88 / 15.31	28.40 / 18.13	18.71 / 11.28	30.90 / 23.91	29.56 / 21.30
<b>0.0 m &lt; Diff. &lt; 0.3 m</b>	17.55 / 8.44	39.07 / 17.28	43.43 / 23.43	14.61 / 10.01	44.95 / 28.16	40.61 / 23.46

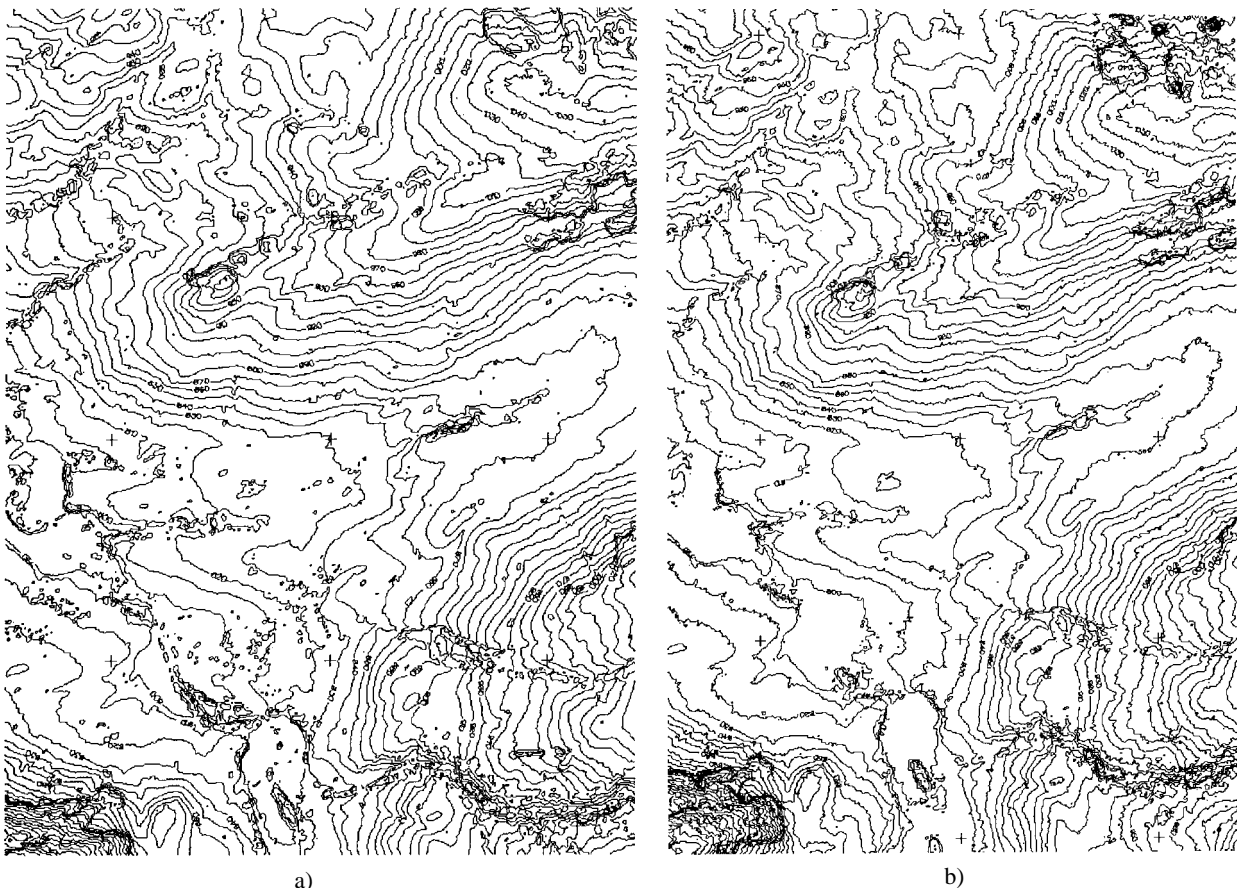


Fig. 5. Contours with 10 m interval: left system B2, right system B1.

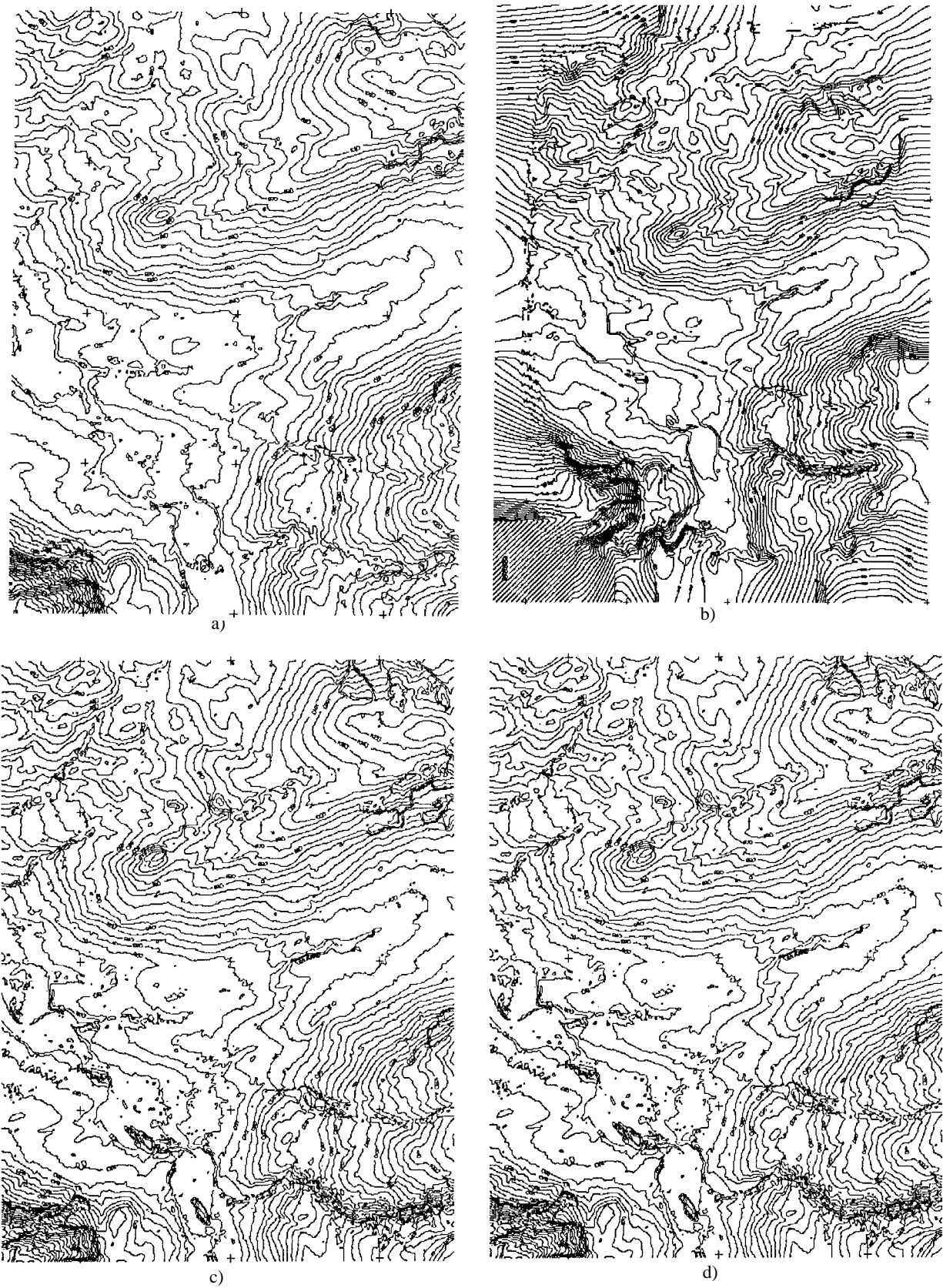


Fig. 6. Contours with 10 m interval. From top left clockwise: systems A, C, D1, D2

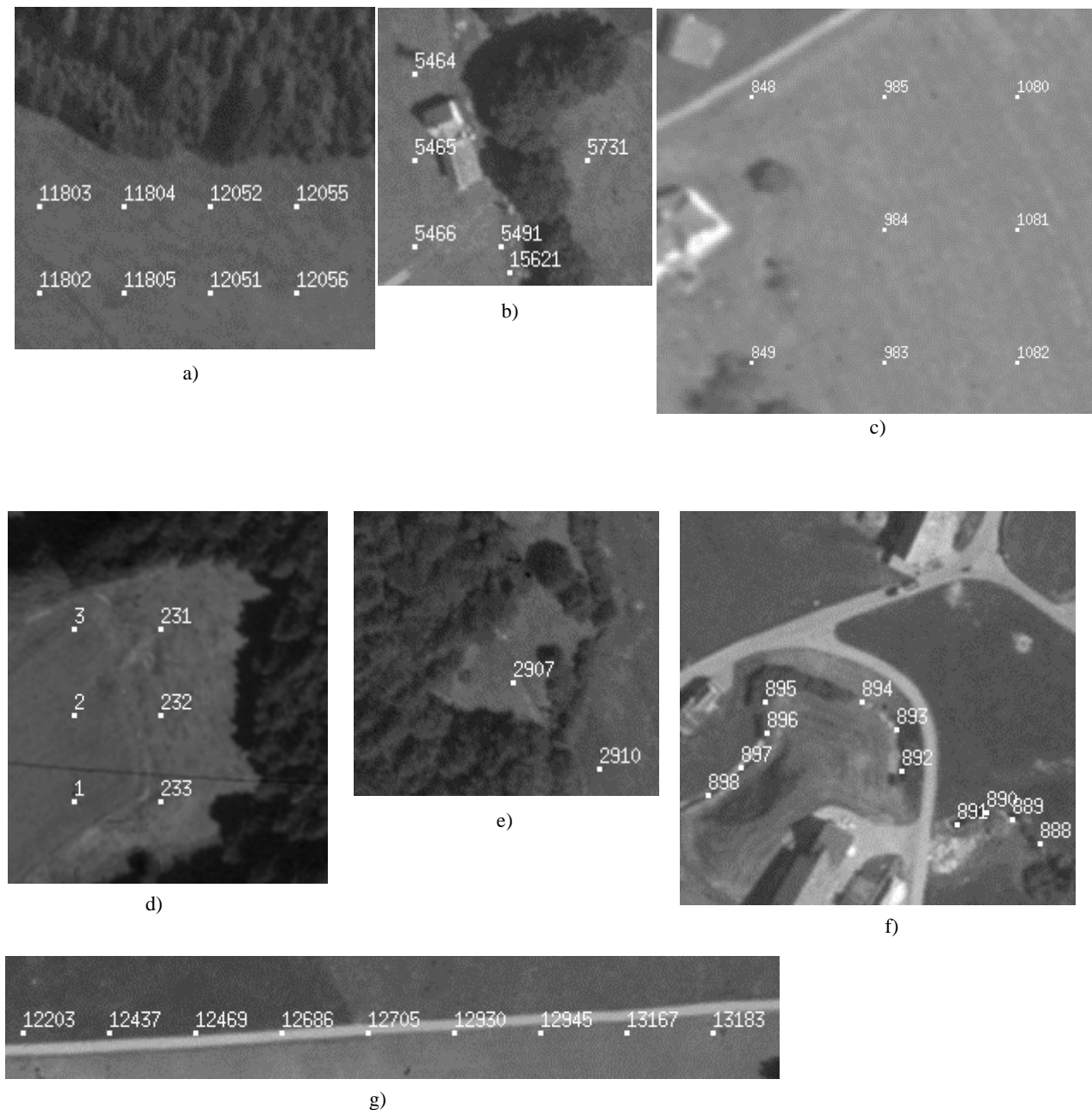


Fig. 7. Different large errors overlaid on an orthoimage: a) little texture, points close to forest (system B1, error at point 12056 = -11.55 m), b) points close to 3-D objects like buildings and trees – occlusions, shadows (system B1, error at point 5731 = -9.98 m), c) textureless areas (system D2, error at point 984 = -8.03 m), d) very large errors at the bottom left DTM region (system C, errors vary from 55 to 65 m), e) point in a forest opening is wrong due to robust local height filtering; instead of eliminating single buildings or trees, and since this ground point is a minority in its neighbourhood, the filtering sets its height similar to the surrounding tree height (system C, error of point 2907 = -7.20 m), f) points along breaklines – surface discontinuities (systems B1, C, error of point 893 = -6.52 m), g) edges parallel to epipolar lines – multiple solutions (system B2, error of point 13167 = -18.11 m).

### 3.3 Evaluation Scores

Table 5 shows the evaluation scores of the systems for the benchmark using all criteria and different evaluation criteria weighting versions (for details see Käser et al., 1998).

Regarding DTM generation, other important evaluation criteria, apart from accuracy, included execution time (varied a lot) and tools to visualise and edit the results (sufficiently good tools only with system A). In the final evaluation, additional criteria, like costs, maintenance etc., were also used.

Table 5. Benchmark test: evaluation scores for various weighting versions (for explanations see Käser et al., 1998)

Version System	Uniform (69/36) <sup>1</sup>				Optimal (90/45) <sup>1</sup>				Minimal (90/30) <sup>1</sup>			
	A	B	C	D	A	B	C	D	A	B	C	D
Orthoimage & Mosaic	43	36	40	-	50	39	57	-	44	36	51	-
DTM	27	18	20	22	29	27	18	32	21	22	16	24

<sup>1</sup> Maximum number of score points for orthoimage & mosaic / DTM.

#### 4. DISCUSSION AND CONCLUSIONS

Good knowledge of the theories involved in digital photogrammetric processes and the algorithms used in the different system modules 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. The analysis of the results was made more difficult due to the poor quality of protocol files and output listings (often long outputs, numbers listed with no or unclear indication of what they represented). In all cases, algorithmic details are known only by 1-2 development experts who usually work at the headquarters. 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. Sometimes, very new versions of software modules were used without sufficient previous testing. 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.

Regarding DTM generation, the main limitation of the systems is their failure to reliably indicate blunders. Other important points include: no good performance at and explicit modelling of breaklines ; no simultaneous use of multiple images and associated geometric constraints to reduce problems due to occlusions, multiple solutions etc. ; poor filtering of buildings and trees (a successful identification and elimination of such objects requires additional cues like colour and texture information), while sometimes terrain features are filtered out instead ; pre- and post-editing tools that are complete, fast and easy-to-use are rare ; image preprocessing, like radiometric equalisation with parallel contrast enhancement could be very beneficial for matching but systems do not offer such functionality ; many matching parameters, not clear, and not

easy to set (instead, the user should set only very few parameters and the programs should compute internally the optimal match parameters for the given scene and image content, even adapting these parameters to variations within the images). It should be also noted that the problems encountered in this test become worse with larger image scales and/or steeper terrain.

Regarding orthoimage generation, the main problem concerns the radiometric equalisation of orthoimages during mosaicking. Especially, an equalisation of colour images without spectral shifts, convenient (and if possibly automatic) determination of the seam lines and treatment of problems like hot spots remain an open problem. This does not mean, that other simpler processes, as it was shown in these tests, always perform well.

None of the systems fulfilled to a large extent the specified needs. Although one and half years have passed since the tests only minor improvements have been performed in between.

#### ACKNOWLEDGEMENTS

The authors would like to thank the 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., Li, H., Stefanidis, A., Sinning, M., Mason, S., 1996. Comparison of Two Digital Photogrammetric Systems with Emphasis on DTM Generation: Case Study Glacier Measurement. In: Int'l Archives of Photogrammetry and Rem. Sensing, Vol. 31, Part B4, pp. 104-110, Vienna, Austria.
- Käser, Chr., Eidenbenz, Chr., Baltsavias, E.P., 1998. Evaluation and Testing of several Digital Photogrammetric Systems for System Acquisition by a National Mapping Agency. In: Int'l Archives of Photogrammetry and Rem. Sensing, Vol. 32, Part 2, Cambridge, UK.