IROS Workshop on Visual Control of Mobile Robots (ViCoMoR 2012)
October 11th, 2012, Vilamoura, Algarve, Portugal

Homography based visual odometry with known vertical direction and weak Manhattan world assumption

Olivier Saurer, Friedrich Fraundorfer, Marc Pollefeys
Computer Vision and Geometry Lab, ETH Zürich, Switzerland

***Abstract*— In this paper we present a novel visual odometry pipeline, that exploits the weak Manhattan world assumption and known vertical direction. A novel 2pt and 2.5pt method for computing the essential matrix under the Manhattan world assumption and known vertical direction is presented that improves the efficiency of relative motion estimation in the visual odometry pipeline. Similarly an efficient 2pt algorithm for absolute camera pose estimation from 3D-2D correspondences is presented that speeds up the visual odometry pipeline as well. We show that the weak Manhattan world assumption and known vertical allow for direct relative scale estimation, without recovering the 3D structure. We evaluate our algorithms on synthetic data and show their application on real data sets from camera phones and robotic micro aerial vehicles. Our experiments show that the weak Manhattan world assumption holds for many real-world scenarios.**

I. Introduction

The Manhattan world assumption is a very strong restriction to a general 3D scene. And yet this assumption is fulfilled for many scenes that contain man-made architectural structures, at least partially. The assumption especially holds true for indoor environments, and also for urban canyons of modern cities. This was successfully demonstrated and exploited in a variety of recent papers [1], [4], [5]. In this work we will refer to the weak Manhattan world, describing a world consisting of vertical planes which are arbitrary oriented around the vertical direction. They are not required to be orthogonal to each other. The only restriction is, that vertical planes are parallel to the gravity vector and the ground planes are orthogonal to the vertical direction.

Especially visual odometry [16] can benefit at a high degree from the Manhattan world assumption. Visual odometry is the means of ego motion estimation of e.g. mobile robots fitted with cameras. One computational bottleneck of visual odometry is the robust motion estimation using RANSAC [2]. The computational complexity of RANSAC depends exponentially on the number of data points needed for hypothesis generation, for unconstrained motion in 3D this would mean 5 data points (feature correspondences) using the 5pt essential matrix algorithm [14]. Constraints on the robot motion (e.g. planar motion), on the environment or using additional sensor data however can reduce the number of necessary data points and make RANSAC more efficient, which is important to achieve real time performance. For the case of a planar motion assumption, which is true for many mobile robot applications, two point correspondences are sufficient to compute an egomotion hypothesis for RANSAC [15].

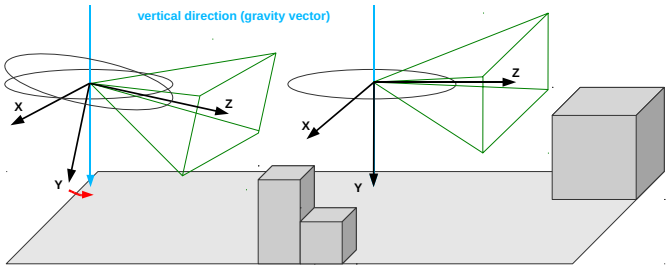

Fig. 1. Knowing the vertical direction, e.g. by measuring the gravity vector with an IMU or from vanishing points, the image can be rotated such that the y-axis of the camera matches the vertical direction. Under the weak Manhattan world assumption this aligns the x-z-plane of the camera with the ground plane and the y-axis with the vertical direction of walls.

In this work we will present two novel relative pose algorithms and a novel absolute pose algorithm which exploits the weak Manhattan world assumption and additionally takes advantage of the knowledge of the vertical direction of the scene structure in the images. The known vertical direction and the assumptions about the environment lead to a simpler formulation for relative 5DOF camera motion, in particular to a 2pt algorithm and a 2.5pt algorithm, in contrast to the standard 5pt method. For successive 6DOF camera pose estimation from 3D-2D matches we propose a new 2pt method, exploiting the known vertical direction.

The vertical direction can be computed from image features, but also from an inertial measurement unit (IMU) (which e.g. measures the earth's gravity vector) attached to a camera. This is the case for almost every state-of-the-art smart phone (e.g. IPhone, Google Nexus S, Nokia N900..) which are all equipped with a camera and an IMU. We conduct synthetic experiments to evaluate the proposed algorithms under different image noise and IMU measurement noise and compare the results to the standard 5-pt relative pose and the 3-pt relative pose with known vertical direction algorithm. We further demonstrate the proposed algorithms on real data from camera phones (which come with IMU sensors) and show visual odometry results for a robotic micro aerial vehicle. The experiments show that the weak Manhattan world assumption holds and can be exploited in real-world scenarios.

II. Related work

Early results on coupling inertial sensors (IMU) and cameras and using the measured gravity normal for ego-motion

estimation have been discussed in [17] or [11]. Most closely related to our paper are the works of [3], [7]. In [3] the authors present an algorithm to compute the essential matrix from 3pt correspondences and a known gravity vector, e.g. from an IMU measurement. This 3pt formulation can speed up RANSAC significantly compared to the standard 5pt algorithm. In [7] vision and IMU information is combined in a similar way to find new algorithms for absolute pose estimation. More recently [8] proposed to combine IMU and vision data from monocular camera to incrementally accumulate the motion and reconstruct the camera trajectory. The incremental approach requires integration of the IMU data over time and is brittle towards IMU inaccuracy.

The Manhattan world assumption [1] has recently been picked up again and successfully been used for multi-view stereo [5], the reconstruction of building interiors [4] and also for scene reconstruction from a single image only [9]. In this paper we combine both ideas, the IMU-vision fusion and a weak Manhattan world assumption. Similar to [3], [7] but now for the case of homographies we derive a formulation for relative motion under weak Manhattan world constraints and a formulation for absolute pose. A similar idea for egomotion estimation has also been described in [12] where the authors derive a homography for vertical walls under a planar motion constraint. However, in our formulation the relative motion is not restricted to motion on a plane.

III. Relative and absolute pose estimation

Knowing the vertical direction in images will simplify the estimation of camera pose and camera motion, which are fundamental methods in any odometry pipeline. It is then possible to align every camera coordinate system with the measured vertical direction such that the y-axis of the camera is parallel to the vertical direction and the x-z-plane of the camera is orthogonal to the vertical direction (illustrated in Fig. 1). Under the Manhattan world assumption this means that the x-z-plane of the camera is now parallel to the world's ground plane and the y-axis is parallel to vertical walls. This alignment can just be done as a coordinate transform for motion estimation algorithms, but also be implemented as image warping such that feature extraction method benefit from it. Relative motion between two such aligned cameras reduces to a 3-DOF motion, which consists of 1 remaining rotation and a 2-DOF translation vector. The absolute camera pose for aligned camera has 4-DOF, again 1 remaining rotation and a 3-DOF translation vector.

The algorithms for estimating the relative pose are derived from computing a homography between a plane in two images and decomposing it. After incorporating the Manhattan world constraint, which restricts the possible planes to vertical and horizontal ones, and after incorporating the vertical direction, which decreases the DOF of the camera orientation, the parameterization of a homography is greatly simplified. This simplification leads to a 2pt and a 2.5pt algorithm for computing homographies and closed form solutions for the decomposition. The homography is represented by

$$\mathbf{H} = \mathbf{R} + \frac{1}{d}\mathbf{t}^T\mathbf{n}, \tag{1}$$

where $\mathbf{R} = \mathbf{R_y}\mathbf{R_x}\mathbf{R_z}$ is a rotation matrix representing the relative camera rotations around the x, y, and z-axis, $\mathbf{t} = [t_x, t_y, t_z]^T$ represents the relative motion, $\mathbf{n} = [n_x, n_y, n_z]^T$ is the plane normal and d is the distance from the first camera center to the plane. In all our derivations, the camera-plane distance is set to 1 and absorbed by $\mathbf{t}$. With the knowledge of the vertical direction the rotation matrix R can be simplified such that $\mathbf{R} = \mathbf{R_y}$ by pre-rotating the feature points with $\mathbf{R_x}\mathbf{R_z}$, which can be measured from the IMU or vanishing points. Under the weak Manhattan world assumption additionally the parameterization of the plane normal $\mathbf{n}$ can be simplified, to be only vertical or horizontal planes.

A. 2pt relative pose for known plane normal

The following algorithm is able to compute the relative pose given 2pt correspondences and the normal of the plane on which the points reside. The derivation will be carried out for a vertical plane but works similar for planes parameterized around other axis.

The homography for a vertical plane can be written as

$$\mathbf{H} = \mathbf{R_y} + [t_x, t_y, t_z]^T[n_x, 0, n_z] \tag{2}$$

where the normal vector is parametrized by $n_x = sin(\phi_n)$ and $n_z = cos(\phi_n)$. The homography then writes as

$$\mathbf{H} = \begin{bmatrix} \cos(\phi_y) + n_x tx & 0 & \sin(\phi_y) + n_z t_x \\ n_x t_y & 1 & n_z t_y \\ n_x t_z - \sin(\phi_y) & 0 & \cos(\phi_y) + n_z t_z \end{bmatrix} \tag{3}$$

$$= \begin{bmatrix} h_{11} & 0 & h_{13} \\ h_{21} & 1 & h_{23} \\ h_{31} & 0 & h_{33} \end{bmatrix} \tag{4}$$

To solve for the 6 entries of $\mathbf{H}$ we solve $\mathbf{x}' \times \mathbf{H}\mathbf{x} = \mathbf{0}$, where $\mathbf{x} = [x\ y\ 1]^T$ and $\mathbf{x}' = [x'\ y'\ 1]^T$ are the point correspondences. By using this relation we get rid of the unknown scaling factor of the homography. Knowing n_x and n_y leads to one additional linear constraint in the entries of the homography, $h_{23} = n_x/n_z h_{21}$.

This leaves 5 entries in $\mathbf{H}$ to be estimated. Each point correspondences gives 2 inhomogeneous linearly independent equations of the form $\mathbf{A}\mathbf{h} = \mathbf{b}$,

$$\begin{bmatrix} 0 & 0 & -x - \frac{n_x}{n_z} & xy' & y' \\ -xy' & -y' & xx' + \frac{n_x}{n_z}x' & 0 & 0 \end{bmatrix} \mathbf{h} = \begin{bmatrix} y \\ -x'y \end{bmatrix} \tag{5}$$

where $\mathbf{h} = [h_{11}\ h_{13}\ h_{21}\ h_{31}\ h_{33}]^T$.

Using 2 point correspondences this gives 4 equations which is a deficient-rank system. The solution is $\mathbf{h} = \mathbf{V}\mathbf{y} + w\mathbf{v}$ (see [6]) where $svd(\mathbf{A}) = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and $\mathbf{v}$ is the last column vector of $\mathbf{V}$. The vector $\mathbf{y}$ is computed by $y_i = b_i/d_i$ where d_i is the $i-th$ diagonal entry of $\mathbf{D}$ and $\mathbf{b}' = \mathbf{U^T}\mathbf{b}$.

This leaves the unknown scalar w which can be computed from the additional constraint, that one of the singular values of the homograph has to be one (i.e., $\det(\mathbf{H}^T\mathbf{H} - \mathbf{I}) = 0$, see [13]). By substituting $\mathbf{h} = \mathbf{V}\mathbf{y} + w\mathbf{v}$ for the entries of $\mathbf{H}$. The determinant is a 4th order polynomial in w which results in 4 solutions for $\mathbf{H}$.

If the plane normal is not known one can sample the ratio n_x/n_z. Each sample represents a hypothesis in the RANSAC loop that are then tested against the other points. Having multiple hypothesis is better than computing the orientation with one additional point sample since this has only a linear instead of an exponential impact on the RANSAC iterations. Knowledge about the approximate orientation of the wall relative to the camera will reduce the number of hypothesis, for example the case when the camera is moving along a corridor it is not always necessary to sample the full $360\,\mathrm{deg}$ for the side walls.

The such parameterized homography can easily be decomposed in the rotation and translation parameters of the relative motion. First step is a proper normalization of the up to scale homography, by dividing it by h_{22}. h_{22} needs to be 1 according to Eq. 3. Using the relation $n_x^2 + n_z^2 = 1$, t_y can be obtained from h_{21} and h_{23} by $t_y = \pm(h_{21}^2 + h_{23}^2)^{\frac{1}{2}}$ which gives two solutions for t_y which differ in the sign. The normal can then be computed as follows that $n_x = h_{21}/t_y$ and $n_z = h_{23}/t_y$. Two pairs of the normal have to be computed for the two t_y. Next we can compute $\sin(\phi_y)$ from a quadratic in the entries $h_{11}, h_{13}, h_{21}, h_{23}$ and the $\cos(\phi_y)$ is obtained from the relation $\sin(\phi_y)^2 + \cos(\phi_y)^2 = 1$. Finally t_x and t_z are obtained as $t_x = (h_{11} - \cos(\phi_y))/n_x$ and $t_z = (h_{33} - \cos(\phi_y))/n_z$.

B. 2pt relative pose for ground plane

This derivation is a special case of the previous one and will work for points on the ground plane. The normal of the ground plane is $\mathbf{n} = [0\ 1\ 0]^T$. This leads to an again simpler formulation for the homography and only 2 point correspondences are enough to compute the full homography. The homography for the ground plane can be written as: $\mathbf{H} = \mathbf{R}_\mathbf{y} + [t_x, t_y, t_z]^T[0, 1, 0]$

The entries of $\mathbf{H}$ are then

$$\mathbf{H} = \begin{bmatrix} \cos(\phi_y) & t_x & \sin(\phi_y) \\ 0 & t_y + 1 & 0 \\ -\sin(\phi_y) & t_z & \cos(\phi_y) \end{bmatrix} \quad (6)$$

$$= \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ 0 & h_{22} & 0 \\ -h_{13} & h_{32} & h_{11} \end{bmatrix} \quad (7)$$

Because of 2 linear constraints in the entries of H, H has only 5 unknowns, for which can linearly be solved using 2 point correspondences. Each point gives 2 constraints of the form $\mathbf{A}\mathbf{h} = \mathbf{0}$,

$$\begin{bmatrix} y' & 0 & -xy' & -y & yy' \\ -xy' & -yy' & -y' & x'y & 0 \end{bmatrix} \mathbf{h} = \mathbf{0} \quad (8)$$

where $\mathbf{h} = [h_{11}\ h_{12}\ h_{13}\ h_{22}\ h_{32}]^T$.

The rotation and translation parameters of the relative motion can be read off the homography matrix directly, after proper normalization. Inspection of $\mathbf{H}$ shows that the following relation $h_{11}^2 + h_{13}^2 = 1$ has to be fulfilled. The proper normalization is by dividing $\mathbf{H}$ by $(h_{11}^2 + h_{13}^2)^{\frac{1}{2}}$. The rotation matrix R and the translation vector $\mathbf{t}$ are then:

$$\mathbf{t} = [h_{12},\ h_{22} - 1,\ h_{32}]^T, \quad \mathbf{R} = \begin{bmatrix} h_{11} & 0 & h_{13} \\ 0 & 1 & 0 \\ -h_{13} & 0 & h_{11} \end{bmatrix} \quad (9)$$

C. 2.5pt relative pose with unknown plane normal

The 2.5pt algorithm is an extension of the 2pt described in section III-A. The homography is designed as in Eq 2. However, when the plane normal $\mathbf{n}$ is not known we can't make use of the same linear constraint, thus all the 6 parameters of $\mathbf{H}$ have to be estimated. To do this, one more equation is needed which can be taken from a third point. Thus we stack the constraint equations of 2 points and 1 of the equations from a third point into an equation system of the form $\mathbf{A}\mathbf{h} = \mathbf{b}$. The two equations from one point are as follows:

$$\begin{bmatrix} 0 & 0 & -x & -1 & xy' & y' \\ -xy' & -y' & xx' & x' & 0 & 0 \end{bmatrix} \mathbf{h} = \begin{bmatrix} y \\ -x'y \end{bmatrix} \quad (10)$$

where $\mathbf{h} = [h_{11}\ h_{13}\ h_{21}\ h_{23}\ h_{31}\ h_{33}]^T$.

As in section III-A the solution to this system is of the form $\mathbf{h} = \mathbf{V}\mathbf{y} + w\mathbf{v}$. The unknown scalar w can again be computed from the additional homography constraint $det(\mathbf{H}^T\mathbf{H} - \mathbf{I}) = 0$ (see [13]). The determinant is a 4th order polynomial in w which results in 4 solutions for $\mathbf{H}$.

The interesting fact in this case is that we used only 1 equation of the 2 available ones for computing the homography. While in a RANSAC loop it is however necessary to sample 3 points for this method, it is now possible to do a consistency check on the 3 point correspondences. To be an outlier free sample the one remaining equation has also to be fulfilled by the estimated homography. This can easily be tested and if it is not fulfilled the hypothesis is discarded. This gives a computational advantage over the standard 3pt essential matrix method [3], because inconsistent samples can be detected without testing on all the other point correspondences.

D. 2pt absolute pose

With known vertical the absolute pose problem gets simplified as well and it is possible to compute the remaining 4DOF absolute pose from 2 3D-2D point correspondences. Here we again assume that the camera is pre-rotated by the vertical direction so that the x-z plane is parallel to the ground plane. The camera matrix is defined as $\mathbf{P} = [\mathbf{R}\ \mathbf{t}]$ which results in

$$\mathbf{P} = \begin{bmatrix} cos(\phi_y) & 0 & sin(\phi_y) & t_x \\ 0 & 1 & 0 & t_y \\ -sin(\phi_y) & 0 & cos(\phi_y) & t_z \end{bmatrix}. \quad (11)$$

There are 4 unknowns which are the rotation angle around the y-axis and one 3D translation vector. Using the relation $\mathbf{x} \times \mathbf{PX} = \mathbf{0}$ we can solve for these unknowns, where $\mathbf{X}$ is a homogeneous 3D point of the form $\mathbf{X} = [X\ Y\ Z\ 1]$ and $\mathbf{x}$ is an image point $\mathbf{x} = [x\ y\ 1]$. One 3D-2D correspondence gives 2 linearly independent equations of the form

$$\begin{bmatrix} yZ & -yX & 0 & -1 & y \\ -yX & -yZ & -y & x & 0 \end{bmatrix} \mathbf{p} = \begin{bmatrix} Y \\ -xY \end{bmatrix} \quad (12)$$

where $\mathbf{p} = [cos(\phi_y)\ sin(\phi_y)\ t_x\ t_y\ t_z]^T$. 2 point correspondences give an equation system with 4 equations. Variable elimination can be used to find expressions for t_x, t_y, t_z and eliminate it from the remaining 4th equation. The remaining equation is in $cos(\phi_y)$ and $sin(\phi_y)$ only and the additional constraint $cos(\phi_y)^2 + sin(\phi_y)^2 = 1$ can be used to get an expression in $sin(\phi_y)$. It is quadratic in $sin(\phi_y)$ and can be solved in closed form. Then the other parameters can be found by back-substitution, which leads to 2 solutions for the camera pose. A similar approach has been described in [7], however for our derivation we assume pre-aligned feature correspondences, while in [7] the measured angles of the IMU are included in the equation system. Furthermore our angles are parameterized in sin and cos while in [7] a representation with tan is used.

IV. Relative scale estimation without 3D structure recovery

The formulation of the homography induced by the ground plane Eq. 1 encodes the inverse distance d of the first camera to the plane. Assuming the ground plane is visible over all views, the relative scale can be propagated between views over the plane. The computation of the homography assumes the distance to the ground plane to be 1, as formulated in Eq. 1. The actual distance between the first camera and the ground plane ise encoded in the translation vector of the camera (i.e., y-component of the camera). This distance can then be used to rescale the relative translation vector of the second camera. The implicit encoding of the scale in the homography allows for direct scale estimation without the need of a computationally expensive recovery of the 3D structure.

V. Experiments

We evaluate the accuracy of the presented algorithms on synthetic data under different image noise and IMU noise. We compare the results to the general 5pt algorithm presented in [14] and the general 3pt algorithm proposed by [3] with two know orientation angles. Finally we demonstrate our algorithms on our own real world datasets.

A. Synthetic evaluation

To evaluate the algorithm on synthetic data we chose the following setup. The average distance of the scene to the first camera center is set to 1. The scene consists of two planes, one ground plane and one vertical plane which is parallel to the image plane of the first camera. Both planes consist of

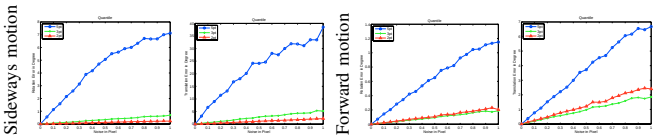

Fig. 2. Evaluation of the 2 point algorithm under different image noise.

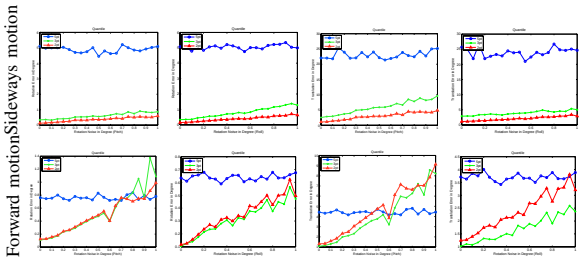

Fig. 3. Evaluation of the 2pt algorithm under different IMU noise and constant image noise with 0.5 pixel standard deviation. (First row sideways motion, second row forward motion.

200 randomly sampled points. The base-line between two cameras is set to be 0.2, i.e., 20% of the average scene distance, and the focal length is set to 1000 pixel, with a field of view of 45 degrees.
Each algorithm is evaluated under varying image noise and increasing IMU noise. Each of the two setups is evaluated under a forward and sideways translation of the second camera. For each configuration we randomly sample 100 cameras.

a) Relative pose:: Fig. 2 and Fig. 3 compare the 2-point algorithm to the general 5-point [14] and 3-point algorithms [3]. Notice, in this experiments the camera poses were computed from points randomly drawn from the ground plane. Since camera poses estimated from coplanar points do not provide a unique solution for the 5pt and 3pt algorithm we evaluate each hypothesis with all points coming from both planes. The solution providing the smallest reprojected error is chosen to be the correct one. This evaluation is used in all our synthetic experiments. Similarly Fig. 4 and Fig. 5 show a comparison of the 2.5pt algorithm with the general 5pt and 3pt algorithm. Here the camera poses are computed from points randomly sampled from the vertical plane only. The evaluation shows that knowing the vertical direction and exploiting the planarity of the scene improves motion estimation. The 2pt and 2.5pt algorithms outperform the 5pt algorithm, in terms of accuracy. Under perfect IMU measurements the algorithms are robust to image noise and perform significantly better than the 5pt algorithm. With increasing IMU noise their performance are still comparable to the 5pt algorithm.

b) Absolute pose:: We compare the presented 2pt absolute pose algorithm to the closed form solution proposed in [10]. We evaluate the algorithm on different noise in the image plane and noise in the roll and pitch measurements of the IMU. The results are shown in Fig. 6 and Fig. 7. With increasing image noise the absolute 2pt algorithm

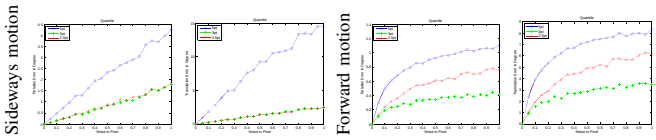

Fig. 4. Evaluation of the 2.5pt algorithm under forward and sideways motion under varying image noise.

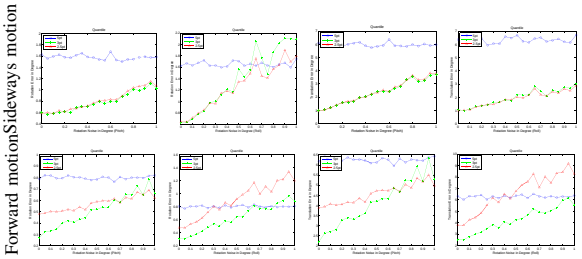

Fig. 5. Evaluation of the 2.5pt algorithm under IMU noise and constant image noise with 0.5 pixel standard deviation. (First row sideways translation, second row forward translation.

outperforms the 4pt algorithm. While with increasing IMU noise their approach has a higher accuracy.

B. Algorithm evaluation on real data

c) Plane detection:: We evaluate our relative motion algorithms on an image pair that contains multiple planes and further demonstrate that the weak Manhattan world assumption holds on real images, where vertical structures might not be perfectly vertical due to construction or due to IMU inaccuracies. Relative motion is computed with the 2pt algorithm as well as with the 2.5pt algorithm. The 2pt algorithm computes the relative motion from matches found on the ground, while the 2.5pt algorithm computes relative motion from matches found on the wall. Fig. 8 shows the inlier sets of the two methods in one image. The motion estimate from the homographies can be refined by constructing the essential matrix of it and finding inliers to the essential matrix (these need not be on planes) and computing a least squares estimate of the essential matrix. Furthermore, the detected inlier sets can be used for plane detection.

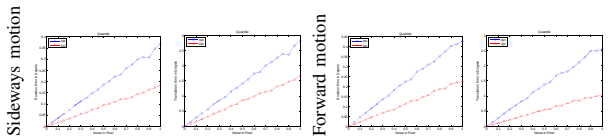

Fig. 6. Evaluation of the 2pt absolute pose algorithm under forward and sideways motion with varying image noise.

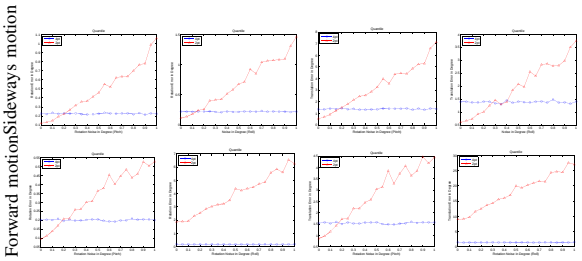

Fig. 7. Evaluation of the 2pt absolute pose algorithm under different IMU noise and image noise of 0.5 pixel standard deviation. (First row, sideways translation, second row forward translation.

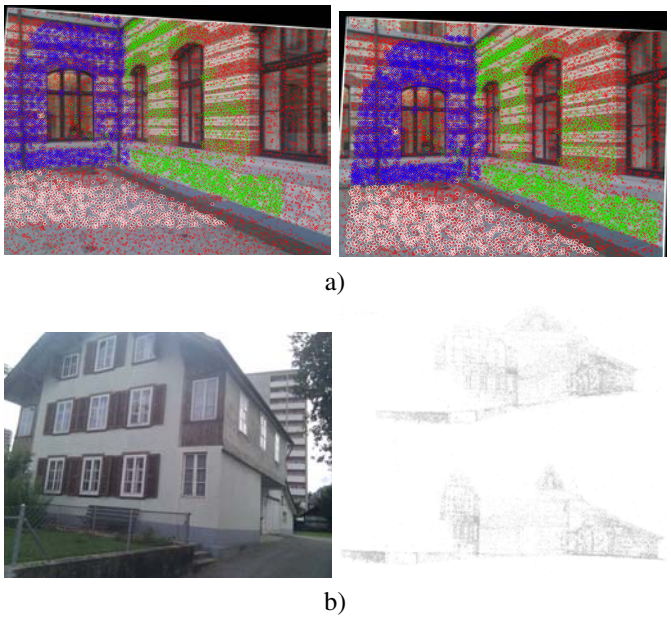

Fig. 8. a) Detected planes using the 2pt and 2.5 algorithm. b) Sample input image and two synthetic views of the reconstructed scene using the absolute 2pt algorithm.

d) Visual Odometry:: We evaluate the 2pt relative pose on our own dataset, recorded with an IMU-camera rig of a micro aerial vehicle. Sample images are shown in Fig. 8a). The experiment consists of 114 images showing a forward motion towards the front wall followed by a backwards motion to the starting point. First, we extract SIFT features and match them to neighboring views. Then, we compute the relative pose between two consecutive views using the 2pt algorithm in a RANSAC scheme. Finally, the solution with most inliers is used to form the camera trajectory by concatenating neighboring poses. Fig. 9 compares the raw odometry obtained from the 2pt algorithm without scaling and without refinement to the odometry obtained after non-linear refinement and proper scaling with the method described in section IV.

e) 2pt absolute pose:: We integrate the 2pt absolute pose algorithm into a standard SfM pipeline and show results of a reconstructed scene in Fig. 8b). The 2pt absolute pose algorithm is used to find a first inlier set from the 3d - 2d correspondences. The full camera pose and 3d points are then refined using bundle adjustment. The dataset was recorded using a Nexus One smartphone, equipped with IMU and camera.

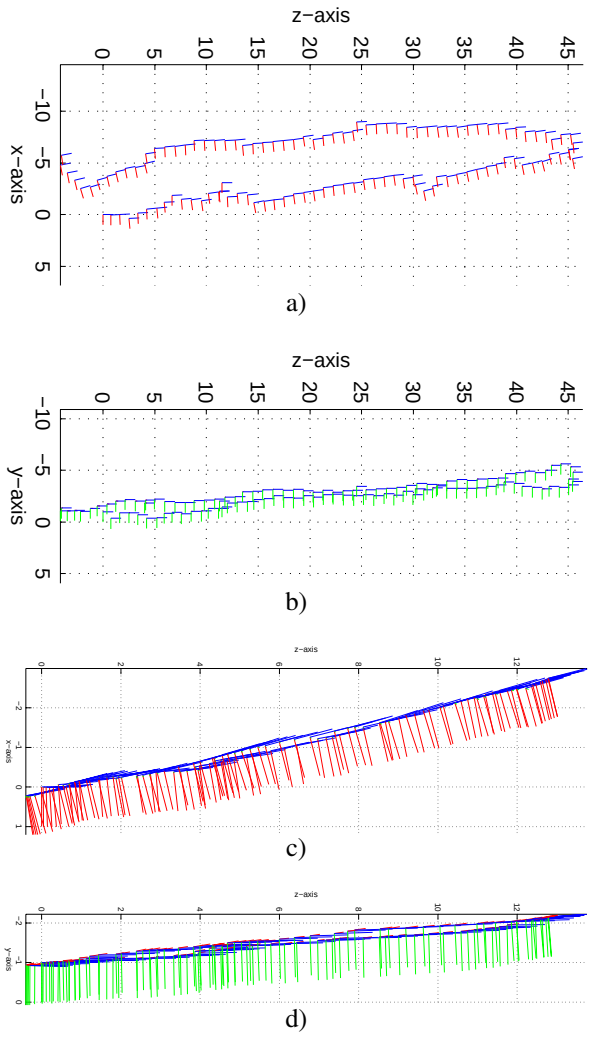

Fig. 9. (a,c) Top view of the camera trajectory. (b,d) Side view of the trajectory. (a,b) Non-refined trajectory obtained from the 2pt algorithm. (c,d) Optimized and properly scaled camera trajectory. (z-axis, motion direction, y-axis gravity direction, blue line optical axis).

VI. Conclusion

In this paper we presented an odometry pipeline that exploits the weak Manhattan world assumption and takes advantage of knowing the vertical direction in images. Our results show that the weak Manhattan assumption holds for real-world scenarios and can be used to derive efficient algorithms for relative motion (2pt, 2.5pt). Furthermore our results confirm that the vertical direction measured from an off-the-shelf IMUs are accurate enough to be used for relative motion estimation and absolute pose estimation (2pt).

References

[1] J.M. Coughlan and A.L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 941–947, 1999.

[2] M. A. Fischler and R. C. Bolles. RANSAC random sampling concensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of ACM*, 26:381–395, 1981.

[3] F. Fraundorfer, P. Tanskanen, and M. Pollefeys. A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In *Proc. 11th European Conference on Computer Vision*, pages IV: 269–282, 2010.

[4] Y. Furukawa, B. Curless, S.M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *Proc. 12th International Conference on Computer Vision*, pages 80–87, 2009.

[5] Y. Furukawa, B. Curless, S.M. Seitz, and R.S. Szeliski. Manhattan-world stereo. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Florida, Miami*, pages 1422–1429, 2009.

[6] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2000.

[7] Z. Kukelova, M. Bujnak, and T. Pajdla. Closed-form solutions to the minimal absolute pose problems with known vertical direction. In *ACCV*, 2010.

[8] Margarita Chli Laurent Kneip and Roland Siegwart. Robust Real-Time Visual Odometry with a Single Camera and an IMU. *Proceedings of the British Machine Vision Conference*, pages 16.1–16.11, 2011. http://dx.doi.org/10.5244/C.25.16.

[9] D.C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Florida, Miami*, pages 2136–2143, 2009.

[10] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *Int. J. Comput. Vision*, 81(2):155–166, feb 2009.

[11] J. Lobo and J. Dias. Vision and inertial sensor cooperation using gravity as a vertical reference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1597–1608, December 2003.

[12] G. López-Nicolás, J.J. Guerrero, and C. Sagüés. Multiple homographies with omnidirectional vision for robot homing. *Robotics and Autonomous Systems*, 58(6):773 – 783, 2010.

[13] Ezio Malis and Manuel Vargas. Deeper understanding of the homography decomposition for vision-based control. Research Report RR-6303, INRIA, 2007.

[14] D. Nistér. An efficient solution to the five-point relative pose problem. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin*, pages II: 195–202, 2003.

[15] D. Ortín and J. M. M. Montiel. Indoor robot motion based on monocular images. *Robotica*, 19(3):331–342, 2001.

[16] Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry [tutorial] part1: The first 30 years and fundamentals. *Robotics Automation Magazine, IEEE*, 18(4):80 –92, dec. 2011.

[17] T. Vieville, E. Clergue, and P.E.D. Facao. Computation of ego-motion and structure from visual an inertial sensor using the vertical cue. In *Proc. 4th International Conference on Computer Vision, Berlin, Germany*, pages 591–598, 1993.