Measuring camera translation by the dominant apical angle

Akihiko Torii[1] Michal Havlena[1] Tomáš Pajdla[1]
[1]CMP, Czech Technical University
Prague, Czech Republic
{torii,havlem1,pajdla}@cmp.felk.cvut.cz

Bastian Leibe[2]
[2]ETH Zürich
Zürich, Switzerland
leibe@vision.ee.ethz.ch

Abstract

This paper provides a technique for measuring camera translation relatively w.r.t. the scene from two images. We demonstrate that the amount of the translation can be reliably measured for general as well as planar scenes by the most frequent apical angle, the angle under which the camera centers are seen from the perspective of the reconstructed scene points. Simulated experiments show that the dominant apical angle is a linear function of the length of the true camera translation. In a real experiment, we demonstrate that by skipping image pairs with too small motion, we can reliably initialize structure from motion, compute accurate camera trajectory in order to rectify images and use the ground plane constraint in recognition of pedestrians in a hand-held video sequence.

1. Introduction

Reliable structure from motion plays an important role in 3D reconstruction [1, 2, 3, 4, 5, 6, 7], self localization [8, 9], and object recognition [10, 11].

Structure from motion works best when camera translation is moderate such that local image features (e.g. LoG and SIFT [12], IBR and EBR [13], Harris-Affine and Hessian Affine [14], MSER [15] and LAF [16], salient regions [17]) provide sufficient number of correct tentative matches and the baseline is sufficiently long to triangulate 3D points from their image projections.

This paper focuses on the problem of reliable detection of too small camera translation from two images and demonstrates that such capability enhances structure from motion and object recognition from a video sequence taken by a moving camera. Since the scale of the reconstruction can't be determined from two images of a moving camera, the amount of the camera translation can be measured only relatively w.r.t. the observed scene.

We propose to measure the amount of the camera translation from pairwise image matches as the dominant apical angle (DAA) of 3D points reconstructed from the matches. The apical angle of a 3D point $\mathbf{X}$ is the angle under which the camera centers are seen from the perspective of the point $\mathbf{X}$.

Recently, the problem of detecting too small translation in structure from motion has been addressed in [4]. Camera motions were considered pure rotations if at least 90% of matches verified by an epipolar geometry were also verified in fitting a pure rotation. Another recent related work [18] looks at a related problem of determining the scale of the motion of a stereo rig with non-overlapping fields of view. We not only are able to detect very small motion but we can measure the amount of relative translation w.r.t. the scene by a linear function of the true size of the translation vector.

We show that the dominant apical angle is a linear function of the length of the true translation for general as well as planar scenes and that it can be reliably estimated in the presence of outliers. We show on simulated data that the measure is accurate and robust. We demonstrate in a real experiment comprising "too small motion" detection, structure from motion, view rectification, and pedestrian recognition, that the proposed measure enables to initialize structure from motion and to increase the accuracy of camera path computation, Figure 1.

2. Measuring amount of camera translation by the dominant apical angle

Consider a pair of calibrated cameras with the normalized camera matrices [19], $\mathtt{P} = [\mathtt{I} \,|\, \mathtt{0}]$ and $\mathtt{P}' = [\mathtt{R} \,|\, -\mathbf{T}]$ and an image point correspondence given by a pair of homogeneous coordinates $(\mathbf{x}, \mathbf{x}')$ represented by unit norm vectors, *i.e.* $\|\mathbf{x}\| = \|\mathbf{x}'\| = 1$. There holds $\lambda'\mathbf{x}' = \lambda\,\mathtt{R}\,\mathbf{x} - \mathbf{T}$, with real λ, λ', rotation $\mathtt{R}$ and translation $\mathbf{T}$.

If there was no noise, pure camera rotation, *i.e.* $\|\mathbf{T}\| = 0$, could be detected by finding $\mathbf{x}' = \mathtt{R}\,\mathbf{x}$ holds true for all correspondences. However, the probability of this situation is zero due to noise in image measurements even if the physical camera really rotates. Thus, in real situations, a non-zero essential matrix $\mathtt{E}$ can always be computed from noisy image matches by, e.g., the 5-point algorithm [20].

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

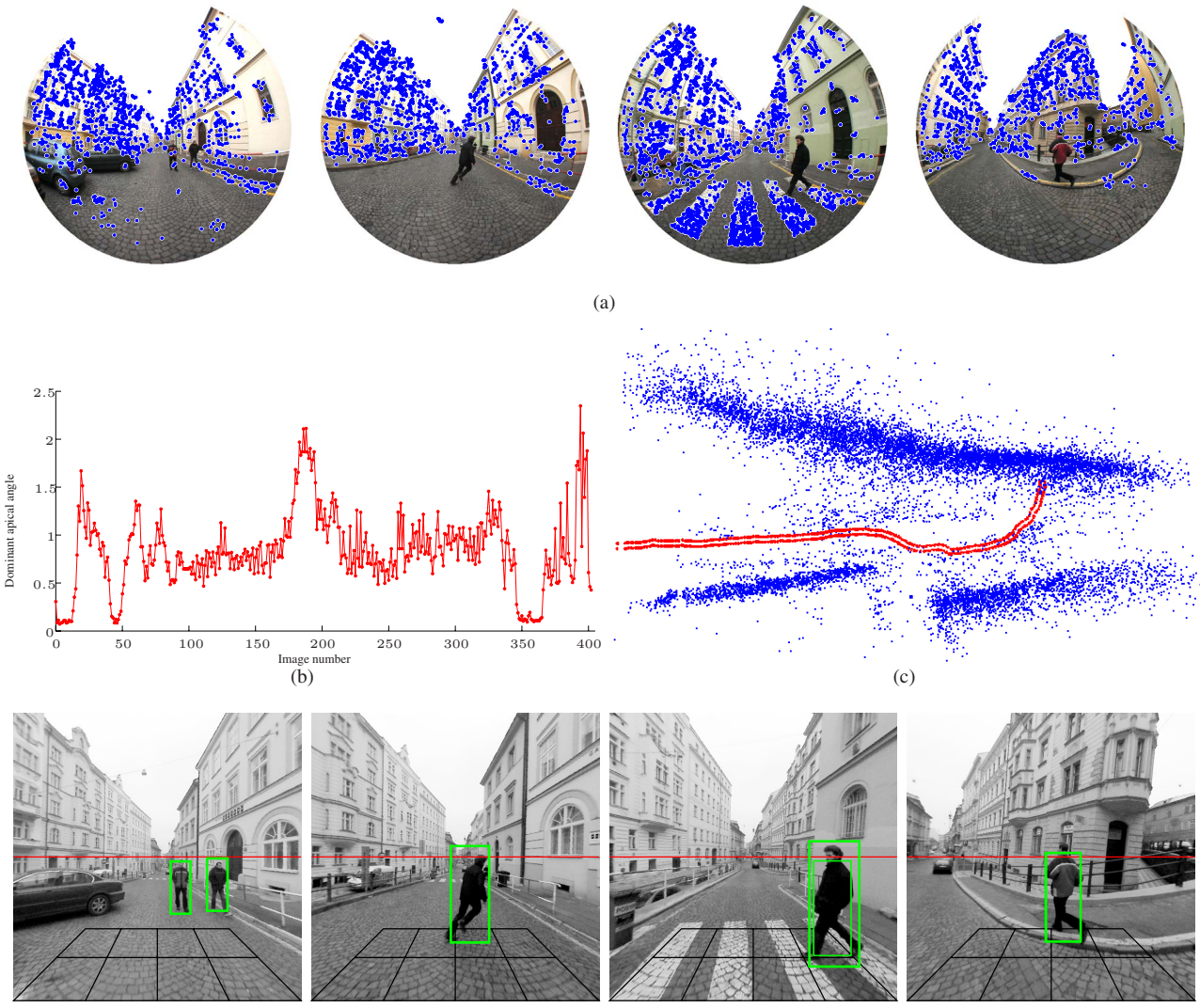


Figure 1. The length of the camera translation relatively to the scene is measured from pairwise image matches (a) as the dominant apical angle under which the camera centers are seen from the perspective of the reconstructed scene points (b). The structure from motion is robustly initialized, the camera path computed (c), the view stabilized and the ground plane tracked in order to reduce false detections of pedestrians (d).

The essential matrix $\mathtt{E}$ can be decomposed into $\mathtt{E} = [\mathbf{t}]_\times \mathtt{R}$, where $\mathtt{E}\,\mathbf{t} = 0$, in four different ways and the right decomposition can be selected to reconstruct all points in front of both cameras [19, $p260$]. The amount of camera translation can be measured only relatively to the distance and the size of the observed scene since large motions w.r.t. a distant scene generate same image correspondences as small motions w.r.t. a close scene.

Having n matches $\{(\mathbf{x}_i, \mathbf{x}'_i)\}_{i=1,\dots,n}$ and the essential matrix $\mathtt{E}$ computed from them, we can reconstruct n 3D points $\{\mathbf{X}_i\}_{i=1,\dots,n}$. Figure 2 shows a point $\mathbf{X}$ reconstructed from image matches $(\mathbf{x}, \mathbf{x}')$. For each point $\mathbf{X}$, there is the apical angle τ, which measures the length of the camera translation from the perspective of the point $\mathbf{X}$. If the cameras are related by a pure rotation, all angles τ are equal to zero. The larger the camera translation is, the larger the angles τ are. The closer the point $\mathbf{X}$ to the midpoint of the camera baseline is, the larger the corresponding τ is. In fact, measuring the apical angles is equivalent to measuring disparities on a spherical retina as the corresponding angles.

For a given $\mathtt{E}$ and matches $\{(\mathbf{x}_i, \mathbf{x}'_i)\}_{i=1,\dots,n}$, one can

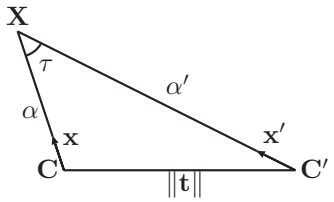


Figure 2. The apical angle τ at the point $\mathbf{X}$ reconstructed from the correspondence $(\mathbf{x},\, \mathbf{x}')$ relatively depends on the length of the camera translation $\mathbf{t}$ and on the distances of $\mathbf{X}$ from the camera centers $\mathbf{C}$, $\mathbf{C}'$.

select the decomposition of $\mathtt{E}$ to $\mathtt{R}$ and $\mathbf{t}$, which reconstruct the largest number of 3D points in front of the cameras. The apical angle τ_i, corresponding to a match $(\mathbf{x}_i, \mathbf{x}'_i)$, is computed by solving the set of linear equations for the relative distances α_i, α'_i

$$\alpha' \mathbf{x}'_i = \alpha\, \mathtt{R}\, \mathbf{x}_i - \mathbf{t} \tag{1}$$

in the least square sense and by using the law of cosines $2\,\alpha_i\,\alpha'_i \cos(\tau_i) = {\alpha_i}^2 + {\alpha'_i}^2 - \|\mathbf{t}\|^2$.

For a small translation w.r.t. the distance to the scene points, it is natural to use the approximation $\alpha_i = \alpha'_i$. Then, the apical angle τ_i becomes a linear function of $\|\mathbf{t}\|$. This is instantly proved by using the approximated equation of the law of cosine $\cos(\tau_i) = 1 - \|\mathbf{t}\|^2/2\alpha_i^2$ and the cosine series expansion $\cos(\tau_i) = 1 - \tau_i^2/2! + \mathrm{O}(\tau_i^4)$.

If all matches were correct, the largest τ would best represent the amount of the translation. However, all matches are rarely correct and thus we need a measure of the translation which is robust. The distribution of values of τ_i depends on the distribution of the points in the scene and on mismatches if they are present. We have observed that for many general 3D as well as planar scenes, the distribution has a dominant mode

$$\tau^* = \arg_{\{\tau_i\}_{i=1}^n} \max\; g(\tau_i) \tag{2}$$

where $g(\tau)$ performs the kernel voting with Gaussian smoothing [21], and that the mode τ^* predicts the length of the translation well. In particular, we have observed that for a fixed scene, the angle τ^* is a linear function of $\|\mathbf{T}\|$. Figures 3(b), 4(b), 5(b) show τ^* as a function of $\|\mathbf{T}\|$ evaluated in simulated experiments for spherical, planar and combined scene points and forward and lateral camera motions. The function is clearly linear. The slope of the linear function depends on the point distribution as well as on the direction of camera translation w.r.t. the scene points but we can see in Figures 3(b), 4(b), 5(b) that the variations of the slope are relatively insignificant. Having large dominant apical angle means that the majority of points can be reliably reconstructed. That provides a certificate of sufficiently large camera translation w.r.t. the size of the scene.

3. Experiments on simulated data

Figures 3, 4, and 5 show the results of simulated experiments for three different scenes, different motion directions, and for the length of the translation increasing from zero to a large value. The amount of camera translation was computed by the method based on RANSAC [22], which is described in **Algorithm 1**. Notice that we use a combination of ordered sampling [23] with kernel voting to maximize the chance of recovering correct epipolar geometry [24]. We also enforce the reconstructed points to be in front of cameras before counting the support size in the RANSAC.

Figure 3 shows an experiment with a general 3D scene consisting of 1000 points uniformly distributed in a hemisphere with the center at $(0, 0, 10)^\top$ and radius 25, Figure 3(a). The first camera was placed at $\mathbf{T}_1 = (0, 0, 0)^\top$ looking towards the scene points. Two motions of the second camera were tested. The backward motion was constructed as $\mathbf{T}_2 = (0, 0, -s)^\top$, *i.e.* we were moving away from the scene. The sideways motion was constructed as $\mathbf{T} = (s, 0, 0)^\top$. In both cases, s changed from 0 to 5.

3D points were projected by normalizing their coordinate vectors, constructed w.r.t. the respective camera coordinate systems, to unit length. To simulate imprecision of the digitization and image measurement, Gaussian noise with standard deviation $\sigma = 3^\circ$, corresponding to 1.3 pixels in a 800×800 image capturing 180° field of view, was added to the normalized vectors.

Figure 3(b) shows the dominant apical angle (DAA) as a function of the length of the true translation. DAA for the backward motion is shown by the blue line with "$+$" markers, whereas DAA for the lateral motion is shown by the red line with "$\times$" markers, both computed from noisy measurements. The green lines with "$\circ$" and "$\square$" markers, respectively, show the respective DAA of the backward and lateral motions computed from exact measurements. We see that the DAA is a linear function of the length of the true motion for translations longer than $0.25\ m$. The slope of the lateral DAA is slightly larger ($2.5\,^\circ/m$) than the slope ($2.0\,^\circ/m$) of DAA for the backward motion in this case. DAA of the zero translation computed from noisy matches is slightly above the zero due to noise in image measurements. Figure 3(c) shows the difference in the estimated camera rotation $\mathtt{R}_{est}$ w.r.t. the true rotation $\mathtt{R}$ evaluated as the angle of rotation of $\mathtt{R}_{est}^{-1}\mathtt{R}$. Notice that the error is constant for all lengths of the translation which shows that the rotation is computed correctly even if the direction of the translation, Figure 3(d), can't be found reliably.

Figures 4 and 5 show the same experiment as above on a planar scene and a 3D scene consisting of two planes. The results are comparable to Figures 3. In particular, we can see that we are able to measure the amount of translation in all three cases. It is interesting to notice that the error in rotation is constant for general 3D scenes, Figures 3(c), 5(c),

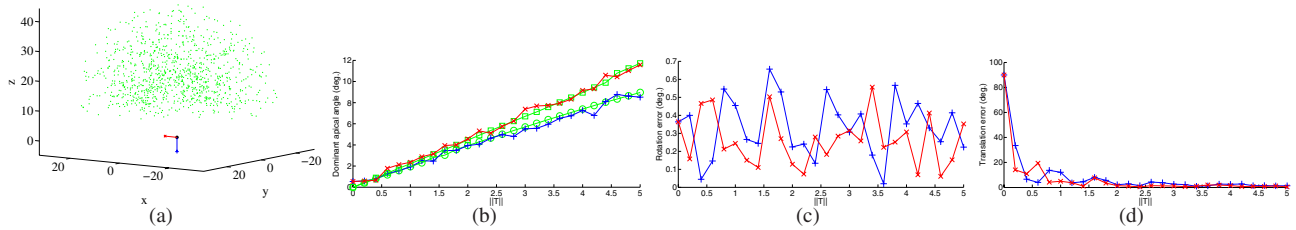


Figure 3. Measuring the length of the camera translation for a general 3D scene. (b) Dominant apical angle. Noisy data: blue line with "+" markers – backward motion, red line with "×" markers – lateral motion; exact data: green lines with "○" and "□" – backward and lateral motions, respectively. (c) Camera rotation error. (d) Camera translation error.

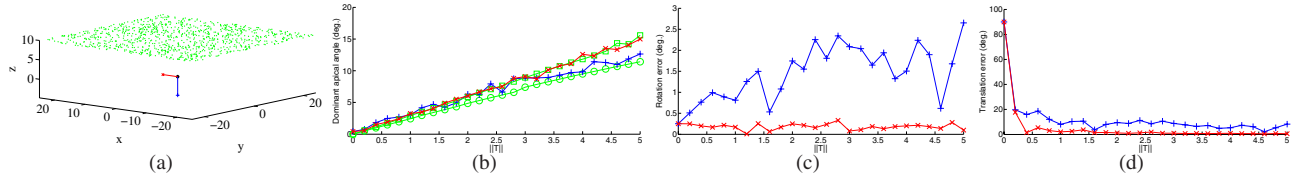


Figure 4. Measuring the length of the camera translation for a planar scene. (b) Dominant apical angle. Noisy data: blue line with "+" markers – backward motion, red line with "×" markers – lateral motion; exact data: green lines with "○" and "□" – backward and lateral motions, respectively. (c) Camera rotation error. (d) Camera translation error.

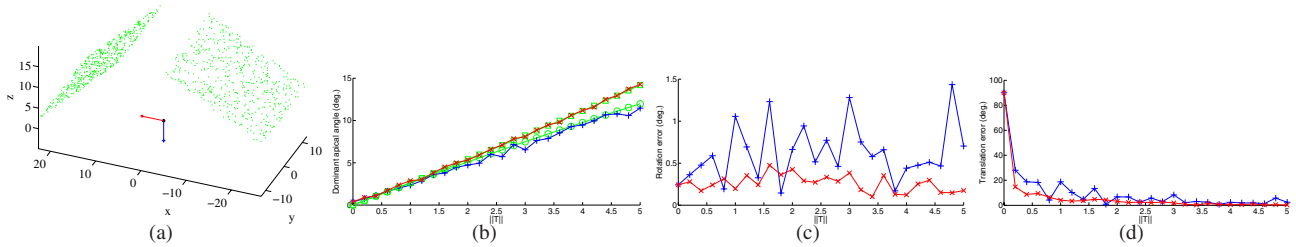


Figure 5. Measuring the length of the camera translation for the scene consisting of two planes. (b) Dominant apical angle. Noisy data: blue line with "+" markers – backward motion, red line with "×" markers – lateral motion; exact data: green lines with "○" and "□" – backward and lateral motions, respectively. (c) Camera rotation error. (d) Camera translation error.

but grows linearly for the planar scene, Figures 4(c). This reflects the fact that the angle which is occupied by scene points determines, to large extent, the quality of rotation estimation from scenes with shallow depth. At the same time, we can see that the quality of estimating the amount of camera translation has not been affected.

4. Experiment on real data

The experiment with real data demonstrates the use of the proposed method for measuring camera translation length in pedestrian detection from a moving camera.

Figure 1(a) shows four images from a 403 image sequence taken by a hand-held pair of cameras consisting of Nikon FC-E9 lens and Kyocera Finecam M410R providing resolution $0.23\ ^\circ/pixel$ at 3 frames per second. We demonstrate here our technique on large circular field of view images but it can be used with standard perspective images as well. The blue dots show image matches between consecutive images in the sequence obtained by **Algorithm 1**. The calibration of camera internal parameters has been done beforehand by the technique [6].

Figure 1(b) shows the dominant apical angle computed on the sequence. The DAA correctly shows that the camera started to translate on the 15-th frame and revealed two other stationary segments in the sequence between frames 41–50 and 346–365.

The structure from motion can greatly benefit from the ability to skip the frames with small translation at the beginning of the sequence since triangulating points from sufficient baseline provides stable 3D structure which can be robustly tracked. After removing the stationary segments

in the sequence, we estimate the motion of the camera rig by the structure from motion [3, 7]. Figure 1(c) shows the trajectory of the cameras (red dots) and the corresponding reconstructed feature points in 3D.

Using the camera trajectories, perspective cutouts with stabilized horizon (red lines in Figure 1(d)) are constructed. The estimated camera translation and rotation is used to track the position of the ground plane (black grid in Figure 1(d)). The position of the ground plane constraints pedestrian detector [10, 11] (green rectangles in Figure 1(d)).

The pedestrian detector is performed on the sequences of the perspective cutouts with the stabilized horizon and without the stabilization, that is, the window of cutout is fixed in the center of the original image. We evaluated that the number of correct detection (true positives), missing detection (false negatives, *i.e*. not detected pedestrians) and false detection (false positives, *i.e*. detected non-pedestrians) in both sequences. The total number of detectable pedestrians is the sum of true positives and false negatives. In both sequences, there were 65 true positives and 28 false negatives. In the sequence without the stabilization, there were 288 false positives. Our stabilization reduces the number of false positives to 146 while keeping other detection unchanged.

Figure 6 shows four examples of the pedestrian detection for the comparison between without the stabilization Figure 6(a) and with the stabilization Figure 6(b). The perspective cutouts with the stabilization obviously shows that the camera trajectories initialized by measuring the size of camera translation are estimated robustly and accurately because the horizontal lines (red lines in Figure 6) are well fit in the center of all cutouts. Furthermore, it is shown that the horizontal stabilization sufficiently reduced the false detections (yellow rectangles in Figure 6(a)).

5. Conclusion

We have introduced a technique for measuring the size of camera translation relatively to the observed scene. Our measure uses the dominant apical angle computed at the reconstructed scene points. The measure is a linear function of the length of the true translation, works for general as well as planar scenes, and is robust against mismatches. The experiments demonstrated that the measure can be used to improve the robustness of camera path computation and object recognition from hand-held cameras.

Acknowledgment

This work has been supported by grants EU FP6-IST-027787 DIRAC, MSM6840770038 DMCM III, STINT Dur IG2003-2 062 and MSMT KONTAKT 9-06-17.

References

[1] A. Akbarzadeh, J.-M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Steweius, R. Yang, G. Welch, H. Towles, D. Nister, and M. Pollefeys. Towards urban 3D reconstruction from video. *3DPVT 2006*.

[2] 2D3. Boujou. www.2d3.com

[3] N. Cornelis, K. Cornelis, and L. Van Gool. Fast compact city modeling for navigation pre-visualization. *IEEE CVPR 2006*, pp. 1339–1344.

[4] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. *IEEE CVPR 2007*.

[5] M. Antone and S. Teller. Scalable, absolute position recovery for omnidirectional image networks. *IEEE CVPR 2001*, Vol. 1, pp. 398–405.

[6] B. Micusik and T. Pajdla. Structure from motion with wide circular field of view cameras. *IEEE PAMI*, 28(7):1135–1149, 2006.

[7] M. Havlena, T. Pajdla, and K. Cornelis. Structure from omnidirectional stereo rig motion for city modeling. *VISAPP 2008*.

[8] T. Goedeme, M. Nuttin, T. Tuytelaars, and L. Van Gool. Omnidirectional vision based topological navigation. *IJCV*, 74(3):219–236, 2007.

[9] K. L. Ho and P. Newman. Detecting loop closure with scene sequences. *IJCV*, 74(3):261–286, 2007.

[10] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3D Scene Analysis from a Moving Vehicle. *IEEE CVPR 2007*.

[11] B. Leibe, K. Schindler, and L. Van Gool. Coupled Detection and Trajectory Estimation for Multi-Object Tracking. *IEEE ICCV 2007*.

[12] D. Lowe. Distinctive image features from scale invariant keypoints. *IJCV* 60(2):91–110, 2004.

[13] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *IJCV* 59(1):61–85, 2004.

[14] K. Mikolajczyk and C. Schmid. Scale and Affine invariant interest point detectors. *IJCV* 60(1):63–86, 2004.

[15] J.Matas, O. Chum, M. Urban, and T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions. *BMVC 2002*, pp. 384–393, 2002.

[16] J. Matas, S. Obdrzalek, and O. Chum. Local affine frames for wide-baseline stereo. *IEEE ICPR 2002*, Vol. 4, pp. 363–366.

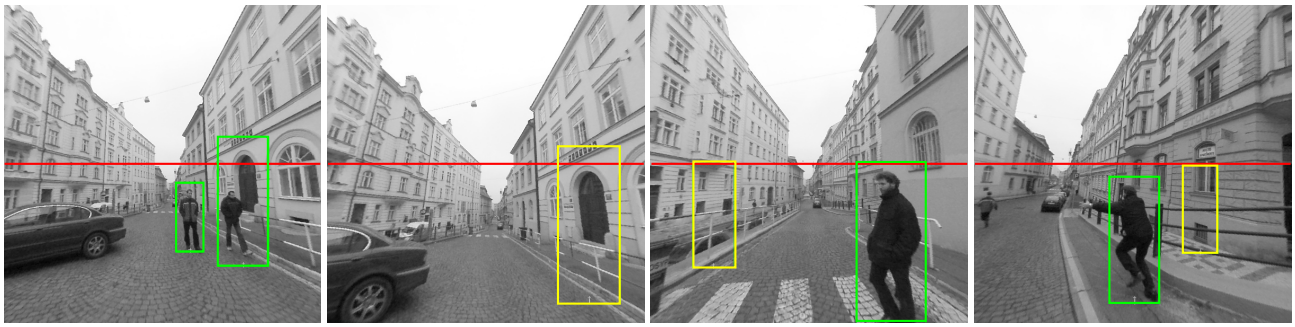
[17] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. *ECCV* pp. 404–416, 2004.

[18] B. Clipp, J-M.Frahm, M. Pollefeys, J-H. Kim, and R. Hartley. Robust 6DOF motion estimation for non-overlapping, multi-camers systems. *WACV 2008*.

[19] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd ed., 2004.

[20] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE PAMI*, 26(6):756–777, 2004.

[21] Li, H. and Hartley, R. A non-iterative method for correcting lens distortion from nine point correspondences. In *OMNIVIS 2005*.



(a)

(b)

Figure 6. Comparison of pedestrian detection. (a) Perspective cutouts using the window fixed in the center of original images. (b) Perspective cutouts with the stabilized horizon computed by using the results of the structure from motion robustly initialized by skipping small translation frames. The perspective cutouts with the stabilized horizon reduce the false detections (yellow rectangles in (a)) because the position of the ground plane feasibly constraints the pedestrian detector [10, 11].

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

[23] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. *IEEE CVPR 2005*, Vol. 1, pp. 220–226, 2005.

[24] A. Torii and T. Pajdla. Omnidirectional camera motion estimation. *VISAPP 2008*.

[25] S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. *BMVC 2002*, Vol. 1, pp. 113–122, 2002.

[26] S. Obdrzalek and J. Matas. Image retrieval using local compact dct-based representation. *DAGM 2003*, LNCS 2781, pp. 490–497, 2003.

[27] H. Stewenius. Groebner Basis Methods for Minimal Problems in Computer Vision. PhD thesis, Lund University, Sweden, 2005.

Algorithm 1 Dominant apical angle computation with robust estimation of epipolar geometry

Input Image pair I_1, I_2.

$N_V := 50$ // the number of soft votes. $N_S := 500$ // the maximum number of random samples.

$\theta := 0.3\,^\circ$ // the tolerance for establishing matches. $\sigma := 3\,^\circ$ // the standard deviation of Gaussian kernel for soft voting.

Output Dominant apical angle τ^*.

I. Detect tentative matches (MSER-INT$\pm$, MSER-SAT$\pm$, APTS-LAP, and APTS-HES) and compute their descriptors (LAF+DCT) [25, 26].

II. Construct the list $M = [\mathbf{m}]_1^N$ of tentative matches with mutually closest descriptors [23]. Order the list ascendingly by the distance of the descriptors. N is the length of the list.

III. Find a camera motion consistent with a large number of tentative matches [24]:

Set D to zero. // Initialize the accumulator of camera translation directions.
for $i := 1, \ldots, N_V$ **do**
  $t := 0$ // The counter of samples.
  while $t \leq N_T$ **do**
    $t := t + 1$ // New sample.
    Select the 5 tentative matches M_5 of the t^{th} sample from the ordered list M [23]
    $\mathtt{E}_t :=$ the essential matrix by solving the 5-point minimal problem for M_5 [20, 27].
    if M_5 can be reconstructed in front of cameras [19, 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 constraint [19, p. 119].
    $N_T := \min(N_T, N_R)$ // Update the termination length.
  end while
  $\hat{t} = \arg_{t=1,\ldots,N_T} \max S_t$ // 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 σ and the 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.
$M^* := [\mathbf{m}^*]_1^{N^*}$ // The inlier matches supporting $\mathtt{E}^*$. N^* is the number of the inlier matches.

IV. Find the dominant apical angle based on the "best" camera motion $\mathtt{E}^*$.

Decompose $\mathtt{E}^*$ into the rotation $\mathtt{R}$ and the translation $\mathbf{t}$ [19, p. 260].
for $i := 1, \ldots, N^*$ **do**
  Compute the apical angle τ_i from the match $\mathbf{m}_i^*$, $\mathtt{R}$ and $\mathbf{t}$ (see Section 2).
end for
Compute the 5^{th} percentile q^{05} and the 95^{th} percentile q^{95} from $[\tau]_1^{N^*}$. // Lower and upper bounds on apical angles to exclude outliers.
for $i := 1, \ldots, N^*$ **do**
  if $q^{05} < \tau_i < q^{95}$ **then**
    Vote in accumulator B by the Gaussian with sigma σ and the mean at τ_i.
  end if
end for
$\tau^* := \arg_{y \in domain(B)} \max B(y)$ // Maximum in the accumulator.

V. Return τ^*.