2017 14th Conference on Computer and Robot Vision

Estimating Camera Tilt from Motion without Tracking

Nada Elassal
Center for Vision Research
York University
Toronto, Canada
nada.elassal@gmail.com

James H. Elder
Center for Vision Research
York University
Toronto, Canada
jelder@yorku.ca

***Abstract*—Most methods for automatic estimation of external camera parameters (e.g., tilt angle) from deployed cameras are based on vanishing points. This requires that specific static scene features, e.g., sets of parallel lines, be present and reliably detected, and this is not always possible. An alternative is to use properties of the motion field computed over multiple frames. However, methods reported to date make strong assumptions about the nature of objects and motions in the scene, and often depend on feature tracking, which can be computationally intensive and unreliable. In this paper, we propose a novel motion-based approach for recovering camera tilt that does not require tracking. Our method assumes that motion statistics in the scene are stationary over the ground plane, so that statistical variation in image speed with vertical position in the image can be attributed to projection. The tilt angle is then estimated iteratively by nulling the variance in rectified speed explained by the vertical image coordinate. The method does not require tracking or learning and can therefore be applied without modification to diverse scene conditions. The algorithm is evaluated on four diverse datasets and found to outperform three alternative state-of-the-art methods.**

***Keywords*-Camera Calibration; Camera Pose; Image Rectification; Optic Flow; Surveillance**

I. Introduction

A major challenge for automatic video analytics is the geometric distortion induced by projection to the image that complicates almost all tasks, including object detection, velocity estimation and crowd analysis. For many systems, camera roll is small and focal length is known or can be estimated, so that the critical remaining unknown is camera tilt (Fig. 2). If camera tilt can be estimated, image and video observations can be corrected for the effects of projection, allowing unbiased analysis (Fig. 3). This correction is critical for traffic analytics [1] and crowd counting [2], for example.

II. Prior Work

Methods for auto-calibrating external camera parameters typically rely upon static features such as families of straight parallel lines, curves [1] or orthogonal structure [3]–[7] in the scene from which vanishing points can be computed. However, in many situations these static features are not present, are confounded by irregularities or are not easily detected due to occlusions and shadows (Fig. 1(a)).

An alternative or complementary approach is to use motion information from the active agents such as pedestrians and vehicles in the scene. Kuo *et al* [8] estimated camera parameters from sequences of key-point features projecting from the main joints of a walking human. However, this approach has only been demonstrated using motion-capture data; reliable automatic detection of these key points from surveillance video in crowded scenes would be challenging.

More relevant to surveillance applications are calibration methods that do not depend upon detailed recovery of object structure. Lv *et al* [9] developed an approach based on tracking the head and foot locations of a pedestrian. Bose *et al* [10] reported a more general approach based on tracking the centroids of multiple objects (pedestrians or cars). Zhang *et al* [11] estimated two horizontal vanishing points from the principal axes of segmented moving vehicles and a vertical vanishing point from the orientation of segmented pedestrians in the scene.

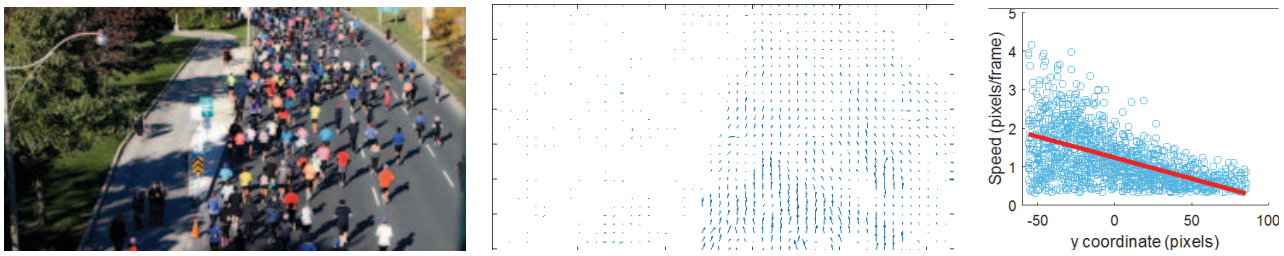
While not dependent upon detailed object structure, these methods do require accurate object segmentation. Dubska *et al* [12] have recently reported a motion-based method that relaxes this requirement. Local feature points on vehicles are tracked to obtain straight motion trajectories that can be used to estimate one ground plane vanishing point.

While not requiring segmentation, this method still requires accurate tracking of features over time, and assumes that the tracked objects are moving at constant speed along straight trajectories in the scene. In this paper, we present a novel method for recovering camera pose, specifically tilt, that does not require tracking and does not depend upon constant speed or straight trajectories.

The proposed method is anchored on a relatively general assumption: we assume zero correlation of object speed in the scene (not the image) with position on the ground plane (in scene coordinates). Importantly, the method makes no assumptions about the directions of motion or the distribution of speeds. Advantages of the proposed approach include:

1) No dependence on the visibility of regular static structures in the scene.
2) No requirements that moving objects be segmented.
3) No dependence on object shape analysis.
4) No requirement that objects or features be tracked over

978-1-5386-2818-8/17 $31.00 © 2017 IEEE
DOI 10.1109/CRV.2017.36



(a) (b) (c)

Figure 1: a) Sample frame from marathon dataset, b) Corresponding optical flow field, c) Scatter plot of speed (magnitude of optical flow vectors) vs y coordinate. (y increases upward.)

time.

5) No assumption about the direction(s) of motion.
6) No assumption that objects move in the same direction.
7) No assumption that individual motions are linear or constant speed.
8) No dependence on learned parameters, meaning that the approach can be applied to a broad range of situations without retraining.

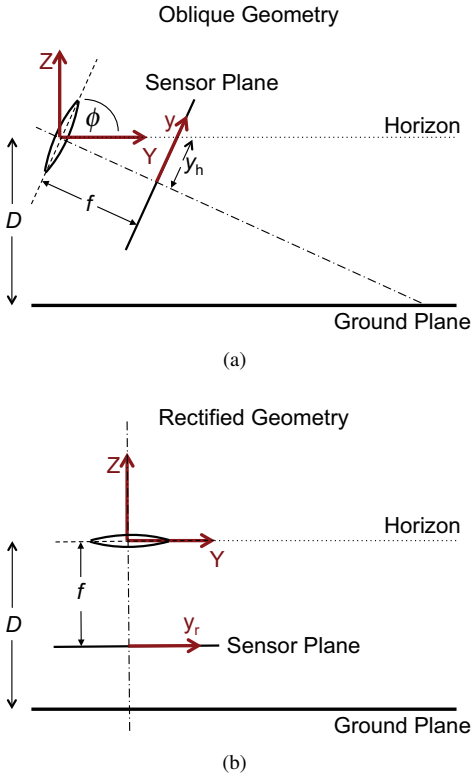
Fig. 1 illustrates the idea behind the approach. Despite stationary motion statistics over the right portion of the ground plane (a), the oblique angle of the camera induces a projective distortion on the optic flow (b), resulting in a decline in image speed with height in the image. This statistical relationship can be captured with a simple affine model (c). The strength of this affine relationship generally increases with the camera tilt angle ϕ relative to the ground surface normal (Fig. 2a).

Given an estimate $\hat{\phi}$ of the tilt angle, the optic flow field can be re-rendered in rectified coordinates (Fig. 2b) and this should result in a reduced correlation between image speed and height in the image. Thus the tilt angle can be estimated by gradient descent on the variance in the rectified optic flow explained by the affine model.

We stress that this algorithm makes no assumption about the azimuthal angle of the camera relative to the motion in the scene. To verify this, we evaluate performance for a range of scenarios (Fig. 3): while in our highway and marathon datasets image motion is primarily along the y-axis of the camera frame, in our outdoor pedestrian dataset the motion directions are diverse, and in our indoor pedestrian dataset the dominant motion is roughly 20 deg from the x-axis, i.e., much closer to the x-axis than the y-axis.

III. Geometry

We assume that camera focal length is known and that imagery has been pre-processed to square the pixels and zero the skew. (These parameters are easily measured in the lab.) We also assume negligible camera roll, which is reasonable for many installations.[1] For notational simplicity we locate the centre of the image coordinate system at the principal point.



(a)

(b)

Figure 2: Camera geometry. The world coordinate frame is centred at the camera and aligned with gravity. Both the X-axis of the world frame and x-axis of the image frame point toward the reader. Tilt angle ϕ (a) is estimated by minimizing the correlation of rectified speed with the rectified y-coordinate y_r (b).

[1] In principal, our method could be generalized to estimate camera roll by searching for the direction in the image that maximizes the correlation with image speed, but we have not yet explored this possibility.

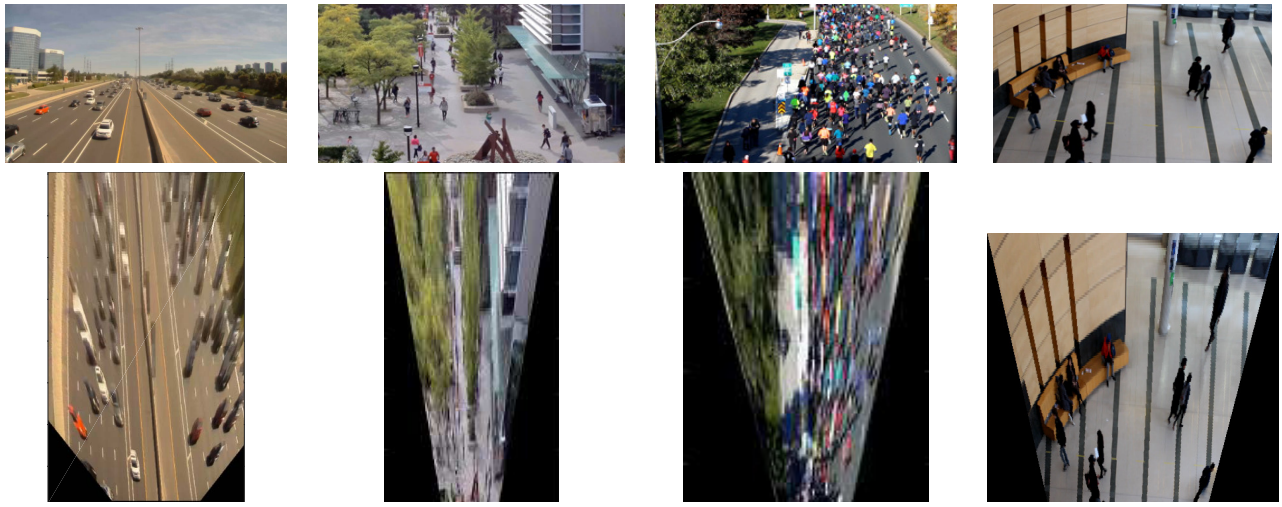


Figure 3: Sample frames and example rectifications computed using our trackless method for our highway, outdoor pedestrian, marathon and indoor pedestrian datasets. Note that rectifications induce distortions for objects not lying flat on the ground plane. These rectifications are shown only to verify the correctness of the tilt estimates: parallel lines on the ground plane should appear parallel in the rectified images.

We assume a planar horizontal ground surface[2] and adopt a right-hand world coordinate system $[X, Y, Z]$ centred at the camera, where the Z-axis is in the upward normal direction (Fig. 2). Without loss of generality, we align the x-axis of the image coordinate system with the X axis of the world coordinate system (both out of the page in Fig. 2).

Under these conditions, a point $[X, Y]^T$ on the ground plane projects to a point $[x, y]^T$ on the image plane according to

$$\lambda[x, y, 1]^T = H[X, Y, 1]^T, \tag{1}$$

where λ is a scaling factor and the homography H is given by ([13], Page 328, Eqn. 15.16):

$$H = \begin{bmatrix} f & 0 & 0 \\ 0 & f\cos\phi & -fD\sin\phi \\ 0 & \sin\phi & D\cos\phi \end{bmatrix} \tag{2}$$

Here f is the focal length in pixels, D is the height of the camera and ϕ is the tilt angle of the camera relative to the ground plane: $\phi = 0$ when the camera points straight down at the ground surface and increases to $\pi/2$ as the camera tilts up toward the horizon.

Conversely, points in the image can be backprojected to the ground plane using the inverse of this homography, $[X, Y, 1]^T = \lambda H^{-1}[x, y, 1]^T$, where

$$H^{-1} = (fD\cos 2\phi)^{-1} \begin{bmatrix} D & 0 & 0 \\ 0 & D\cos\phi & fD\sin\phi \\ 0 & -\sin\phi & f\cos\phi \end{bmatrix} \tag{3}$$

[2]If the surface is planar but not horizontal, our method will estimate the tilt angle relative to the ground but this will of course be offset relative to gravity.

In Euclidean coordinates this backprojection can be written as:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \frac{D}{f\cos\phi - y\sin\phi} \begin{bmatrix} x \\ y\cos\phi + f\sin\phi \end{bmatrix} \tag{4}$$

As a final step, we can apply the homography H of Eqn. (2) with a tilt angle of $\phi = 0$ to the scene points $[X, Y]^T$ computed using (4), transferring these scene points to image points $[x_r, y_r]^T$ taken by a "bird's eye" virtual camera (Fig. 2b), yielding a rectified plan view of the ground surface seen from a height D:

$$\begin{bmatrix} x_r \\ y_r \end{bmatrix} = \frac{f}{f\cos\phi - y\sin\phi} \begin{bmatrix} x \\ y\cos\phi + f\sin\phi \end{bmatrix} \tag{5}$$

Taking the time derivative, we can compute the rectified optic flow field:

$$\mathbf{v}_r = \frac{f}{(f\cos\phi - y\sin\phi)^2} \begin{bmatrix} fx'\cos\phi + (xy' - x'y)\sin\phi \\ fy' \end{bmatrix} \tag{6}$$

This can also be expressed as:

$$\mathbf{v}_r = \begin{bmatrix} x'_r \\ y'_r \end{bmatrix} = \frac{f}{(y_h - y)^2\sin\phi} \begin{bmatrix} x'y_h + xy' - x'y \\ fy'/\sin\phi \end{bmatrix} \tag{7}$$

where $y_h = f\cot\phi$ is the image projection of the horizon (Fig. 2a).

Our key assumption is that the rectified speed $v_r = |\mathbf{v}_r|$, when averaged over rectified image location (x_r, y_r) and time, is invariant with the vertical image coordinate. Thus an estimate of the tilt angle ϕ can be evaluated by measuring the correlation of $v_r\,(x_r, y_r|\phi)$ with y_r.

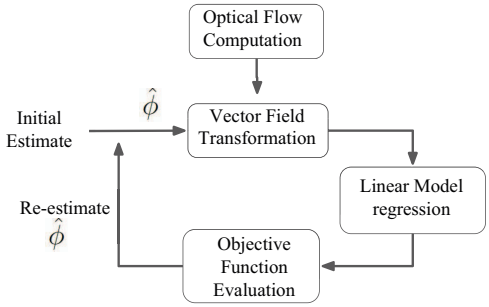


Figure 4: Algorithm overview

IV. Algorithm

Figure 4 provides an overview of the algorithm.

A. Objective Function

Given an estimate $\hat{\phi}$ of the tilt angle, we can compute the rectified speeds $v_r\left(x_r, y_r|\hat{\phi}\right)$ using Eqn. 7. To assess correlation with y_r we use the simple affine model $\hat{v}_r = ay_r + b$ and determine maximum likelihood estimates of the parameters (a, b) by linear regression. The strength of this correlation is measured by the proportion of variance $R^2\left(\hat{\phi}\right)$ explained by the model, given the estimated tilt $\hat{\phi}$:

$$R^2\left(\hat{\phi}\right) = 1 - \frac{\mathbb{E}\left[\left(v_r - \hat{v}_r\right)^2\right]}{\mathbb{E}\left[\left(v_r - \bar{v}_r\right)^2\right]} \tag{8}$$

where $\bar{v}_r$ is the average rectified speed over the rectified image and some interval of time. We seek the tilt angle ϕ^* that minimizes $R^2\left(\hat{\phi}\right)$.

B. Optimization

We estimate ϕ^* by iterative minimization of Eqn. 8 using MATLAB's **fminsearch** (Nelder-Mead simplex method). In our experiments we repeat the search from a coarse regular sampling of initial estimates $0 \leq \hat{\phi} \leq \pi/2$, selecting the ϕ^* that yields the minimum R^2. However in practice we find that given sufficient input data (> 100 frames) the error function becomes convex and a single search initiated at $\hat{\phi} = \pi/4$ suffices.

C. Optical Flow Computation

We employ the optical flow algorithm of Xu et al [14]: Fig. 1(b) shows an example for the marathon dataset. Since we are only concerned with motion on the ground plane, motion vectors above our current estimate of the horizon $\hat{y}_h = f \cot \hat{\phi}$ are ignored.

Ground plane motion will generally be sparse and spatially interleaved with noise due to small environmental motions, camera vibration etc. that does not correlate with y_r and thus could reduce accuracy. This problem can be mitigated by filtering out all but the largest $p\%$ of motion vectors from each image frame: Fig. 6 shows an example for the marathon data set. However, this leaves the problem of estimating the optimal threshold p.

To avoid supervised learning of this parameter, which may not generalize to novel scenarios, we employ an adaptive method to select p individually for each video sequence. In particular, we select the threshold p ($0\% < p \leq 100\%$) that

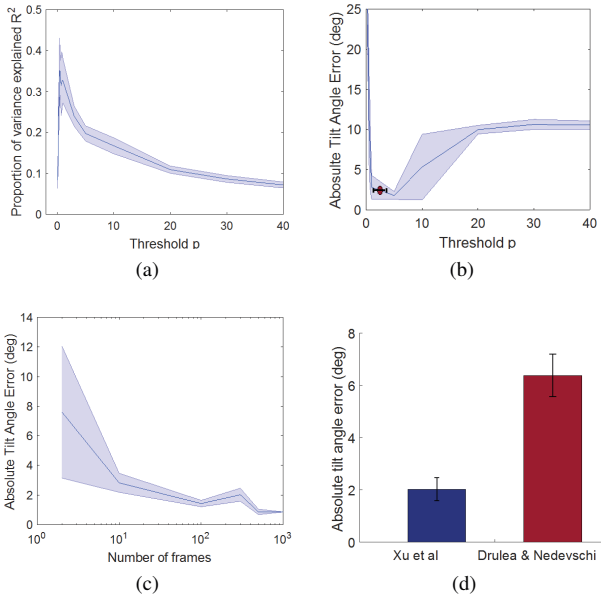


Figure 5: Analysis of algorithm dependence on key variables, for the highway dataset. Shading and error bars indicate standard error of the mean. (a-c) show results using the optic flow algorithm of Xu et al [14]. (a), (b) and (d) show results for 300-frame sequences. a) Average proportion of variance R^2 explained by the affine model as a function of the optic flow threshold p. b) Average tilt error of trackless algorithm as a function of the optic flow threshold p. The red dot indicates the threshold p chosen automatically by the algorithm. c) Average tilt error of trackless algorithm as a function of the number of video frames analyzed. d) Average tilt error of trackless algorithm based on the optical flow algorithms of Xu et al [14] and Drulea & Nedevschi [15].

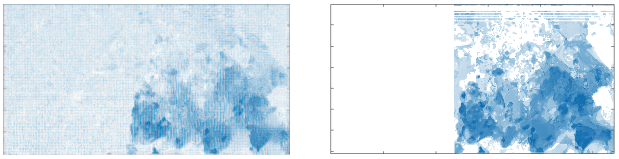


Figure 6: Example optical flow vector field from the marathon dataset, before (a) and after (b) noise removal. The automatically selected speed threshold is p=28%.

maximizes $R^2(0)$, the variance in unrectified image speed $v\,(x, y)$ explained by correlation with the vertical image coordinate y (Fig. 1(c)). Fig. 5(a) shows that for our highway dataset, the proportion of variance explained peaks when a relatively small fraction (2%) of the motion vectors are employed and Fig. 5(b) shows that this threshold yields nearly minimal error in the resulting tilt angle estimate. Thus by selecting the threshold that maximizes variance explained in the unrectified image, we adaptively optimize the accuracy of the algorithm.

We have also evaluated dependence on the optic flow method employed (Fig 5(d)), comparing the methods of Xu et al [14] and Drulea & Nedevschi [15], both highly ranked on the Middlebury dataset [16]. While both work reasonably well, we find the algorithm of Xu et al [14] more accurate for this application and dataset.

Fig. 5(c) shows how accuracy varies as a function of integration time. For the highway dataset, performance is very good for durations of 50 frames (1.7 sec) or more.

V. Evaluation

A. Implementation

We employed the optical flow method of Xu et al [14] (code downloaded from *www.cse.cuhk.edu.hk/leojia/projects/flow*), with parameters matching those used by the authors for evaluation on the Middlebury Benchmark: regularization strength: 6, occlusion handling: 1 and large motion: 0. The average run time for optical flow computation is 45 sec per frame. We implemented our algorithm in MATLAB and have not yet optimized the code for speed (run times listed below). All experiments were conducted on a 4-core desktop computer. Our code and datasets will be released for public use.

B. Datasets

We evaluate our proposed method on 4 diverse datasets (Fig. 3) recorded with 3 different camera/lens systems to assess the generality of the approach: 1) a highway scene where the moving agents are vehicles, 2) an outdoor campus scene where the moving agents are pedestrians, 3) an urban marathon scene where the moving agents are runners and 4) an indoor scene where the moving agents are pedestrians.

The highway and outdoor pedestrian datasets were recorded with a Point Grey Cricket camera equipped with a 16 mm lens. The marathon and indoor pedestrian datasets were recorded with a Canon EOS Rebel T3i camera equipped with a 40 mm lens. All camera/lens systems were calibrated in the lab using a standard calibration procedure to determine focal length f (Table I) and principal point. Frame rate was 30 fps for all datasets and each was partitioned into 5 clips of 300 frames each.

For the highway dataset, ground truth camera tilt angle was estimated manually from the horizon image height y_h using the relation $y_h = f \cot \phi$ (Eqn. 4). For the other three datasets, ground truth tilt angle was measured directly using a digital inclinometer.

C. Evaluation & Comparison with Static Methods

Table I shows quantitative performance of the proposed method on our four datasets. The mean threshold parameter p ranged from 2% to 28% over these datasets, and mean tilt error ranged from 0.46 deg to 1.48 deg.

Fig. 3 shows example rectifications based on these estimated tilt angles. Since rectification can grossly distort objects (e.g, people, cars) that do not lie flat on the ground plane, it is unlikely that the rectified video would be directly useful for visual analytics. However, the rectified imagery *is* useful for visual verification: if tilt estimates are accurate, lines that are parallel on the ground plane should appear parallel in the rectified view. These estimated tilt angles can then be used by downstream algorithms to convert measured image quantities (e.g., image speed in pixels/sec) to scene quantities (e.g., ground plane speed in m/s).

To compare our method to prior approaches, we were able to secure code directly from the authors of three prior methods that use static features (lines or curves) to estimate vanishing points and tilt angle [1], [5], [17]. All methods were provided with focal length and principal point, and were required only to estimate tilt angle.

We note that the availability of regular static structure varies widely in our datasets, and so we expect that the performance of these methods will vary widely as well. In the highway dataset there is one strong family of horizontal parallel lines but other structure is weak. For the outdoor pedestrian dataset there is moderate structure in three orthogonal directions, but it is not dominant. In the marathon scene the main family of parallel horizontal lines is largely occluded by the runners. Finally, in the indoor pedestrian dataset there is one strong family or horizontal lines and some significant vertical structure. On the other hand, all scenes contain significant ground plane motions, but of varying kinds and in various directions.

Fig. 7 shows the resulting performance of these prior static methods alongside ours. For the highway dataset, mean absolute error for our method was 1.06 deg, much better than the methods of Tal & Elder and Wildenauer & Hanbury (errors of 4.73 deg and 55.28 deg, respectively), which expect regular static structure in more than one orthogonal direction. However performance of our motion method was not quite as good as the method of Corral-Soto & Elder (0.55 deg), which we note was designed specifically for highway applications and depends on only a single family of parallel lane boundaries to estimate tilt angle.

For the outdoor pedestrian dataset, mean absolute error for our method was only 0.46 deg. We note that our method is highly accurate here despite the substantial variations in the directions and speeds of motion of the pedestrians in this video, highlighting the fact that our method does not require

Table I: Parameters and experiment results of the proposed method on four datasets. Run times are exclusive of the time required to compute the optic flow map.

Dataset	Image size (pixels)	Focal length (pixels)	True tilt angle (deg)	p (%)	Mean error (deg)	Run time per frame (msec)
Highway	275 x 155	174	87.8	2	1.06	106
Outdoor Pedestrian	320 x 165	953	81.0	10	0.46	383
Marathon	324 x 156	700	76.4	28	1.48	416
Indoor	320x182	584	60.7	5	0.68	350

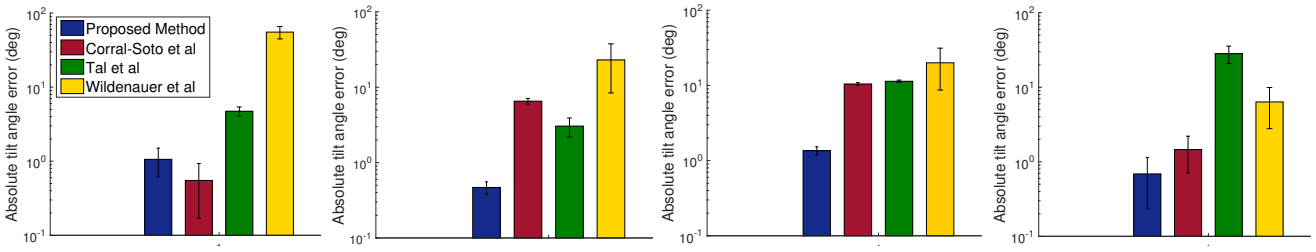


Figure 7: Performance of the proposed method compared with three state-of-the-art static methods from Corral-Soto & Elder [1], Tal & Elder [5] and Wildenauer & Hanbury [17] on (a) the highway dataset, (b) the outdoor pedestrian dataset, (c) the marathon dataset and d) the indoor dataset.

that motions of individual agents be similar. The large errors produced by the prior static methods we evaluated (6.53, 3.05, 23.41 deg) presumably reflect the relative sparseness of static regularities in the scene.

For the marathon dataset, mean absolute error for our method was 1.48 deg, substantially better than competing static methods (errors of 10.45, 11.35 and 19.99 deg). Again, this superiority reflects the relatively strong motion signals and the fact that the static cues are largely occluded by the runners.

For the indoor pedestrian dataset, mean absolute error for our method was only 0.68 deg. Note that in this dataset the dominant direction of motion is roughly 20 degrees counterclockwise from the x-axis of the image. The excellent performance of the proposed algorithm for this example illustrates the invariance of the method to the dominant direction of motion. The method of Corral-Soto & Elder [1] also performs relatively well here (1.45 deg), even though it was originally designed for highway applications. This is presumably due to the strong linear structure of the floor tiles, which are qualitatively similar to traffic lanes. Unfortunately the other two static methods [5], [17] are much less accurate (28.22 and 6.35 deg), despite the fairly strong presence of linear structure in one horizontal and one vertical direction.

In summary, we find that our proposed trackless motion-based method performs quite consistently over a range of scenes and motion distributions. The static methods we have tried, on the other hand, are quite sensitive to the nature and density of regular linear structure. This is not to say that these static methods are not useful. The method of Corral-Soto & Elder [1] should work well for scenes with at least one clearly visible family of parallel curves in the ground plane (as seen for the highway and indoor pedestrian datasets) and the methods of Tal & Elder and Wildenauer & Hanbury [5], [17] should work well for scenes with clearly visible linear structure in three orthogonal directions. However our results do suggest that when this static structure is less visible, either due to the nature of the scene or occlusion by moving agents such as vehicles or pedestrians, the proposed trackless motion-based method for tilt estimation may be more reliable, and thus may complement these static methods.

D. Comparison with Motion-Based Methods

It would be ideal to compare our proposed method with the prior motion-based methods from Dubska *et al* [12] and Zhang *et al* [11] directly on a common dataset as well. Unfortunately, despite contacting authors we were unable to obtain code or datasets used for either method. The best we can do at this stage is to compare performance of these prior algorithms on their proprietary datasets, as reported by the authors, with the performance of our algorithm on our own dataset. (We caution that since there could be systematic differences in the difficulty of the datasets, this comparison should not be used to formally rank the algorithms.)

Comparison with these prior motion-based methods is further complicated by the fact that in this prior work the

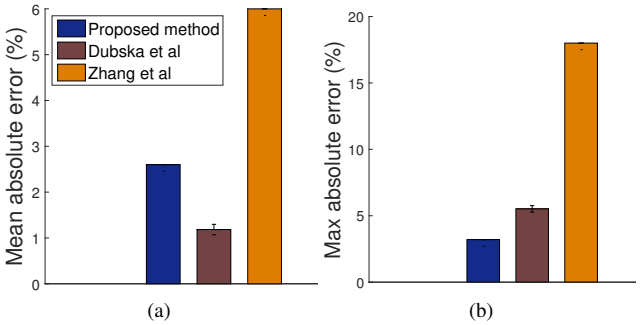


Figure 8: Performance of the proposed method on the highway dataset compared with two state-of-the-art motion-based methods from Dubska *et al* [12] and Zhang *et al* [11] **on different datasets**. a) Mean absolute deviation of point pairs (%), b) Maximum absolute deviation of point pairs (%).

authors did not have ground truth tilt estimates. Instead, they manually identified point pairs in the image known to lie in a horizontal plane and to be equidistant in the scene, and then computed the average absolute percentage deviation of the distances in the rectified imagery from their mean: a more accurate homography estimate should lead to lower average deviation.

Since equidistant horizontal point pairs are not easily identifiable in our pedestrian and marathon datasets, we restricted our attention to the highway dataset. We identified 8 point pairs, each pair projecting from fixed points at the same height on the same vehicle, over 10 consecutive frames. We then projected these points to the rectified image using our estimated homography matrix and measured the mean absolute deviation of their separation over the 10 frames, as in Dubska *et al* [12].

Fig. 8 shows mean absolute deviation of distance between point pairs in rectified imagery for our algorithm on the highway dataset, compared with the errors reported by Dubska *et al* [12] and Zhang *et al* [11] on their respective datasets. We find that by this measure our method has a mean error of 2.6%, lying between the performance reported by Dubska *et al* [12] (1.18%) and that reported by Zhang *et al* [11] (6%) on their respective datasets. However, it appears that our method may be more reliable, as it has a maximum error rate of only 3.2%, compared to 5.5% for the method of Dubska *et al* [12] and 18% for the method of Zhang *et al* [11]. We emphasize that our method also does not require explicit tracking or vanishing point estimation, as required by these prior methods, and thus has potentially lower computational requirements and greater generality. However, we must emphasize that no definitive statements can be made until these methods are compared on a common dataset.

E. Discussion

The proposed trackless motion-based method for camera tilt estimation was found to work reliably over four very different camera/lens systems, scenes, active agents and patterns of motion, with average absolute tilt errors ranging from 0.46 to 1.48 deg. For the highway dataset the static method of Corral-Soto & Elder [1] based on curve parallelism was found to be slightly more accurate. This dataset represents an ideal scenario for the method of Corral-Soto & Elder, where the family of parallel lane markings are clearly visible. Our motion-based method performs almost as well (0.55 deg vs 1.06 deg error) and much better than the other two state-of-the-art static methods (4.73 deg and 55.28 deg error) we assessed [5], [17] .

For the other three datasets (outdoor pedestrian, marathon and indoor) where parallel families of lines and curves are either not as clearly visible or occluded by moving agents, our proposed method outperformed all three methods based on static structural features [1], [5], [17] by a large margin, highlighting the relative generality of our approach.

Unfortunately direct comparison of the proposed trackless motion-based method against prior motion-based methods from Dubska *et al* [12] and Zhang *et al* [11] was not possible due to lack of a common dataset and/or shared code. However, informal comparison on different but similar datasets (Fig. 8) suggests that for traffic data, our approach may be comparable to the method of Dubska *et al* (higher mean error, lower max error) and much better than the method of Zhang *et al*. We also note that these two competing motion-based approaches depend upon explicit appearance modeling, feature tracking and vanishing point estimation, and have been tailored specifically to traffic applications. Our approach, on the other hand, works for general dynamic scenes on a ground plane and does not require explicit tracking or vanish point estimation.

Just as static methods do poorly when regular static features are sparse, motion-based methods such as the trackless method proposed here will do less well when motion is sparse. Characterizing exactly how performance varies with sparsity of motion remains a topic for future work, however note that our automatic method for denoising the optic flow field selects as little as 2% of optic flow vectors with good results, suggesting that dense motion is not necessarily required. Furthermore, in a motion-based method, the quality of the estimate can be improved continuously over time as additional independent motion vectors are observed, something that is not possible for static methods. However, since there will always be some scenes where static methods work best and others where motion-based methods work best, the ultimate system would likely employ both approaches and arbitrate between them over time.

VI. Conclusion and Future Work

In this paper we have presented a novel and very general method for recovering camera tilt from image motion in an unsupervised manner. Unlike prior methods, the proposed algorithm does not depend upon the visibility of regular static structures in the scene and does not require segmentation, shape analysis or feature tracking, thus reducing the required computation. Our method does not require that objects move in the same direction or at constant velocities. Rather, it rests on the much more general assumption of zero correlation of ground plane speed with ground plane position, averaged over time. A novel method for automatically and adaptively selecting the optimal subset of motion vectors generated by the objects moving in the scene means that the algorithm does not require training. This allows the algorithm to be applied to diverse scenarios without reconfiguration. We have demonstrated this generality on four diverse datasets recorded with different camera/lens systems and have found the method to consistently perform well relative to competing state-of-the-art methods based on motion features and on regularities of static structures.

In future work, we intend to characterize the accuracy of the method as a function of motion density and to extend the method to recover roll angle by searching for the direction in the image that maximizes the correlation with image speed. Another goal is to develop compatible methods for estimating focal length, important for camera installations employing zoom lenses.

References

[1] E. R. Corral-Soto and J. H. Elder, "Automatic single-view calibration and rectification from parallel planar curves," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 813–827.

[2] N. Elasal and J. H. Elder, "Unsupervised crowd counting," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2016.

[3] M. Hodlmoser, B. Micusik, and M. Kampel, "Camera auto-calibration using pedestrians and zebra-crossings," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1697–1704.

[4] N. K. Kanhere and S. T. Birchfield, "A taxonomy and analysis of camera calibration methods for traffic monitoring applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 441–452, 2010.

[5] R. Tal and J. H. Elder, "An accurate method for line detection and manhattan frame estimation," in *Proceedings of the Asian Conference on Computer Vision Workshops (ACCV Workshops)*. Springer, 2012, pp. 580–593.

[6] E. Tretyak, O. Barinova, P. Kohli, and V. Lempitsky, "Geometric image parsing in man-made environments," *International Journal of Computer Vision*, vol. 97, no. 3, pp. 305–321, 2012.

[7] J. Deutscher, M. Isard, and J. MacCormick, "Automatic camera calibration from a single manhattan image," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2002, pp. 175–188.

[8] P. Kuo, J.-C. Nebel, and D. Makris, "Camera auto-calibration from articulated motion," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2007, pp. 135–140.

[9] F. Lv, T. Zhao, and R. Nevatia, "Self-calibration of a camera from video of a walking human," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, vol. 1, 2002, pp. 562–567.

[10] B. Bose and E. Grimson, "Ground plane rectification by tracking moving objects," in *Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (PETS)*, 2003, pp. 94–101.

[11] Z. Zhang, M. Li, K. Huang, and T. Tan, "Practical camera auto-calibration based on object appearance and motion for traffic scene visual surveillance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[12] M. Dubska, A. Herout, R. Juranek, and J. Sochor, "Fully automatic roadside camera calibration for traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1162–1171, 2015.

[13] S. Prince, *Computer Vision: Models, Learning and Inference*. Cambridge, UK: Cambridge University Press, 2012.

[14] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1293–1300.

[15] M. Drulea and S. Nedevschi, "Motion estimation using the correlation transform," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3260–3270, 2013.

[16] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[17] H. Wildenauer and A. Hanbury, "Robust camera self-calibration from monocular images of Manhattan worlds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2831–2838.