Pose Estimation from Airborne Video Sequences Using a Structural Approach for the Construction of Homographies and Fundamental Matrices

Eckart Michaelsen[1] and Uwe Stilla[2]

[1] FGAN / FOM, Gutleuthausstr. 1, 76275 Ettlingen, Germany
mich@fom.fgan.de
http://www.fom.fgan.de
[2] Photogrm. & Rem. Sens. / TU Munich, Arcisstr. 21, 80333 München, Germany
stilla@bv.tum.de
http://www.bv.tum.de

Abstract. A structural knowledge-based search method is utilized for the estimation of geometric transforms from airborne video sequences. Examples are projective planar homographies and constraints such as the fundamental matrix. These estimations are calculated from correspondences of interest points between two images. Different approaches are discussed to cope with the problem of outlier-correspondences. To ensure any-time performance the search process is implemented in a data-driven production system. The pose estimation from planar homographies is compared to estimations from fundamental matrices. A fusion of both approaches is proposed. The image processing is performed by bottom-up structural analysis using an assessment-driven control. Examples are from the thermal spectral domain.

1 Introduction

Pose trajectory estimation from moving cameras is an important task for scene reconstruction as well as navigation. Research in this field was stimulated by development of mobile autonomous robots. Particularly, methods using projective geometry were utilized [3][6][9]. Recently, unmanned aircraft equipped with video cameras are gaining increased attention for civil as well as military applications like traffic monitoring [16] or surveillance tasks. The appearance of a scene viewed from an aircraft depends on the flight altitude and the height of the sensed objects. If this ratio is large, the scene will appear flat. This implies a different approach than a spatial scene.

Flat scenes are treated by planar homographies. These may be estimated by e.g. minimizing the sum of absolute errors [1]. Given a Gaussian distribution on the displacements of the corresponding image positions it can be shown that the minimization of the sum of the squared errors is the optimal solution [9]. Actually, the direct linear transform (DLT) methods proposed today minimize an "algebraic" squared error sum that is not identical with the squared displacement error in the 2-d image coordinates. However, it has been shown that this error minimization approximates the Gaussian minimization very closely provided that the coordinates are normalized in a proper way [6]. The main disadvantage of minimization of squared error sums is

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 486–494, 2004.

the sensitivity to the inclusion of outliers into the calculation. An outlier is a correspondence that has been erroneously constructed. It does not follow the distribution assumptions underlying the estimation. Because of its possibly large displacement and particularly because of squaring, it may have a large weight in the computation where it should be neglected. Outliers cannot be avoided if automatic estimation is the task. Therefore so-called "robust" methods are proposed.

Section 2 presents and compares three robust estimation methods to solve the problem of planar homography estimation with DLT squared error sum minimization. The term "outlier" and its meaning in the context of homography estimation from airborne videos is further investigated in Section 3. Section 4 compares the pose estimation from planar homographies to estimations from fundamental matrices. A fusion of both approaches is proposed in Section 5. The image processing is performed by data-driven structural analysis and an assessment-driven control. All example data are taken from the thermal spectral domain to ensure independence of the daylight.

2 Robust Estimation of Planar Homographies

Robust estimation methods may be classified into approaches that assume the existence of mutually exclusive sets of inliers and outliers (Section 2.2) and others that assign weights to the correspondences (Section 2.1).

2.1 Iterative Re-weighting Least Squares (IRLS)

An example of the assigning of weights to the correspondences is iterative re-weighting least squares [7]. The inverse of the residual of the least squares solution of each correspondence of the complete sample is used to re-weight its influence. Correspondences yielding a large residual error will be punished and correspondences yielding a small error will gain more influence. If a large portion of the correspondences is expected to be wrong, a local minima problem may occur. The convergence of IRLS to the desired minimum is theoretically not guaranteed. It may end up with zero-error and thus infinite weight on an arbitrary minimal sample and random small weights on all other members. However, in our examples we found that it does converge slowly but robustly to a good solution. IRLS-estimation of 2D-homographies is available in public code libraries [17]. Proposals are made for the handling of occlusion outliers and lighting changes within the IRLS-method [8].

2.2 Random Sample Consensus

The standard method for inliers-outliers discrimination is the random sample consensus approach (RANSAC) [4]. The calculation is performed on minimal samples which are randomly picked from the complete sample of correspondences. The result of the calculation is tested on all the other correspondences giving a residual error. If this error is smaller than a threshold t_s, the correspondence will be termed to be in consensus with the actual sample. After repeating this procedure a sufficient number

of times the search is terminated. The termination criterion bases on a minimum size of the current best consensus m and a maximal number of cycles n_c. There is an elaborate theory for the choice of these parameters (t_s, m, n_c) from the expected portion of outliers, a standard deviation of the error of the position of inliers, and a significance level [6]. The sample with the highest consensus is chosen and the corresponding consensus set is used to determine the estimation by mean squared error minimization.

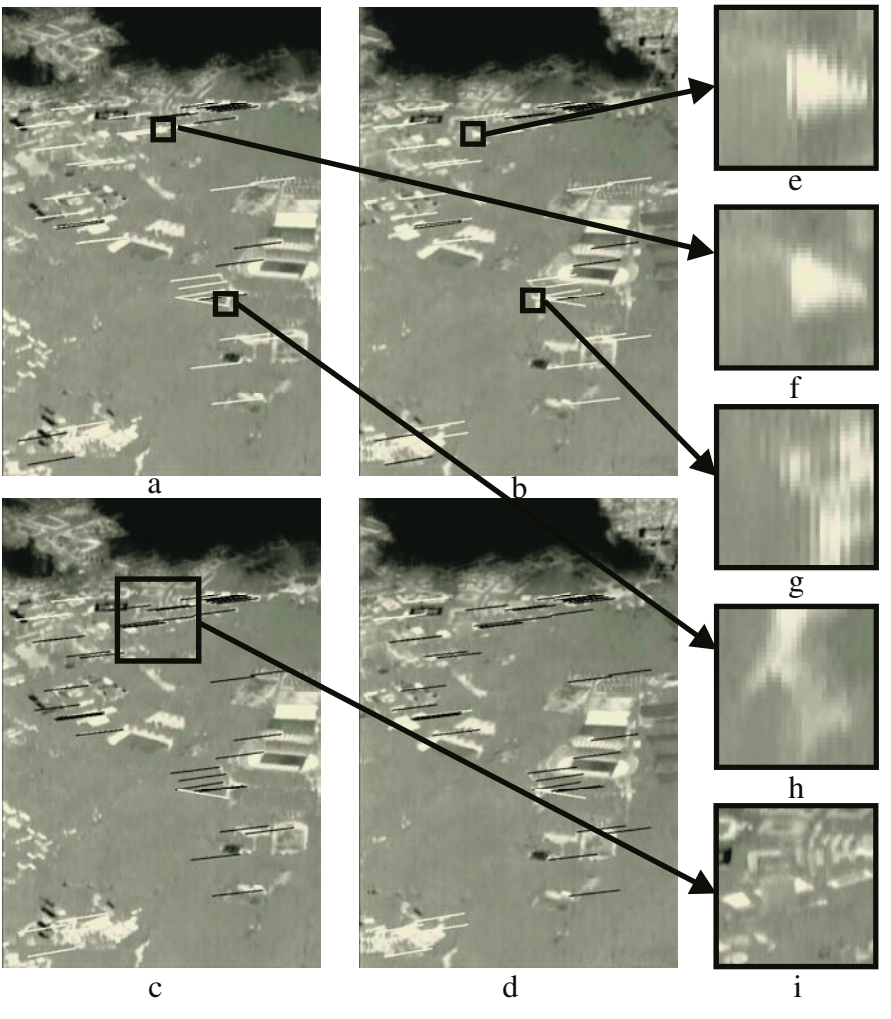

Fig. 1. Image pair of a thermal video sequence and corresponding points. a,b) Best GSAC-sample in black, other correspondences in white; c,d) RANSAC-sample with innlier (black) and outlier (white); g,h) incorrect correspondence excluded by both methods; e,f) incorrect correspondence found as RANSAC-inlier; i) greater section around that location showing that it results from partial occlusion.

2.3 Good Sample Consensus

Some authors proposed to modify random sampling by taking also the quality of the samples into account [11], [13]. This obvious idea is not new and has already been touched in the original paper of RANSAC [4]. For such improvements a criterion always has to be defined that assesses the suitability of a sample for the intended calculation. For example the position of the corresponding points within the images will be important for estimating homographies. They should cover as much area of the images as possible. Moreover, more than two collinear points should be avoided. Fig. 1 shows images with large homogenous regions and the structure concentrated in few regions. Because RANSAC only counts the number of mutually consistent correspondences, it may concentrate too much on densely structured regions and tend to under-estimate the importance of good but rather isolated correspondences elsewhere, e.g. the correspondences in the lower left corner that are missing in the inliers set of RANSAC.

In GSAC the assessment criterion is used to control the search for a good sample on which the solution is based. The correspondences in the lower left corner are now included. Section 5 explains how GSAC can be implemented by a structural method.

3 Classification of Correspondences

1) Correspondences Consistent with the Homography Model: A correspondence between structures in two different images will be called *correct* if the location on the object in the scene that caused them is the same. Additionally they must fulfill the model constraint. For homography estimation only those objects that are located on the assumed plane can cause correct correspondences. In urban terrain this plane will be at the average height of the buildings. Of course, we will have to tolerate small deviations from this constraint. The residual error will mainly result from the localization error of the 2d-structures and may be modeled as normally distributed.

2) Correspondences Consistent with a More General Geometric Model: In an urban area there may be tall buildings that are jutting out of the plane. Corresponding structures resulting from the roofs of such tall buildings will violate the homography constraint. Still, they may be correct in the sense that they come from the same physical property. Their deviation from the homography follows a different rationale: They will be located close to the epipolar line which goes through the point determined by the homography and through the epipole. They should be excluded from the estimation of the scene plane, but they may be included into the estimation of the camera rotation and epipole.

3) Correspondences from Moving Objects: Video sequences taken by a moving sensor yield image pairs that were obtained at different time instants. Moving objects in the scene may cause semantically correct correspondences that neither follow the planar homography nor the epipolar constraint. However, such correspondences from moving objects are required for applications like traffic monitoring. Fig 2 shows such a correspondence resulting from a moving vehicle. The correspondence is indicated as a white line, while other correspondences are drawn in black.

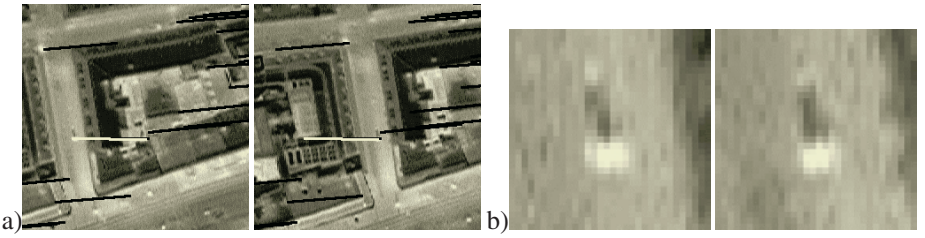

Fig. 2. Nadir looking sequence taken from urban terrain; example of a correct correspondence that is an outlier to the epipolar constraint estimation (a moving car on the ground); a) larger sections with surroundings, b) corresponding structures.

4) False Correspondences: Using an automatic method to construct the correspondences we cannot avoid the handling of semantically false correspondences. Most often these will result from occlusion phenomena. Fig. 1e and Fig. 1f show an example, where the outlines of a warm flat building roof (white) are partially occluded by a tall building in front of it (grey) – compare Fig. 1i. There is a structure correspondence located on the T-junctions caused by this occlusion. Such correspondences do not follow any predictable error function. They may accidentally be inside the error bounds of a homography estimation like the RANSAC-estimation displayed in Fig. 1a and Fig. 1b.

4 Robust Estimation of Epipolar Constraints

A central theorem of projective geometry states that from a pair of views of a scene the mutual orientation and translation of the cameras can be calculated from at least seven corresponding point-pairs *(x, x')* and that the position of the corresponding points in the 3d scene also follows from this reconstruction [6]. This is a constructive argument that is based on the inference of the fundamental matrix F from the correspondence data. This matrix formulates the epipolar constraint by stating $x^T F x' = 0$. It can also be estimated from this simple linear equation using at least eight correspondences. Such estimation is depicted in Fig. 3e and Fig. 3f. Problems with instability of the solution will occur, if all the correct correspondences are located in one plane. This happens in flat terrain. For testing this automatically, samples that are inconsistent with the homography model (see Section 3 class 2) are searched. Fig. 3e shows a sample of correspondences that gives the epipolar constraint depicted in Fig. 3f. But, this sample is smaller than the best GSAC-sample for the homography displayed in Fig. 3c. Moreover, in forward looking situations like the one presented in Fig. 3 the epipole (black/white cross) may be inside the frame, and scene reconstruction is impossible for the area around the epipole.

The effort for a random search for suitable minimal samples (containing seven or eight correspondences) is a rising polynomial with degree seven or eight with growing portion of outliers. Therefore, some authors introduced an intermediate part-of hierarchy into the samples [2]. Others propose a “plane plus parallax” approach [14]. First, one estimates a homography H that maps those points located on a dominant

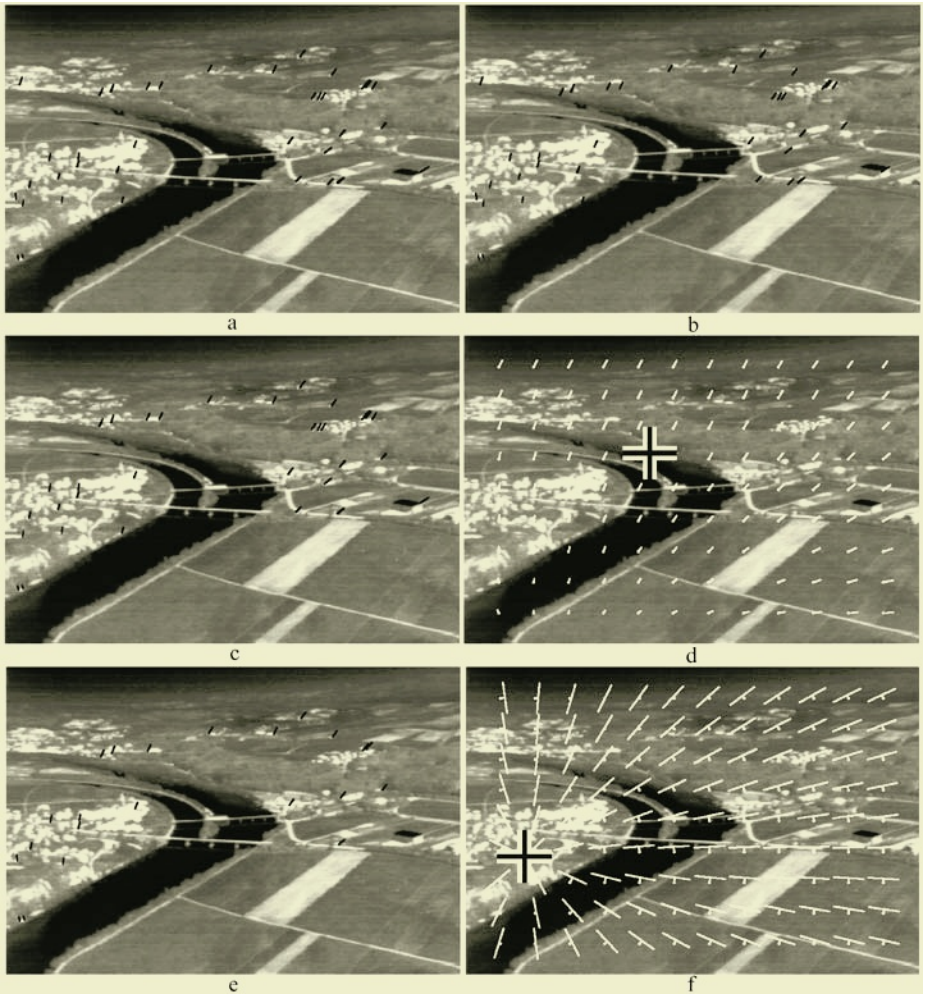

Fig. 3. Example of estimations from a forward looking oblique sequence; a) all correspondences on one frame b) same on other frame c) GSAC consensus correspondence set for homography estimation d) homography displayed as vectors on a grid array e) consensus correspondence set for epipolar estimation f) epipolar constraint; for each point on the grid the epipolar line is indicated and connected to the point.

plane from one image to the other $Hx=x'$ (Section 2). Then the homography can be decomposed in the form $H=R-nt^T$. R is the camera rotation matrix and the outer product nt^T results from the plane normal n and the camera translation t [3].

If we use normalized camera coordinates the 3-d vector t can also be interpreted as epipole. Multiplying the skew-matrix constructed from this vector with the rotation matrix R will give the fundamental matrix F belonging to the image pair. This matrix estimation for the epipolar constraint may then be refined using additional correspondences of the type mentioned in Sect. 3 class 2). In Fig. 3d the estimated homography is presented as a white vector field and the calculated epipole. In spite of the consid-

erably non-planar structure of the valley scene this turns out more stable than the direct epipolar constraint estimation. Particularly, the difference between the epipole estimations in Fig. 3d and Fig. 3f is considerable. A small rotation of the camera combined with such a displacement of the epipole give roughly the same point movements for forward looking geometries.

5 Production Nets for GSAC-Estimationof Geometric Entities

The pose estimation is partitioned into several steps and intermediate results. The overall structure of the process can be depicted by a so-called production net (Fig. 4). This bipartite graph contains productions and concepts (object types) as nodes. Arcs go from an object concept to a production whenever the objects are input to the production. Arcs go from a production to a concept whenever these concepts are constructed by the production. The productions contain constraints that incoming objects must fulfill to fit into the construction of the out-going objects of a higher concept. They also contain the functions that are necessary to construct these objects. The constructive part also contains an assessment part that evaluates the newly built object. Details of the control mechanism have been published in [15]. The assessment criteria used here are named in Fig. 4.

The processing starts with the application of an interest operator on the images that marks locations where neither homogeneity nor an aperture problem is likely [5]. Pixels trespassing a threshold form the primitive objects **P** of the structural analysis. Production p_1 groups such primitive objects into interest objects **I** using vicinity as its constraint, center of gravity as its function and total mass for its assessment. The position of such an interesting location **I** is determined with sub-pixel-precision. Given such an object production p_2 will search the other image for corresponding partners. It will construct new objects correspondence **C** for all such objects and assess them by means of correlation. Each such object may vote for a translation transform. Production p_3 forms objects **T** of a pair of correspondence objects **C**. Since they may be used to vote for similarity transforms, these objects are assessed according to the distance between the two locations in the image. Large distances give more precision for such estimation. Production p_4 gathers two such objects **T** and forms a quadruple object **Q** from them.

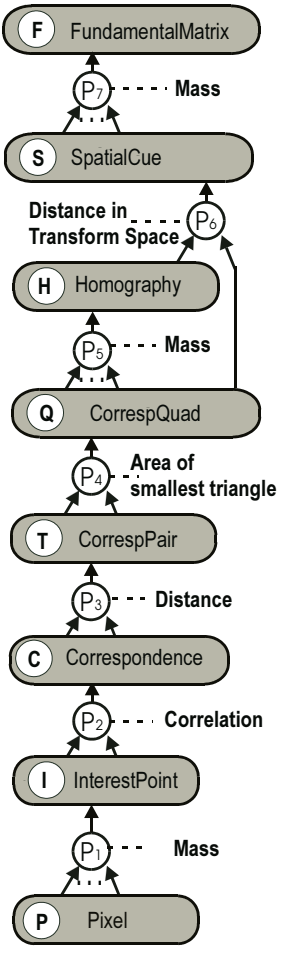

Fig. 3. Production-net with assessment criteria for bottom-up data-driven control.

From these a cue to the homography can be calculated. It is not only important that all four correspondences in such an object **Q** must be inliers of the type discussed in Section 3 class 1). Also no three of the four points are allowed to be collinear. They should cover as much area as possible. Therefore the area of the smallest of the four triangles in the quad is chosen as the assessment criterion. Production p_5 clusters the homography estimations from several consistent objects **Q**. The result is a new object **H** that is calculated via DLT squared error sum minimization from the sample of correspondences preceding the objects **Q** in the cluster. Thus Productions p_4 and p_5 implement the GSAC-rationale outlined in Section 2.3. An object **H** is assessed not only according to the number of correspondences in it, but also according to the assessment of the preceding objects **Q.** Production p_6 searches well directed for the outliers of the cluster process implemented by Production p_5. There may be correspondences in them that belong to the type described in Section 3 class 2). This results in spatial cue objects **S**. Such objects contain a fundamental matrix estimation. They are assessed according to the inconsistence of the homographies preceding them. Of course such cues need affirmation because it may result from correspondences of the types discussed in Section 3 classes 3) and 4). Production p_7 clusters consistent objects **S** into a well founded fundamental matrix estimation object **F** where the calculation is based again on DLT with the sample of the preceding correspondences.

The control scheme forms hypothesis of each newly constructed object and all the productions to which an arc goes from its type. These hypotheses get a priority according to the assessment of the object. All hypothesis compete for computational resources. In this manner homographies and fundamental matrices are already estimated from prominent and well positioned correspondences while other less important interest point objects still wait for an opportunity to search for correspondences in the other image. The process may be terminated at any time followed by choosing the best object **H** or **F** obtained up to this time instance according to the same assessment criteria.

6 Conclusion

Camera pose estimation using fundamental matrices as well as planar homographies can be obtained from the same images. The decision of which method should be preferred depends on the situation. Intermediate results give criteria for the choice. For a selected method the best sample of correspondences has to be searched. A structural knowledge-based approach combines both methods, uses well directed search and avoids early decisions. During the search run weights are assigned to entities like correspondences between structures in different images, pairs of such correspondences, quadruples and larger sub-sets. Each intermediate result is evaluated and the control of the whole system is based on these evaluations. Thus spurious calculations are avoided. Originally the production net approach has been invented for dealing with costly object recognition tasks in an accumulative way using affirmative intermediate results [10]. In pose estimation intermediate results may also be mutually competing. Thus, the assessments are a key issue in balancing such system. Tests comparing GSAC homography estimation performance to RANSAC and IRLS on a collection of example data are under way and will be published in [12].

References

1. Ben-Ezra, M., Peleg, S., Werman, M.: Real Time Motion Analysis with Linear Programming. CVIU, Vol.78, (1999) 32-52.
2. Chum, O., Matas, J., Obdrzalek, S. Epipolar Geometry from Three Correspondences. CVWW'03, Vatlice, Czech Republic, (2003).
3. Faugeras, O.: Three-Dimensional Computer Vision. MIT Press, Cambridge, Mass, (1993).
4. Fischler, M. A., Bolles, R. C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. Assoc. Comp. Mach., Vol. 24, (1981) 381-395.
5. Foerstner, W.: A Framework for Low Level Feature Extraction. In: Eklundh J.-O. (ed). Computer Vision – ECCV 94. Vol. II, B1, (1994) 383-394.
6. Hartley, R., Zisserman A.: Multiple View Geometry in Computer Vision. Proc. Cambridge University Press, Cambridge, (2000).
7. Holland, P. W., Welsch, R. E.: Robust regression using iteratively reweighted least-squares. Comm. Statist. Theor. Meth., Vol. 6 (1977) 813-827.
8. Jurie, F., Dhome, M.: Real Time Robust Template Matching. BMVC-2002 (2002) 123-132.
9. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.: An Invitation to 3-D Vision. Springer, Berlin, (2000).
10. Michaelsen E., Stilla U.: Probabilistic Decisions in Production Nets: An Example from Vehicle Recognition. In: Caelli T., Amin A., Duin R. P. W., Kamel M. Ridder D. de (eds): Structural, Syntactic and Statistical Pattern Recognition SSPR/SPR 2002, LNCS 2396, Springer, Berlin (2002) 225-233.
11. Michaelsen E., Stilla U.: Good Sample Consensus Estimation of 2d-Homographies for Vehicle Movement Detection from Thermal Videos. In: Ebner H., Hiepke C., Mayer H., Pakzad K. (eds.): Photogrammetric Image Analysis PIA´03. Intern. Arch. of Photogr. and Rem. Sens., Vol. 34, Part 3/W8 (2003) 125-130.
12. Michaelsen E., Stilla U.: Sensor Pose Inference from Airborne Videos by Decomposing Homography Estimates. Accepted for ISPRS 2004, Commission III, WG III/I (2004)
13. Torr P. H. S., Davidson C.: IIMSAC: Synthesis of importance sampling and random sample consensus. IEEE – PAMI, Vol. 25, no. 3 (2003) 354-364.
14. Sawhney, H. S.: Simplifying motion and structure analysis using planar parallax and image warping. ICPR 94, IEEE-Press, Los Alamitos, Ca., Vol. I (1994) 403-407.
15. Stilla U., Michaelsen E., Lütjen K.: Automatic Extraction of Buildings from Aerial Images. In: F. Leberl, R. Kalliany, M. Gruber (eds.), Mapping Buildings, Roads and other Man-made Structures from Images, IAPR-TC7, Wien, Oldenburg, (1996) 229-244.
16. Sujew, S., Ernst, I.: LUMOS Airborne Traffic Monitoring. In: Int. Workshop on Airborne Traffic Measurement, DLR, Berlin, (2003).
17. Robust Estimation Library, rrel, university of Manchester (accessed 24 Dec. 2003), http://paine.wiau.man.ac.uk/pub/doc_vxl/contrib/rpl/rrel/html/