OMNIDIRECTIONAL CAMERA MOTION ESTIMATION

Akihiko Torii and Tomáš Pajdla
Center for Machine Perception, Department of Cybernetics, Faculty of Elec. Eng.
Czech Technical University in Prague, Karlovo nám. 13, 121 35, Prague, Czech Republic

Keywords: Camera Motion Estimation, Omnidirectional Images, Epipolar Geometry.

Abstract: We present an automatic technique for computing relative camera motion and simultaneous omnidirectional image matching. Our technique works for small as well as large motions, tolerates multiple moving objects and very large occlusions in the scene. We combine three principles and obtain a practical algorithm which improves the state of the art. First, we show that the correct motion is found much sooner if the tentative matches are sampled after ordering them by the similarity of their descriptors. Secondly, we show that the correct camera motion can be better found by soft voting for the direction of the motion than by selecting the motion that is supported by the largest set of matches. Finally, we show that it is useful to filter out the epipolar geometries which are not generated by points reconstructed in front of cameras. We demonstrate the performance of the technique in an experiment with 189 image pairs acquired in a city and in a park. All camera motions were recovered with the error of the motion direction smaller than 8°, which is 4 % of the 183° field of view, w.r.t. the ground truth.

1 INTRODUCTION

Projections of scene points into images acquired by a moving camera are related by epipolar geometry (Hartley and Zisserman, 2004). In this work, we present a practical algorithm which improves the state of the art automatic camera relative motion computation and simultaneous image matching. Such an algorithm is a useful building block for autonomous navigation and building large 3D models using structure from motion.

In contrary to existing structure from motion algorithms, e.g. (2d3 Ltd, ; Davison and Molton, 2007; Cornelis et al., 2006; Williams et al., 2007), which solve the problem when the camera motion is small or once 3D structure is initialized, we aim at a more general situation when neither the relationship between the cameras nor the structure is available. In such case, 2-view camera matching and relative motion estimation is a natural starting point to camera tracking and structure from motion. This is an approach of the state of the art wide base-line structure from motion algorithms e.g. (Brown and Lowe, 2003; Martinec and Pajdla, 2007) that start with pair-wise image matches and epipolar geometries which they next clean up and make them consistent by a large scale bundle adjustment.

In this paper, we improve the state of the art approach to automatic computation of relative camera motion (Hartley and Zisserman, 2004; Nistér and Engels, 2006) and simultaneous image matching (Pritchett and Zisserman, 1998; Tuytelaars and Gool, 2000; Schaffalitzky and Zisserman, 2001; Matas et al., 2004) by combining three ingredients which altogether significantly increase the quality of the result.

To illustrate the problem we shall now discuss four interesting examples of camera motions which have gradually increasing level of difficulty.

Figure 1(a) shows an easy pair which can be solved by a standard RANSAC estimation (Hartley and Zisserman, 2004). 57%, i.e. 1400, of tentative matches are consistent with the true motion. Figure 1(c) shows a dominant peak in the data likelihood $p(M|\mathbf{e})$ of matches given the motion direction (Nistér and Engels, 2006) meaning that there is only one motion direction which explains a large number of matches.

Figure 2(a) shows a more difficult pair that contains multiple moving objects, large camera rotation and considerable occlusion in the scene. Only 8%, i.e. 120, tentative matches were consistent with the true motion. Figure 2(c) shows that there are many

Torii A. and Pajdla T. (2008).
OMNIDIRECTIONAL CAMERA MOTION ESTIMATION.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 577-584
DOI: 10.5220/0001084505770584

motion directions with high support, in this case from wrong tentative matches.

Figure 3(a) shows an even more difficult pair since only 1.4%, i.e. 50, tentative matches are consistent with the true motion. There are very many wrong tentative matches on bushes where local image features are all small and green. Thus, many motion directions get high support from wrong matches. The true motion has the highest support but its peak is very sharp and thus difficult to find in limited time.

Figure. 4(a) shows a very difficult pair that contains large camera rotation and many repetitive features which generate wrong tentative matches. In this case, the motion supported by the largest number of tentative matches is incorrect. Notice that the peak of the likelihood of matches in Fig. 4(c) does not correspond to the direction of the true motion.

All the above examples can be solved correctly by the technique presented in this paper.

The state of the art technique for finding relative camera orientations from image matches first establishes tentative matches by pairing image points with mutually similar features and then uses RANSAC (Fischler and Bolles, 1981; Hartley and Zisserman, 2004; Chum and Matas, 2005) to look for a large subset of the set of tentative matches which is, within a predefined threshold θ, consistent with an epipolar geometry (Hartley and Zisserman, 2004). Unfortunately, this strategy does not always recover the epipolar geometry generated by the actual camera motion. This has been observed, e.g., in (Li and Hartley, 2005).

Often, there are more models which are supported by a large number of matches. Thus the chance that the correct model, even if it has the largest support, will be found by running a single RANSAC is small. Work (Li and Hartley, 2005) suggested to generate models by randomized sampling as in RANSAC but to use soft (kernel) voting for a physical parameter, the radial distortion coefficient in that case, instead of looking for the maximal support. The best model is then selected as the one with the parameter closest to the maximum in the accumulator space. This strategy works when the correct, or almost correct, models met in the sampling provide consistent values of the parameter while the incorrect models with high support generate different values of the parameter. Here we show that this strategy works also when used for voting in the space of motion directions.

It has been demonstrated in (Chum and Matas, 2005) that ordering the tentative matches by their similarity may help to reduce the number of samples in RANSAC. Paper (Chum and Matas, 2005) brought two main contributions. First, PROSAC sampling strategy has been suggested which allows to uniformly sample from the list of tentative matches ordered ascendingly by the distance of their descriptors. It allows to start by drawing promising samples first and often hit sufficiently large configuration of good matches early. The second contribution concerned a modification of the RANSAC stoping criterion (Hartley and Zisserman, 2004, p. 119) to be able to deal with very long sets of tentative matches without the necessity to know their number beforehand.

When working with perspective images, it is generally accepted (Hartley and Zisserman, 2004) that the best way to evaluate the quality of an epipolar geometry is to look at image reprojection errors. This is, for two images, equivalent to evaluating the distances of image points to their corresponding epipolar lines. We compared the image reprojection error with the residuals evaluated as the angle between rays and their corresponding epipolar planes, which we refer as the *angular error* here. In our experience, when cameras are calibrated, the angular error can safely be used instead of the image reprojection error. To be absolutely correct, every ray should be accompanied by a covariance matrix determining its uncertainty. The matrix depends on (i) image measurement error model and (ii) on the point position in the image. The point position determines how the unit circle around the point maps into the cone around the ray. In this paper we neglected the variability of the covariance matrix across the field of view and assumed it to be a scaled identity.

Next we describe how we combine ordered sampling of tentative matches, soft voting, and the orientation (cheirality) constraint (Hartley and Zisserman, 2004) on minimal five points used for computing camera motions to get an algorithm which solves all camera motions.

2 THE ALGORITHM

Algorithm 1 presents the pseudocode of the algorithm used to generate results described in this work. Next we describe the key parts of the algorithm in detail.

2.1 Detecting Tentative Matches and Computing their Descriptors

MSER (Matas et al., 2004), Harris-Affine and Hessian Affine (Mikolajczyk et al., 2005) affine covariant feature regions are detected in images. These features are alternative to popular SIFT features (Lowe, 2004) and work comparably in our situation. Parameters of the detectors are chosen to limit

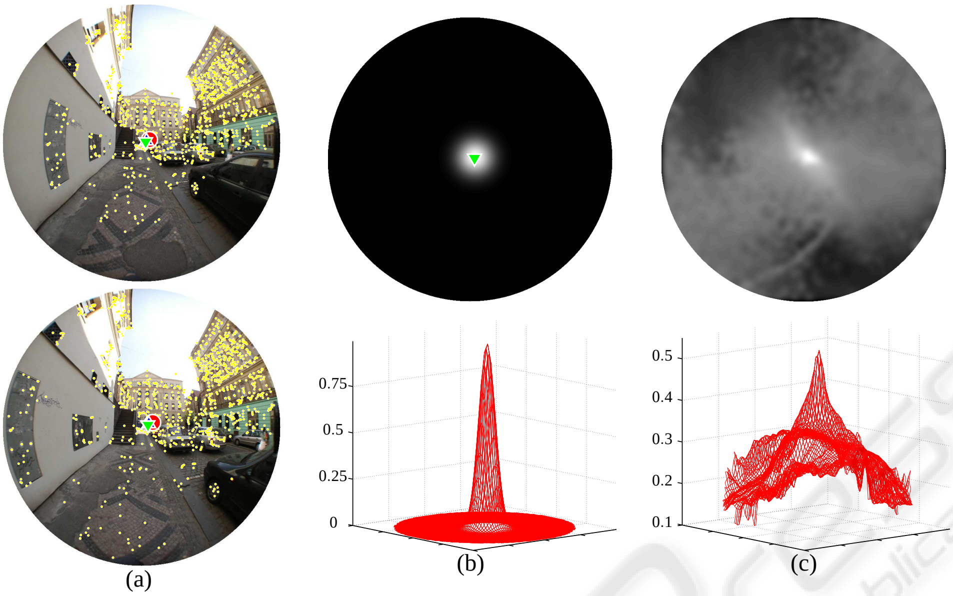

Figure 1: An easy example of camera motions. (a): Fist (top) and second (bottom) images. Red ○, blue △, and green ▽ (our result) represent the true epipole, the epipole computed by maximizing the support, and the epipoles computed by soft voting for the position of the epipole, respectively. Small dots show the matches giving green ▽. (b): Voting space for the motion direction in the first image generated by 50 soft votes casted by the result of 500-sample PROSAC, visualized on the image plane (top) as a 3D plot (bottom). White represents large number of votes. The peak corresponds to green ▽ (our result). (c): The maximal support for every epipole (i.e. CIF image from (Nistér and Engels, 2006)). White represents hight support. The image space has been uniformly sampled by 10000 epipoles and for each epipole the size of support of the best model found by 500-sample PROSAC has been recorded.

the number of regions to 1-2 thousands per image. The detected regions are assigned local affine frames (LAF) (Obdržálek and Matas, 2002) and transformed into standard positions w.r.t. their LAFs. Discrete Cosine Descriptors (Obdržálek and Matas, 2003) are computed for each region in the standard position. Finally, mutual distances of all regions in one image and all regions in the other image are computed as the Euclidean distances of their descriptors and tentative matches are constructed by selecting the mutually closest pairs.

MSER region detector is approximately 100 times faster than the Harris and Hessian Affine region detector but MSERs alone were not able to solve all image pairs in our data. MSERs perform great in urban environment with contrast regions, such as windows, doors and markings. However, they often provide many useless regions on natural scenes because they tend to extract contrast regions which often do not correspond to real 3D structures, such as regions formed by tree branches against the sky or shadows casted by leaves.

2.2 Ordered Randomized Sampling

We use *ordered sampling* as suggested in (Chum and Matas, 2005) to draw samples from tentative matches ordered ascendingly by the distance of their descriptors. We keep the original RANSAC stopping criterion (Hartley and Zisserman, 2004) and we limit the maximum number of samples to 500. We have observed that pairs which could not be solved by the ordered sampling in 500 samples got almost never solved even after many more samples. Using the stopping criterion from (Chum and Matas, 2005) often leads to ending the sampling prematurely since the criterion is designed to stop as soon as a large non-random set of matches is found. Our objective is, however, to find a globally optimal model and not to stop as soon as a local model with large support is found.

We have observed that there are often several alternative models with the property that the correct model of the camera motion has a similar or only slightly larger support than other models which are not correct. Algorithm 1 would provide almost identical results even without the RANSAC stopping criterion but the criterion helps to end simple cases sooner

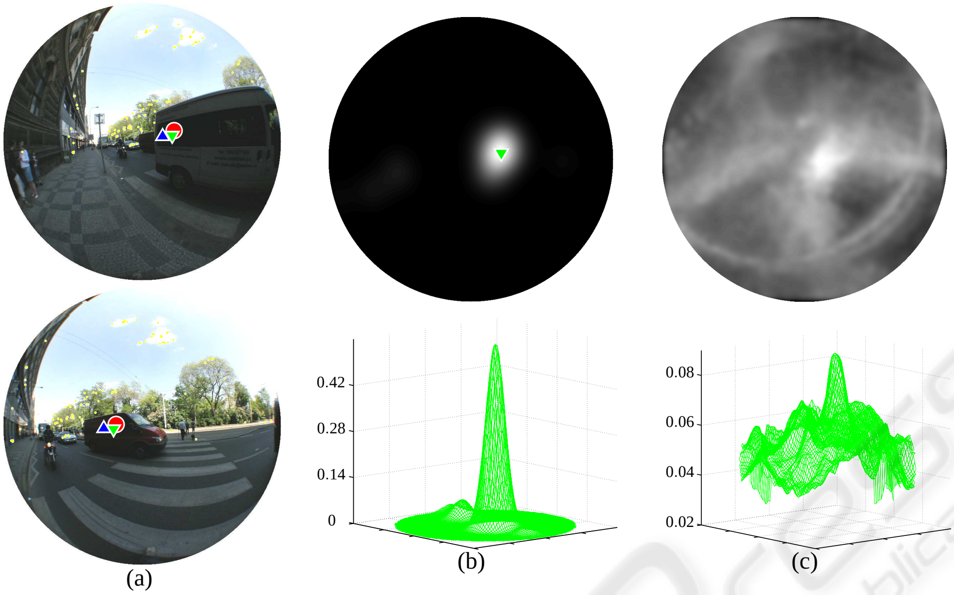

Figure 2: A more difficult example of camera motions. See Fig. 1.

than after 500 samples.

Having a calibrated camera, we draw 5-tuples of tentative matches from the list $M = [\mathbf{m}]_1^N$ of tentative matches ordered ascendingly by the distance of their descriptors. From each 5-tuple, relative orientation is computed by solving the 5-point minimal relative orientation problem for calibrated cameras (Nistér, 2004; Stewénius, 2005).

Row (*sim*) in Fig. 5 shows that many more correct motions have been sampled in 500 samples of PROSAC using ordered matches than by using the same number of samples on a randomly ordered list of matches, row (*rnd*).

2.3 Orientation Constraint

An essential matrix can be decomposed into four different camera and point configuration which differ by the orientation of cameras and points (Hartley and Zisserman, 2004). Without enforcing the constraint that all points have to be observed in front of the cameras, some epipolar geometries may be supported by many matches but it need not be possible to reconstruct all points in front of both cameras.

For omnidirectional cameras, the meaning of infrontness is a generalization of the classical infrontness for perspective cameras. With perspective cameras, a point X is in front of the camera when it has a positive z coordinate in the camera coordinate system. For omnidirectional cameras, a point X is in front of the camera if its coordinates can be written as a positive multiple of the direction vector which represents the half-ray by which X has been observed.

In general, it is beneficial to use only those matches which generate points in front of cameras. However, this takes time to verify it for all matches. On the other hand, it is fast to verify whether the five points in the minimal sample generating the epipolar geometry can be reconstructed in front of both cameras and to reject such epipolar geometries which do not allow it.

Row (*oc*) in Fig. 5 shows that the number of incorrectly estimated motions decreased when such epipolar geometries were excluded by this orientation constraint.

Furthermore, the orientation constraint in average reduces the computational cost because it avoids evaluating residuals corresponding to many wrong camera motions.

2.4 Soft Voting

In this paper, we vote in two-dimensional accumulator for the estimated motion direction. However, unlike in (Li and Hartley, 2005; Nistér and Engels, 2006), we do not cast votes directly by each sampled epipolar geometry but by the best epipolar geometries recovered by ordered sampling of PROSAC. This way the votes come only from the geometries that have very high support. We can afford to compute more, e.g. 50, epipolar geometries since the ordered sampling is much faster than the standard

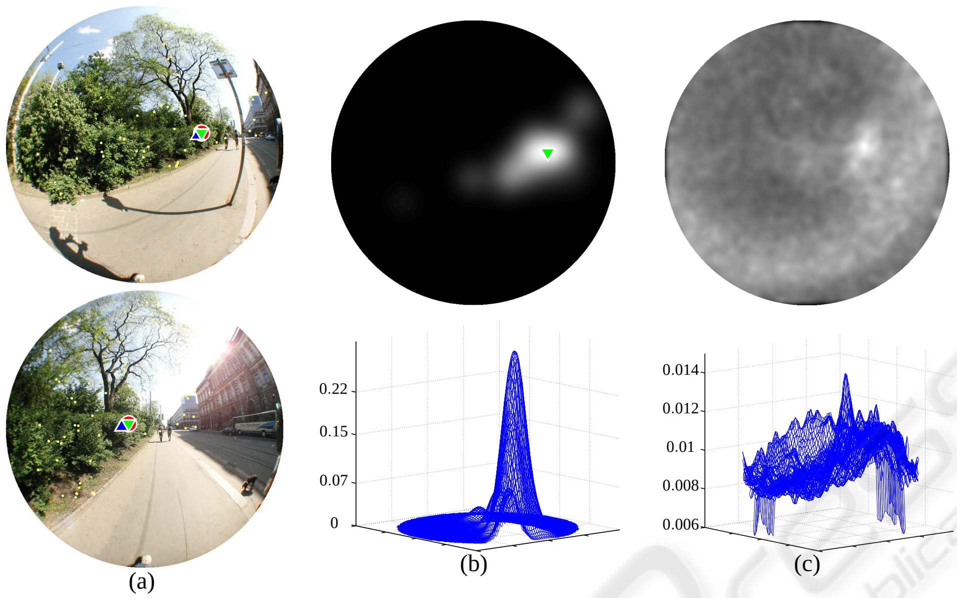

Figure 3: Even more difficult example of camera motions. See Fig. 1.

RANSAC. Altogether, we need to evaluate maximally $500 \times 50 = 25000$ samples to generate 50 soft votes, which is comparable to running a standard 5-point RANSAC for expected contamination by 84 % of mismatches (Hartley and Zisserman, 2004, p. 119). Yet, with our technique, we could go up to 98.5 % of mismatches with comparable effort. The relative camera orientation with the motion direction closest to the maximum in the voting space is finally selected.

Figure 5 shows the improvement of using soft voting for finding the relative motion when casting 50 soft votes. On several difficult image pairs, such as Fig. 4, the motion supported by the largest number of tentative matches was incorrect but the soft voting provided a motion close to the ground truth.

3 EXPERIMENT

3.1 Image Data

Experimental data consist of 189 image pairs obtained by selecting consecutive images of an image sequence. We do not use the fact that images were taken in a sequence and our method works for any pair of images. The distance between two consecutive images was 1-3 meters. Most of the camera motions have rotations up to 15° but large rotations of 45° are also present. Images were acquired by Kyocera Finecam M410R with Nikon FC-E9 fisheye-lens with 183° view angle. The image projection is equiangular and was internally calibrated (Mičušík and Pajdla, 2006) beforehand. Images were digitized in resolution 800 pixels/183°, i.e. 0.2°/pixel, which is comparable to 240×180 pixels for more standard 40° view angle. Acquisition of images started in a narrow street with buildings on both sides, then continued to a wider street with many driving cars, and finally lead to a park with threes, bushes and walking people.

3.2 Ground Truth Motion

For most image pairs, the “true” camera motions were recovered by running the Algorithm 1 a number of times and checking that (i) the true motion has been repeatedly generated by correctly matched 5-tuples of matches and that (ii) the motion direction pointed to the same object in both images. In a few image pairs, for which we could not get a decisive number of consistent results, the true motion has been generated from a 5-tuple of correct matches selected manually. We estimate the precision of our ground truth motion estimation to be higher than 4 % of the view field, which corresponds to 8° and 32 image pixels.

3.3 Result

Figure 5(*SV*) shows the quality of the estimated camera motion by Algorithm 1. The algorithm looks for the motion with motion direction closest to the global maximum in the accumulator after casting soft votes

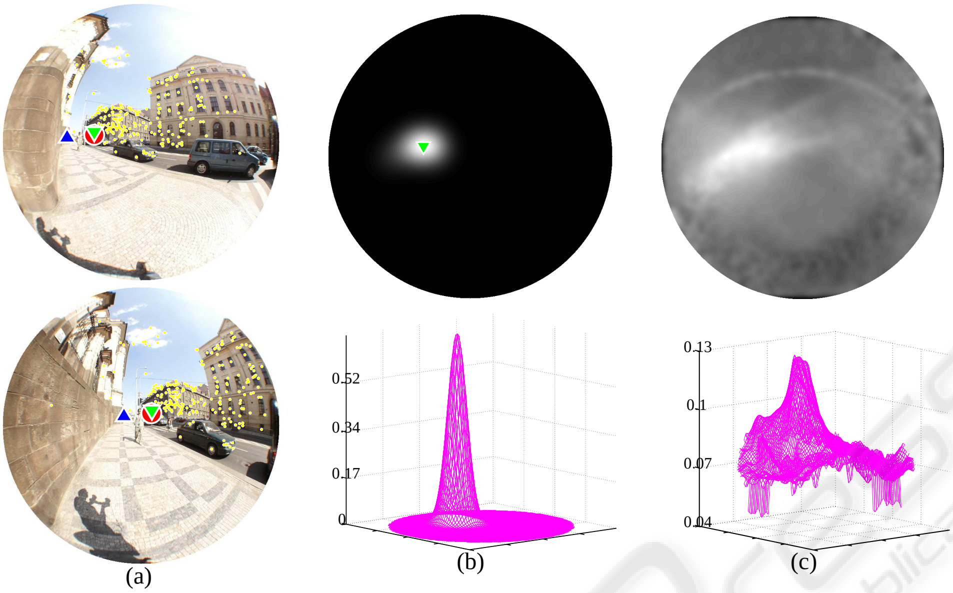

Figure 4: A very difficult example of camera motions. See Fig. 1. The motion direction with the largest support of 294 matches (b) is wrong. Our algorithm 1 finds the correct motion which is supported "only" by 287 matches.

from 50 motions. The 50 motions for soft votes are estimated in 500 samples by the ordered sampling based on residuals evaluated as the angle between rays and their corresponding epipolar planes. All motions in our test data were estimated with motion directions within 8°, i.e. 4 % of the view angle, from the ground truth.

4 CONCLUSIONS

We have presented a practical algorithm which can compute camera motions from omnidirectional images. We have improved the state of the art by combining ordered sampling from tentative matches ordered by their descriptor similarity with the orientation constraint and soft voting. We used angular residual error which better fits to the geometry of omnidirectional cameras. Our algorithm was able to correctly compute motions of all tested 189 image pairs.

ACKNOWLEDGEMENTS

The authors were supported by EC project FP6-IST-027787 DIRAC and by Czech Government under the reseach program MSM6840770038. Any opinions expressed in this paper do not necessarily reflect the views of the European Community. The Community is not liable for any use that may be made of the information contained herein.

REFERENCES

2d3 Ltd. Boujou. *http://www.2d3.com*.

Brown, M. and Lowe, D. G. (2003). Recognising panoramas. In *ICCV '03*, Washington, DC, USA.

Chum, O. and Matas, J. (2005). Matching with PROSAC - progressive sample consensus. In *CVPR '05*, volume 1, pages 220–226, Los Alamitos, USA.

Cornelis, N., Cornelis, K., and Gool, L. V. (2006). Fast compact city modeling for navigation pre-visualization. In *CVPR '06*, pages 1339–1344, Washington, DC, USA.

Davison, A. J. and Molton, N. D. (2007). Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

Hartley, R. I. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition.

Li, H. and Hartley, R. (2005). A non-iterative method for correcting lens distortion from nine point correspondences. In *OMNIVIS '05*.

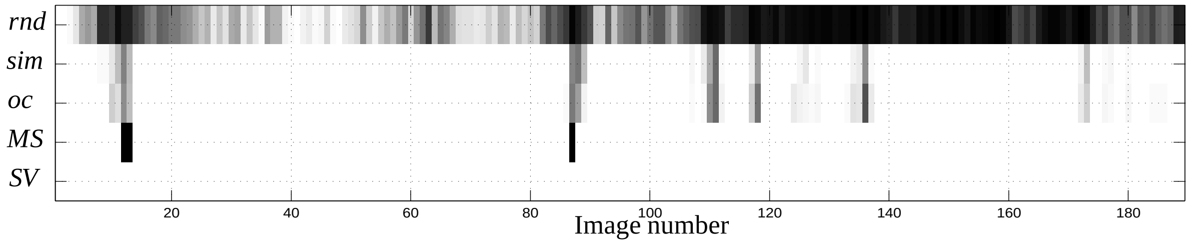

Figure 5: Camera motion estimation obtained by the ordered sampling, orientation constraint, and soft voting. The graph shows the number of camera motion directions that are not further than 8° from the ground truth as a function of the image pair $(i, i+1)$ with i on the horizontal axis. Lighter color represents higher numbers. The motion directions were computed from the first 500 5-point samples drawn from tentative matches ordered randomly, row (*rnd*), and tentative matches ordered by their similarity, row (*sim*). Ordering of tentative matches by similarity greatly increased the number of correctly estimated epipoles. Row (*oc*) shows the number of epipolar geometries for which the 5 points of their generating minimal sample get reconstructed in front of both cameras (orientation constraint). Using the orientation constraint further improves the result. Row (*MS*) shows whether the motion direction, which is supported by the largest number of tentative matches in 50 motions, is (black) or is not (white) more than 8° apart from the true motion direction. Row (*SV*) shows whether the estimated motion direction, which is closest to the global maximum in the accumulator, is (black) or is not (white) more than 8° apart from the true motion direction. This is the best strategy for estimating the camera motion.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Martinec, D. and Pajdla, T. (2007). Robust rotation and translation estimation in multiview reconstruction. In *CVPR '07*, Minneapolis, MN, USA.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767.

Mičušík, B. and Pajdla, T. (2006). Structure from motion with wide circular field of view cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1135–1149.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V. (2005). A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72.

Nistér, D. (2004). An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–777.

Nistér, D. and Engels, C. (2006). Estimating global uncertainty in epipoloar geometry for vehicle-mounted cameras. In *SPIE, Unmanned Systems Technology VIII*, volume 6230.

Obdržálek, Š. and Matas, J. (2002). Object recognition using local affine frames on distinguished regions. In *BMVC '02*, volume 1, pages 113–122, London, UK.

Obdržálek, Š. and Matas, J. (2003). Image retrieval using local compact dct-based representation. In *DAGM '03*, number 2781 in LNCS, pages 490–497, Berlin, Germany.

Pritchett, P. and Zisserman, A. (1998). Wide baseline stereo matching. In *ICCV '98*, pages 754–760, Bombay, India.

Schaffalitzky, F. and Zisserman, A. (2001). Viewpoint invariant texture matching and wide baseline stereo. In *ICCV '01*, Vancouver, Canada.

Stewénius, H. (2005). *Gröbner Basis Methods for Minimal Problems in Computer Vision*. PhD thesis, Centre for Mathematical Sciences LTH, Lund University, Sweden.

Tuytelaars, T. and Gool, L. V. (2000). Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC '00*, Bristol, UK.

Williams, B., Klein, G., and Reid, I. (2007). Real-time slam relocalisation. In *ICCV '07*, Rio de Janeiro, Brazil.

Algorithm 1 Camera motion estimation by ordered sampling from tentative matches with geometrical constraints.

```
Input:  Image pair $I_1$, $I_2$.
    $\theta := 0.3^\circ$ ... the tolerance for establishing matches
    $\sigma := 4^\circ$ ... the standard deviation of Gaussian kernel for soft voting
    $N_V := 50$ ... the number of soft votes
    $N_S := 500$ ... the maximum number of random samples.
    $\eta := 0.95$ ... the termination probability of the standard RANSAC (Hartley and Zisserman, 2004, p. 119).
Output: Essential matrix $\mathtt{E}^*$.

1. Detect tentative matches and compute their descriptors.
 1.1 Detect affine covariant feature regions MSER-INT+, MSER-INT-, MSER-SAT+,
     MSER-SAT-, APTS-LAP, and APTS-HES in left and right images, Sec. 2.1.
 1.2 Assign local affine frames (LAF) (Obdržálek and Matas, 2002) to the regions and transform the regions
     into a standard position w.r.t. their LAFs.
 1.3 Compute Discrete Cosine Descriptors (Obdržálek and Matas, 2003) for each region in the standard position.
2. Construct the list $M = [\mathbf{m}]_1^N$ of tentative matches with mutually closest descriptors. Order the list ascendingly
   by the distance of the descriptors. $N$ is the length of the list.
3. Find a camera motion consistent with a large number of tentative matches:
   Set $D$ to zero. // Initialize the accumulator of camera translation directions.
   for $i := 1, \ldots, N_V$ do
     $t := 0$ // The counter of samples.   $n := 5$ // Initial segment length.
     $N_T := N_S$ // Initial termination length.
     while $t \le N_T$ do
       if $t = \lceil 200000\binom{n}{5}/\binom{N}{5}\rceil$ (Chum and Matas, 2005) then
         $n := n+1$ // The maximum number of samples for the current initial segment reached, increase
         the initial segment length.
       end if
       $t := t+1$ // New sample
       Select the 5 tentative matches $M_5$ of the $t^{th}$ sample by taking 4 tentative matches from $[\mathbf{m}]_1^{n-1}$ at
       random and adding the $5^{th}$ match $\mathbf{m}_n$.
       $\mathtt{E}_t :=$ the essential matrix by solving the 5-point minimal problem for $M_5$ (Nistér, 2004; Stewénius,
       2005).
       if $M_5$ can be reconstructed in front of cameras (Hartley and Zisserman, 2004, p. 260) then
         $S_t :=$ the number of matches which are consistent with $\mathtt{E}_t$, i.e. the number of all matches $\mathbf{m} =$
         $[\mathbf{u}_1, \mathbf{u}_2]$ for which $\max(\measuredangle(\mathbf{u}_1, \mathtt{E}_t\mathbf{u}_2), \measuredangle(\mathbf{u}_2, \mathtt{E}_t^\top\mathbf{u}_1)) < \theta$.
       else
         $S_t := 0$
       end if
       $N_R := \log(\eta)/\log\left(1 - \binom{S_t}{5}/\binom{N}{5}\right)$ //The termination length defined by the maximality con-
       straint (Hartley and Zisserman, 2004, p. 119).
       $N_T := \min(N_T, N_R)$ // Update the termination length.
     end while
     $\hat{t} = \arg_{t=1,\ldots,N_S}\max S_t$ // The index of the sample with the highest support.
     $\hat{\mathtt{E}}_i := \mathtt{E}_{\hat{t}}$, $\hat{\mathbf{e}}_i :=$ camera motion direction for the essential matrix $\mathtt{E}_{\hat{t}}$.
     Vote in accumulator $D$ by the Gaussian with sigma $\sigma$ and mean at $\hat{\mathbf{e}}_i$.
   end for
   $\hat{\mathbf{e}} := \arg_{\mathbf{x}\in domain(D)}\max D(\mathbf{x})$ // Maximum in the accumulator.
   $i^* := \arg_{i=1,\ldots,50}\min\measuredangle(\hat{\mathbf{e}}, \hat{\mathbf{e}}_i)$ // The motion closest to the maximum.
   $\mathtt{E}^* := \hat{\mathtt{E}}_{i^*}$ // The “best” camera motion.
4. Return $\mathtt{E}^*$.
```