Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 199–207

DOI: 10.15439/2015F182
ACSIS, Vol. 5

Towards the automatic motion recovery using single-view image sequences acquired from bike

Jakub Kolecki
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: kolecki@agh.edu.pl

***Abstract*—This paper describes the design, implementation and results of the image-based ego-motion estimation algorithm. As a source data the images captured from the bike platform are used. The device is supposed to be a part of a mobile mapping system prototype. Firstly the feature detection and matching is carried out providing the set of characteristic points in all images in the sequence. The 5-point solution based on the Gröbner basis is used to solve for essential matrices and to reject outliers. Least-square relative pose model fitting is accomplished using quaternion-based bundle adjustment. In the next step the modified Horn formula is used to recover bike trajectory up to the absolute orientation. Within this step the scene structure recovery is provided in the form of a point cloud. Finally ground control information is used to obtain data geo-referencing and the accuracy analysis. Obtained results provide satisfying robustness and accuracy. However some improvements and development scenarios are suggested.**

I. Introduction

CURRENTLY imaging sensors are extensively used as components of mobile mapping systems (MMSs), mobile robots and unmanned aerial vehicles. Each camera is a source of usually large number of images, captured with the specified frequency. Acquired image sequences may be processed to provide automatically extracted mapping information using algorithms refered as a dense point cloud generation or structure from motion (SFM). Additionally images may be utilized to estimate motion trajectory of the vehicles. Such application is often called the visual odometry. The real-time trajectory estimation is applied in the navigation. The visual navigation can take place autonomously or together with the inertial/GNSS sensors, completing a multi-sensor navigation system. Parallel navigation and mapping are sometimes combined together in the SLAM procedure.

Basically mapping applications do not require real time computation of a trajectory. The accurate trajectory computation is conducted in the post processing and is a crucial step in the mobile mapping workflow as it greatly influences the accuracy of final products. Processing of the image sequences can be divided into two main steps:

- feature detection and matching
- ego-motion estimation, based on the detected features

Researchers conducted within founds of Department of Mining Surveying and Environmental Engineering (AGH University of Science and Technology) no. 11.11.150.949

In the machine vision feature detectors try to imitate humane vision to search for some characteristic points (keypoints), that are suitable to be traced in subsequent images. The feature correspondence is tested using descriptors and detectors that try to simulate the mental process [1]. Corners are a typical example of features suitable for tracing. The result of a feature detection is a list of keypoints' IDs and their coordinates provided in the image 2D coordinate frame. Evaluation of particular feature detectors are not within the scope of this work, but generally a set of automatically measured keypoints has a large number (sometimes over 50%) of outliers i.e. missmatched features. The first approach to deal with outliers is to prevent false matches using the external information about image geometry. This information can come from positioning sensors such as GPS/INS systems [2], [3]. If the orientation of two images acquired with calibrated camera is approximately known, location of corresponding features is held down to the neighborhood of epipolar lines associated with those points (Fig. A1). However more robust approach to prevent false matches is to use multi-view camera configuration providing multi-view image sequences [4], [5], [6]. Commonly two cameras are applied. As a result of a system calibration the accurate orientation of the second camera in the coordinate frame of the first camera is known. This enables the accurate epipolar line location for a certain keypoint (Appendix A).

In the case of single-view sequences, the keypoint matching cannot really benefit from the epipolar constraint (Appendix A). If no GPS or IMU are available, only the approximate motion characteristic is known, constraining corresponding keypoint searching to region of interests (ROI) rather then lines. As a result a significant number of outliers may occur. Approaches to the outlier rejection are based on the epipolar constriant imposed on the fundamental matrix (F) or the essential matrix (E) (Appendix A). Snavely, Seitz and Szeliski [7] propose the estimation of the F matrix using the 8-point algorithm inside the RANSAC [8]. As a result a set of 8 image points that best fit the fundamental matrix model is found in each image pair. At the same time outliers can be detected and rejected. Bartelsen and Mayer [9] prefer to use the essential matrix instead of the fundamental matrix. In contrast to the fundamental matrix, the essential matrix estimation requires the knowledge about the camera calibration but remains robust to the critical configurations met in the

978-83-60810-66-8/$25.00

planar scenes. It also requires smaller number of coresponding points. The 5-point algorithm developed by Nister [10] and the locally optimized RANSAC [11] are proposed as a solution method. Finally the E or F model fitting can by carried out based on inliers only, using least-square approaches.

Applying the epipolar constraint cannot eliminate badly matched points that are located near the corresponding epipolar lines. A 3D information is necessary to detect the remaining, relatively small number of outliers. To solve the problem, the ray intersection (Appendix B) is carried out for each image model to calculate the spatial coordinates of the tie points (Fig. A.1). Sequential orientation of the subsequent images [2] using for example the DLT approach [12] inside the RANSAC procedure or formation of image triplets [9] allows to complete the rejection of outliers. Overlapping triplets can be linked to recover orientation of every image in the sequence. However the recovered orientation suffers from the drift effect. Besides it can be determined only up to the absolute scale, rotation and translation. Aforementioned 7 parameters, if needed, can be estimated based on ground control information such as the GPS coordinates of the image projection centers [9] or coordinates of the control points. Finally to estimate the orientation of the images more accurately the least square bundle adjustment can be applied. For the real time scenario the good solution to the drift reduction is the detection of loop closures [13].

A terrestrial mobile mapping can be carried out from almost every vehicle. Cars are used commonly in case of commercial systems. However data acquisition using cars is restricted to streets and their surroundings only. To overcome those limitations systems designed for smaller vehicles are developed, among which bikes seem to fill the gap between hand-held systems, mobile robots and cars. Bikes can access many more location than cars and move faster then pedestrians and mobile robots. Probably the most famous bike system for mobile data acquisition was developed by Google in the Street View project. Besides, students from Stuttgart University designed the prototype of the bike mobile mapping system with the laser scanner and two-antenna GNSS/INS unit [14].

Similarly to other mapping or visual odometry systems, bike systems can be a source of image sequences. Automatic processing of image data acquired from bike can be used to determine trajcetory and finally to reconstruct the geometry of objects. However it should be noticed that the bike movement is different to the car movement. When cycling it is more difficult to keep constant speed and direction than in case of driving. The turn rate of bike is small when compared to car. In case of single-view sequences with no information about approximate image geo-referencing, motion estimation from image data is supposed to be much more challenging task than in case of the multi-view systems and the smoother movement.

Works addressed in this paper aim to provide the solution to the problem of the automatic orientation of a single-view image sequences. It is beyond the scope of this study to examine the hardware potential as well as the achieved processing time. However the study offers some important insight into the analytical approaches and their practical aspects, providing at the same a kind of overview of the existing solutions. The following sections describe the consecutive steps of the algorithm, starting from keypoint detection and matching proceeding to the relative orientation and finally to the absolute orientation. The fifth section describes the experiments. Then the results and discussion are provided.

II. Feature detection and matching

As this step of the algorithm is still under the development, only the outline of the matching strategy is provided. In the urban environment corner points are likely to occur in almost every image and can be detected using one of many available detectors. Proposed algorithm uses Kovesi's implementation [15] of Noble's version [16] of Harris feature detector [17]. After completing the detection, features are matched using the monogenic phase approach [18], [19] with the set of parameters proposed by Kovesi [15]. Assume a certain corner is detected and matched in the initial pair in the sequence. Matching algorithm searches for the corresponding point in the third and subsequently in the next images. At the same time new corners appear in consecutive images. As the approximate image-to-image distance and the scene depth are known, some simple geometric constrains like row and column limits can be imposed on a searching area greatly reducing the computation time.

Corner points are generally detected as the maxima in the image that is the result of applying Harris operator. The absolute maximum value (AMV) constraint can be set to limit the number of keypoints. However applying this simple constraint leads to nonuniform distribution of detected points. This is the result of variations in the scene content. Some parts of the scene like trees, windows, cars etc. are "rich" in keypoints, while others like flat walls contain no keypoints at all. To overcome this problem the image is divided into blocks of equal sizes. In the addressed case study the 24 blocks (4 $\times$ 6) are used . Besides absolute value of Harris maxima, two other parameters are set to provide more favorable distribution of detected features: maximum number of points in each block (PIB) and minimum distance between points (DBP). Decreasing the AMV and the PIB and at the same time increasing the DBP leads to the more uniform distribution of points. Exemplary values of feature detection parameters, applied in the refereed case study were as follows: AMV = 10 (Kovesi suggested 200 [15]), PIB = 60, DBP = 50.

As the result of matching procedure a list of points and their pixel coordinates is provided, allowing computation of the relative orientation of consecutive images.

III. Relative orientation

A. Essential matrix computation

The relative orientation of two images acquired with calibrated camera is encoded in the 3 $\times$ 3 essential matrix E. Derivation of the essential matrix from the relative orientation parameters - translation and rotation (t,R) and the camera matrix can be found e.g. in Krauss [20], or in the simpler

form in the Appendix A. However the solution of the inverse problem is not so trivial as there are 4 possible solutions that have to be tested for cheirality [21]. The essential matrix as well as the fundamental matrix satisfies the well known complanarity constraint [20], [21]:

$$x'^{\top} E x = 0 \tag{1}$$

where x and x' are the column vectors of homogenous image coordinates. The rank deficiency of essential matrix implies the following constraint:

$$\det(E) = 0 \tag{2}$$

The E matrix has two non-zero singular values that are equal. This constraint can be expressed in the algebraic form as [22]:

$$2EE^{\top}E - \mathrm{tr}(EE^{\top})E = 0 \tag{3}$$

It is advantageous to compute the essential matrix using one of few available close-form solutions using minimal, i.e. 5, number of corresponding points [10], [23], [24], [25]. Using the close-form solution requires no prior approximation and can be easily tested for outliers using the RANSAC procedure. The algorithm developed by Stéwenius, Engels and Nistér [25] was adopted within proposed solution because of its relatively simple implementation. Using equation (1) and finding its four-dimensional null-space, the essential matrix can be parametrized with three unknowns x, y, z:

$$E = xE_1 + yE_2 + zE_3 + E_4 \tag{4}$$

Inserting equation (4) into (2) and (3) produces the system of 10 3rd degree polynomial equations in three unknowns. This system is solved using the Gröbner basis. Up to 10 solutions for E exist but only the real ones are of the further interest.

B. Detection of outliers

In the proposed approach the essential matrix is estimated inside the RANSAC procedure. This enables detection of outliers. Assume that for the subsequent image pairs in the sequence N samples are chosen, each consisting of 5 point pairs. As there are up to 10 real solutions for E, in the worst case there can be $10N$ possible solutions. According to the typical RANSAC each point in the image pair is classified as inlier or outlier according to the specified threshold value. In the addressed solution a kind of locally optimized RANSAC [11] is used. All points in each sample get a score that is inversely proportional to the distance from the model value. If the distance is higher than the threshold, the score is zero and the point is classified as outlier. The sample with the highest total score wins.

The easiest way to score each point is to calculate distance to the epipolar line using (1) (see Appendix A). However it may happen that due to mismatching, a point that is projected near the epipolar line lies behind the camera. To avoid treating such points as inliers it was decided to recover the rotation matrix (R) and the translation vector (t) of the second image from each real E [21] and to calculate coordinates of keypoints in three dimensional coordinate frame of the first image using intersection of rays (Fig. A1, Appendix B). Now only the keypoints with negative Z coordinates are going to be tested further. Given R, t and estimated 3D coordinates of keypoints, the projections to images are found so that the 2D euclidean distances to the measured locations can be computed.

Assume three consecutive image pairs: [k,k+1], [k+1,k+2], [k+2,k+3]. A certain feature is matched correctly in pair no. 1. Subsequently this feature is matched incorrectly in pair no. 2. As a result it is classified as outlier. Nevertheless this keypoint may not be rejected because it could happen that incorrectly matched feature in image k+2 is correctly matched with the feature in image k+3. As a consequence this keypoint is recognized as two separate keypoints and gets separate id's in images forming pairs 1 and 3. It won't be used in relative orientation of pair no. 2.

C. Estimation by least square fitting

The sample with the best score provides good estimations of R and t but it does not take into account all inliers. To utilize all available information, the least square adjustment can be carried out using all points classified as inliers. The image coordinates of keypoints are treated as observations and explicitly related to the parameters in the form of well known colinearity equations (see e.g. [20]). As a consequence the system of nonlinear equation is formed. The elements of R are not treated as parameters directly. To avoid possible singularities resulting from parametrization in terms of Euler angles [26], the entries of the R matrix are expressed as functions of the elements [27] of a quaternion. Both quaternion and t are assumed to have unit norms that leads to the additional constrains imposed on the parameters. Finally the 3D coordinates of all tie points complete the set unknowns. The R, t and 3D coordinates of tie points resulting from the RANSAC should be accurate enough to linearize the equation system and subsequently solve it in only one iteration.

IV. Sequence orientation

A. Model-to-model transformation

As a result of the relative orientation the R and t are provided for each image pair in the sequence. Such relatively oriented image pair is called a model. Assume image k forming the model with image k+1.The consecutive model is formed by images k+1 and k+2. Common points are now used to stitch both models. In this way the orientation of a short, 3 image, sequence is recovered. Subsequently the algorithm proceeds to the transformation of the third model based on the reference points that appear in the previously created block. Model stitching is carried out further, until all images are oriented.

Each model is oriented according to the Horn algorithm [26] and involves estimation of 7 parameters: 3 for the rotation, 3 for the translation and finally the scale. The Horn approach consists of the following steps. At first the centroids in both point sets are calculated. Coordinates of all points are reduced to respective centroids. Secondly the rotation that maximizes

the dot product of vectors pointing from centroids to corresponding points is found. The rotation is parametrized in terms of four elements of the unit quaternion. Once the quaternion is known, it is possible to align vectors in both frames to make them nearly parallel. Vectors won't never be exactly parallel due to outliers. Finally the scale is recovered using translated and rotated vectors. Three different approaches to scale computation are proposed depending on which point set is assumed to have better accuracy. The approach of Horn deals with minimal 3-point case as well as with greater number of points. It fulfills the condition of the least sum of squared residuals. Finally no approximation of the parameters is needed.

It should be mentioned that in addition to the terrain points (keypoints), adjacent models have one additional common point, namely the projection center of the common image - k+1 in the later example. This point lies far away from the rest of points and certainly has a worse reliability. Applying a standard Horn solution would cause that even a small errors in 3D coordinates of the tie points can result in large residual of projection center locations incorporating relatively large errors to the estimated motion. Therefore during the minimization of the dot product the utilization of the model frame coordinates is preferred to the usage of coordinates reduced to their centroids. In fact such modification means that the translation is simply calculated, not estimated, hence the estimation of remaining four parameters is more reliable i.e. less sensitive to the influence of erroneous tie point locations.

During the model-to-model transformation the sparse point cloud of keypoints is being formed. Besides points used inside the Horn algorithm, each stitched model incorporates a set of new points. If this points appear also in the next model, they are used as the reference. However some points exist that appear only in one model. The correctness of the location of such points cannot be fully checked. As a result some erroneous points in the sparse point cloud appear.

B. Absolute orientation

Until now the image sequence was oriented up to the scale, absolute rotation and absolute translation. If the missing parameters are to be recovered, the external information need to be utilized. Basically there are two approaches to provide the external orientation to images: direct measurement and geo-referencing through ground control points (GCPs) referred as the indirect approach. To measure the external orientation directly one can use GPS and inertial sensors. If the GPS is used alone, the estimation of absolute orientation parameters takes place using the coordinates of projection the centers recovered in the previous step and the reference trajectory line recorded by a receiver [9]. Geo-referencing through control points usually requires the manual measurement of terrain features, the coordinates of which are known from other survey. In case the terrain coordinates of control points are known from a geodetic survey the indirect approach is assumed to be more accurate. In the presented study the second approach was utilized as no GPS measurements were available. After completing the absolute orientation it is possible to smooth the results and increase accuracy by performing the bundle adjustment. In such a case the loop closures, if only present, can be taken into account for the further accuracy increase.

V. Experiments

A. Preparatory works

The accuracy of the motion recovery from single view sequences strongly depends on the imaging geometry. In case of corridor sequences, when camera looks forwards or backwards, the intersection angles between correspondent rays are narrow, leading to the large errors of tie point locations. As a result the recovered camera orientation tends to drift quickly. In contrast to the corridor sequences a sequences with camera looking perpendicular to the moving direction (aside-looking sequences) should allow to achieve a better accuracy. In the following tests only the motion recovery from the aside looking sequences is covered, however the algorithm is supposed to deal with the geometry of any kind.

To test the proposed approach the decision was made to acquire the image sequence of the test-field area located at the AGH University Campus (Fig. 1). This test field is equipped with the number of natural GCPs, that are to be used to evaluate the accuracy. GCPs are located mostly at the building façades. It was decided to use the wide angle camera to be able to capture the façades from top to bottom. In addition to the large overlap, even in case of wide baseline, the wide angle lens provides increased accuracy of depth component of the tie point location in space. This is of the fundamental importance for the process of model stitching as the drift is supposed to accumulate slower. Besides, the obtained sparse point cloud would have better accuracy than in case of using the narrow angle lenses. In addition to better accuracy the wide angle lens performs better when imaging in motion. It guarantees a large depth of field allowing imaging with small aperture and short exposure time. Taking all the above into consideration the Nikon D5200 camera with the Sigma 10-20 mm f/3.5 rectilinear lens was chosen as the imaging sensor. In addition to the acquisition of a high resolution 24 megapixel (4000×6000) images the sensor of the camera allows HD video recording. The focal length was set to 12 mm providing the horizontal viewing angle about $90\,^{\circ}$. The principal distance was fixed by blocking the focusing ring. The camera was calibrated to determine the interior orientation parameters and the distortion.

B. Data acquisition

Initially the tests involving acquisition of HD videos were made, but because of low quality of extracted frames it was decided to switch the camera to the time-lapse mode, choosing the highest possible frequency of 1 Hz. However it came out quickly that capturing images with the 1 Hz frequency makes the camera buffer stuck - the shutter is not released until the last image is saved. Lowering the frequency would either lengthen the imaging base, possibly leading to the problems with feature matching or force decreasing the cycling speed

Fig. 1. Planned trajectory line imposed on the image of the test-field

resulting in extension of the overall acquisition time. The reasonable solution to avoid the above mentioned effects was to switch to the lower resolution of 13.488 megapixel.

The camera was fastened to the bike using a specially constructed device consisting of the 3 DOF head, allowing sequence acquisition from freely selectable viewing angle. The camera was inclined to look slightly upwards and perpendicular to the cycling direction. The test sequence was acquired in the aperture priority mode. The aperture value was set to 5 resulting in the exposure time between 1/2000 and 1/1000 second. The planned trajectory line is shown in the Fig 1. The test sequence was to have the shape of a loop. With the aim of comparison a part of the loop was to be cycled twice. The decision had to be made which side the camera should look at. Choosing the right direction provides convergent image configurations within all the turns and a good overlap. However in the case of the test-field the test were carried out in (Fig. 1) it was better to look left as to capture the façades lying closely to the trajectory line, providing advantageous distribution of keypoins. The disadvantages of such configurations are the occurrence of the divergent images within the turns leading to decrease in the overlap and occurrence of the narrow angles of intersecting rays.

After applying all the above mentioned settings the sequence of 195 images was acquired.

C. Data processing

After collecting the data, the keypoint detection and matching algorithm was tested. The rough motion characteristic was known allowing to restrict the location of possible matches to the ROIs of a fixed size. The imposed constraint was supposed to reduce the number of possible outliers. The keypoint matching is followed by the relative pose estimation of consecutive image pairs. The least square relative orientation was tested but due to a very long computation time it was not applied in the final solution. Four image pairs, each located within the turns were not oriented properly due to the very high outlier rate and improper keypoint distribution. In this case the problem was fixed by adding some tie points manually.

Having the relative pose of the subsequent models estimated, the sequence formation was carried out. The first model in the sequence was chosen as a starting model. As the relative pose estimation constrains the base vector to equal 1, all the linear quantities calculated within this stage such as translations, residuals, errors are expressed in the unit of the length of the first base. The threshold parameter of the RANSAC procedure, i.e. the linear residual of the tie point, was set to 0.1.

For now the orientation of all of the images in the sequence was estimated up to the absolute quantities (translation, rotation, scale). To solve for the missing parameters and provide the accuracy analysis the 26 natural control points were used. Each control point was measured in the selected model (image pair). The accuracy assessment was provided by the residuals of the control point coordinates. Finally the sparse point cloud provided in the global coordinate frame was examined visually to look for previously undetected mismatches. The extent of the inconsistencies observed as a result of cycling the same part of the loop twice were to be analysed deeply.

VI. Results

Despite applying the ROI-restricted matching a large number of outliers was observed in almost all image pairs (Fig. 2). During the RANSAC-based estimation of the relative pose it came out that the number of outliers considerably exceeds 50%. Besides the limitations of the monogenic phase matcher the reason of such a high outlier rate could be simply the content of the scene. For instance a number of corners appearing on the similar windows' frames are hard to be matched correctly. In addition the epiploar lines are nearly parallel to the horizontal edges of windows' elements so that even solving for relative pose cannot eliminate some outliers. It can also be noticed that there are quite a lot of trees in front of the façades (Fig. 1, Fig. 2, Fig. 3). As a results a number of a false keypoints is detected at the intersections of branches and twigs. Some of them are also incorrectly matched. The similar happens for keypoints detected in the reflections appearing in the window panes (Fig. 2). Also a lot of corner points detected at the grainy structure of the asphalt are matched incorrectly. Using the 5-point algorithm inside the RANSAC allows to eliminate most of the outliers. Fig. 2 and Fig. 3 are provided as an example.

It was decided to examine the influence of the drift (accumulation of the errors within the sequence formation stage) on the accuracy of the absolute orientation. The results are provided in the table 1. At the beginning the sequence of 10 images was oriented using four GCPs. The centimeter-level errors were obtained. Afterwards the number of images was increased until the appearance of a next group of available control points. Finally the sequence of 158 images was oriented. As no GCPs were measured in the further images, the last row of the table 1 represents the accuracy of the absolute orientation of the whole sequence. Performed analysis shows that generally

Fig. 2. One of the images from the southern part of the sequence and the vectors showing the displacement of the keypoints to the the next image. Results before applying the 5-point relative pose estimation algorithm inside RANSAC.

Fig. 3. One of the images from the southern part of the sequence and the vectors showing the displacement of the keypoints to the the next image after automatic rejection of outliers

while increasing the length of the sequence errors tend to increase, however not in the regular way. The worst results were obtained for the Y coordinate and the best for the Z.

Fig. 4 shows the sparse cloud of tie points obtained as a result of the sequence formation. The colour of points changes from blue to red as to show the inconsistencies in the point cloud resulting from the orientation drift. The first matched keypoint in the first image pair is coloured in blue. The last matched keypoint in the last model is coloured in red. The black spots represent the location of projection centres. The black line represents the trajectory. The total length of the trajectory is 232.58 m. The arrows show the cycling direction

The façades are clearly visible in the cloud as well as the kerbs and the trees. During the data acquisition there was not as many cars parked as it can be seen in the Fig. 1. Few of them can also be visible in the cloud. There are also quite a lot of points that seem to be located inside the buildings. This tie points may represent mismatched keypoints

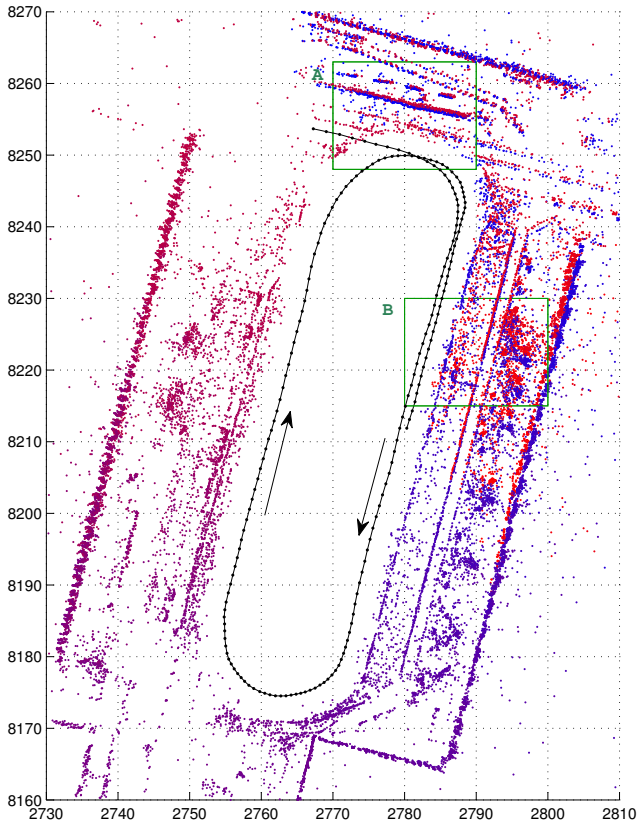

Fig. 4. The trajectory line and the point cloud. The arrows point the cycling direction. Green rectangles A and B mark the areas that will be referenced further. Orientation: north, units: meters.

located only in two images, occupying consistent epipolar lines. Such points pass the RANSAC testing, carried out within the model formation, but cannot be tested in the model stitching procedure. As a result of capturing certain parts of the scene twice, the inconsistencies in the resultant point cloud appear. To show them in details two parts of the cloud bounded by green rectangles are shown in the greater scale in the Fig. 5 and Fig. 6.

The thickest strip of points in the Fig. 5 represents the front edge of the hedge, part of which is visible at the bottom of the Fig. 1. Points forming four segments parallel to the hedge are likely to be located at the crowns of the trees and the items used to shape them in the espalier-like form - see the very bottom of the Fig.1. In front of the hedge there are some points at the pavement. Some of them form a linear features that may represent kerbsides. Two linear features that lie behind the trees represent the railings of the ramp that belongs to the building the image in the Fig. 1 was captured from. Looking at the points located at the hedge it can be noticed that the red points are shifted with respect to the blue ones. The shift is about 40 cm in the south-north direction. When looking at the trees and railings and finally at the façade (Fig. 4) this shift seem to decrease.

TABLE I
ACCURACY OF THE ABSOLUTE ORIENTATION OF THE SEQUENCES OF A DIFFERENT LENGTH. THE LAST COLUMN PROVIDES THE RMS ERRORS OF THE 3D CONTROL POINT LOCATION.

Num. of images	**Distance [m]**	**Num. of points**	$RMSE_X$ **[mm]**	$RMSE_Y$ **[mm]**	$RMSE_Z$ **[mm]**	$RMSE_P$ **[mm]**
10	15.31	4	6	16	12	21
34	42.94	7	63	33	30	77
55	75.82	9	87	71	60	127
73	95.48	12	99	111	60	160
89	107.46	16	212	224	83	320
107	126.83	16	290	265	103	406
141	177.39	23	272	403	126	502
158	193.56	26	267	405	240	542

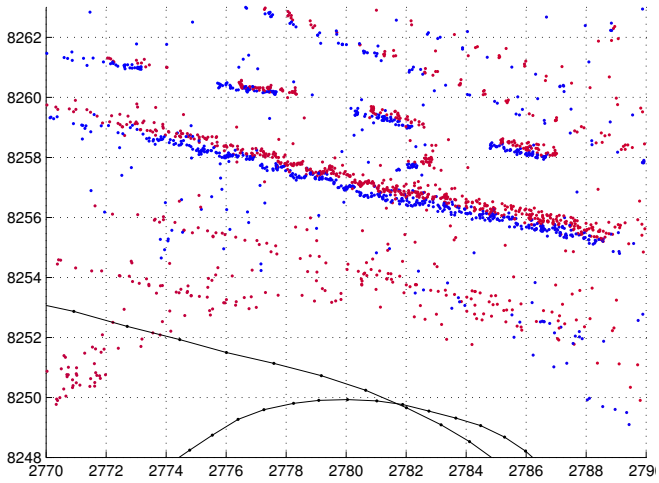

Fig. 5. Inconsistency of the point cloud within the area A

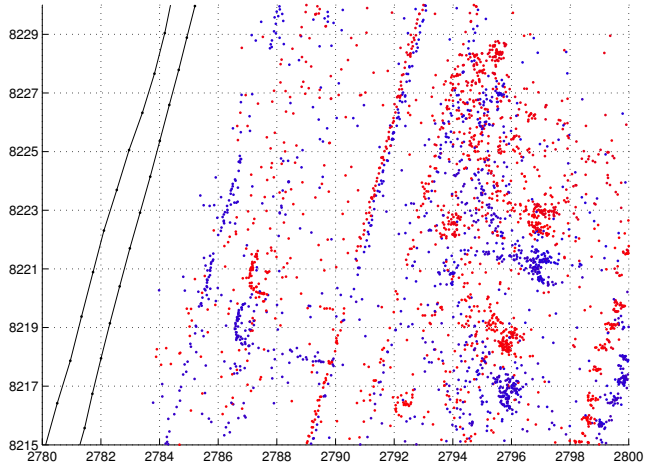

Fig. 6. Inconsistency of the point cloud within the area B

In the Fig. 6 the points located at the car body can be observed as well as linear features representing kerbs. There are groups of points representing trees that grow in the front of the building the façade of which can be seen in the bottom right corner of the figure. Looking at the points representing the car one can notice a considerable inconsistency in the point cloud. The red points are shifted about 2 meters with respect to blue points. The reason why the shift increased to such a high value can be the unfavourable imaging geometry (divergent camera axes, decreased overlap) at the north-east turn. Finally it can be found quite unexpected that no drift in the heading component of the angular orientation can be noticed - the blue and red linear features visible in Fig. 5 and Fig. 6 stay almost exactly parallel.

VII. DISCUSSION

In this paper the new solution to the automatic orientation of single-view image sequences is proposed and the results of the tests conducted based on the data acquired from bike are presented. The solution assumes the calibrated camera case to achieve more robust performance in outlier detection and better accuracy. Modifications to the model-to-model stitching procedure are proposed as to achieve better reliability of the sequence formation, which result in more robust trajectory estimation.

The conducted tests allow the examination of certain steps of the proposed solution. The first step i.e. the feature detection and matching seem to perform quite well for images with a similar angular orientation. However it tends to fail for images captured within turns so that even the manual point measurement was to be carried out to fix the problem. To improve the keypoint matching firstly a more robust feature descriptors and matcher can be applied. Secondly the matching should be integrated with the relative pose solution. The E matrix is quite accurately estimated using the proposed method, even in the presence of outliers, so that the equations of epipolar lines can be used to impose a stronger constraints for the feature re-matching (Appendix A). Afterwards features can be re-matched after solving for the essential matrix. Then the refined essential matrix is to be estimated and the solution can proceed in the iterative manner. Having the robustness improved one can think about improving the accuracy by using the sub-pixel corner measurement. Additionally the motion recovered within the process of model stitching can be smoothed using the bundle adjustment, however this approach may take quite a lot of computation time.

Assuming no real time application it is also better to select a different starting model for the sequence formation. Probably choosing the model near to the middle of the sequence would reduce the error as the drift is to accumulate on distances of about the half distance of the sequence.

The improvement of the orientation procedure is going to be followed by integration of other sensors like GPS or IMU. It would demand changing the imaging sensor from SLR to the industrial camera. Besides providing the time synchronization interface the industrial camera allows imaging at the higher frequency and at the same moment allowing real-time image data processing. Generally at this stage of the research the obtained results can be found satisfactory and consist a good starting point for developing a bike MMS equipped with the visual orientation unit. There exists a field to improve the robustness, accuracy and operational performance of the solution by improving both algorithms, implementation and a hardware.

APPENDIX

A. Essential matrix the and epipolar constraint for a calibrated camera

Assume that the two images of approximately the same scene were taken with calibrated camera (Fig. A.1). Assume this two images form photogrammetric model. Relative orientation of the second image with respect to the first image can be parametrized by the orthonormal rotation matrix (R) of the second camera frame and translation vector (t) of the second camera projection center (O'). The translation vector simply equals the base vector (b). The relative orientation can be estimated up to scale factor. Usually b is assumed to be the unit vector.

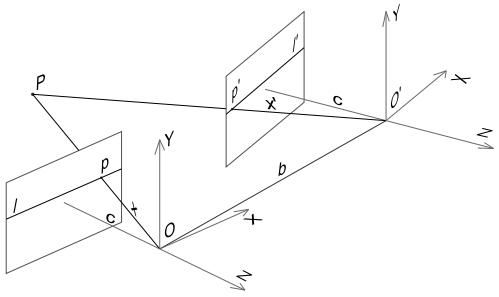

Fig. A.1. Two metric images forming the model

Point P is located in the scene and projected into images to form points p and p'. Assume two vectors x and x' that originate in projection centers O and O' and point at points p and p'. Cooridinates of those vectors are given in the reference frame of respective cameras so that the third coordinate is equal to the principal distance of camera with the minus sing and for metric images is assumed to be the same, i.e.:

$$x = \begin{bmatrix} \xi \\ \eta \\ -c \end{bmatrix}, x' = \begin{bmatrix} \xi' \\ \eta' \\ -c \end{bmatrix} \tag{A.1}$$

Now the coplanarity constraint reads as follows:

$$x^\top (b \times Rx') = 0 \tag{A.2}$$

Coordinates of b fill the elements of skew-symmetric matrix B:

$$B = \begin{bmatrix} 0 & -b_z & b_y \\ b_z & 0 & -b_x \\ -b_y & b_x & 0 \end{bmatrix} \tag{A.3}$$

so that:

$$x^\top BRx' = 0 \tag{A.4}$$

and consequently:

$$x'^\top Ex = 0 \tag{A.5}$$

where E is the essential matrix. The projection center O and point P define a ray that is projected into second image as the line l' (Fig. A1). The equation of this line is obtained by inserting the coordinates of x into equation (A.5). In the similar way the equation of the epipolar line l can be derived.

B. Intersection

Assuming the R and t are known, it is now possible to estimate the coordinates of point P in the model coordinate frame, i.e. the coordinate frame of the first camera (Fig. A.1). If points p and p' represent two correctly matched keypoints, theoretically both rays should meet in point P. However due to measurement errors rays are not going to intersect. Knowing the x and x' vectors (A.1) the location of point P can be determined by least square solution. Collinearity equations for two corresponding keypoints are as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_P = \lambda_1 \begin{bmatrix} \xi \\ \eta \\ -c \end{bmatrix} \tag{B.1}$$

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_P = \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} + \lambda_2 R \begin{bmatrix} \xi' \\ \eta' \\ -c \end{bmatrix} \tag{B.2}$$

where λ_1 and λ_2 are unknown scale coefficients. After elimination of λ_1 and λ_2 from (B.1) and (B.2) followed by term's rearrangement the observed 2D coordinates of the keypoint can be written using explicitly elements of the relative orientation as the functions of unknowns:

$$\begin{bmatrix} \xi \\ \eta \\ \xi' \\ \eta' \end{bmatrix} = \begin{bmatrix} -c\frac{X_P}{Z_P} \\ -c\frac{Y_P}{Z_P} \\ -c\frac{R_{1,1}(X_P-b_x)+R_{2,1}(Y_P-b_y)+R_{3,1}(Z_P-b_z)}{R_{1,3}(X_P-b_x)+R_{2,3}(Y_P-b_y)+R_{3,3}(Z_P-b_z)} \\ -c\frac{R_{1,2}(X_P-b_x)+R_{2,2}(Y_P-b_y)+R_{3,2}(Z_P-b_z)}{R_{1,3}(X_P-b_x)+R_{2,3}(Y_P-b_y)+R_{3,3}(Z_P-b_z)} \end{bmatrix} \tag{B.3}$$

This system of equations can be rewritten in the linear form. The solution provides coordinates of point P. In case of erroneous measurements in x and x' the projection of estimated P point into images won't coincide with points p and p' so that the residual vectors will appear.

REFERENCES

[1] A. Śluzek, M. Paradowski,"Is Visual Similarity Sufficient for Semantic Object Recognition?", *Computer Science and Information Systems (FedCSIS) Federal Conference on. IEEE,* Wrocław, 2012, pp. 167-173.

[2] R. J. Handley, J. P. Abbott, C. R. Surawy, "Continuous Visual Navigation - An Evolution of Scene Matching", *Proceedings of the 1998 National Technical Meeting of The Institute of Navigation,* Long Beach, CA, 1998, pp. 217-224.

[3] C. V. Tao, M. A. Chapman, B. A. Chaplin, "Automated Processing of Mobile Mapping Image Sequences", *ISPRS Journal of Photogrammetry and Remote Sensing,* 55, 2001, pp. 330-346, DOI: http://dx.doi.org/10.1016/S0924-2716(01)00026-0

[4] D. Griessbach, D. Baumbach., S. Zuev,"Vision Aided Intertial Navigation," EUROCow, Castelldefels, 2010.

[5] F. Fraundorfer, D. Scaramuzza, "Visual Odometry - Part II: Matching, Robustness, Optimization and Applications", *IEEE Robotics and Automation Magazine*, June, 2012, pp. 78-90, DOI: http://dx.doi.org/10.1109/MRA.2012.2182810

[6] Y. Xu, F. Chen, "Real-time and Robust Visual Navigation Localization Algorithm based on ORB", *Applied Mechanics and Materials,* Vol. 241-244, 2012, pp. 478-482, DOI: http://dx.doi.org/10.4028/www.scientific.net/AMM.241-244.478

[7] N. Snavely, S. M. Seitz, R. Szeliski, "Modeling the World from Internet Photo Collections", *International Journal of Computer Vision*, 80(2), 2007, pp. 189-210, DOI: http://dx.doi.org/10.1007/s11263-007-0107-3

[8] M. A. Fischler, R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* 24(6), 1981, pp. 381-395.

[9] J. Bartelsen, H. Mayer, "Orientation of Image Sequences Acquired from UAVS and with GPS Cameras", EUROCow, Castelldefels, 2010.

[10] D. Nister, "An efficient solution to the five-point relative pose problem", IEEE PAMI, 26(6), 2004, pp.756-770, DOI: http://dx.doi.org/10.1109/TPAMI.2004.17

[11] O. Chum, J. Matas, J. Kittler, "Locally Optimized RANSAC," *Pattern Recognition - DAGM,* Springer Verlag, Berlin, 2003, pp. 249-256, DOI: http://dx.doi.org/10.1007/978-3-540-45243-0_31

[12] Y.I. Abdel-Aziz, H.M.Karara, "Direct linear transformation from comparator coordinates into object-space coordinates in close-range photogrammetry," *Proceedings of the ASP/UI Symposium on Close-Range Photogrammetry,* Falls Church, VA, 1971 pp. 1-18.

[13] K. L. Ho, P. Newmann, "Detecting Loop Closure with Scene Sequences," *International Journal of Computer Vision*, 74(3), 2007, pp. 261-286, DOI: http://dx.doi.org/10.1007/s11263-006-0020-1

[14] FARO, http://blog-uk.faro.com/2013/08/mobile-mapping-system-do-it-yourself/

[15] MATLAB and Octave Functions for Computer Vision and Image Processing, http://www.csse.uwa.edu.au/ pk/Research/MatlabFns/index.html

[16] A. Noble, "Descriptions of Image Surfaces", PhD thesis, Department of Engineering Science, Oxford University, 1989, p. 45.

[17] C.G. Harris and M.J. Stephens, "A combined corner and edge detector", *Proceedings Fourth Alvey Vision Conference,* Manchester, 1988, pp 147-151.

[18] M. Felsberg and G. Sommer, "A New Extension of Linear Signal", *Processing for Estimating Local Properties and Detecting Features,* DAGM Symposium, Kiel, 2000.

[19] M. Felsberg and G.Sommer, "The Monogenic Signal", *IEEE Transactions on Signal Processing,* 49(12), 2001, pp. 3136-3144, DOI: http://dx.doi.org/10.1109/78.969520

[20] K. Kraus, "Photogrammetry - Geometry from Images and Laser Scans", Walter de Gruyter, Berlin, 2007

[21] R. Hartley, A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, 2003, DOI: http://dx.doi.org/10.1017/CBO9780511811685

[22] J. Philip, "A Non-Iterative Algorithm for Determining all Essential Matrices Corresponding to Five Point Pairs", *Photogrammetric Record,* 15(88), 1996, pp. 589-599.

[23] Z. Kukelova, M. Bujnak, T. Pajdla, "Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems", BMVC 2008.

[24] D. Batra, B. Nabbe, M. Hebert, "An alternative formulation for five point relative pose problem", *IEEE Workshop on Motion and Video Computing,* 2007, DOI: http://dx.doi.org/10.5244/C.22.56

[25] H. Stewénius, C. Engels, and D. Nister, "Recent developments on direct relative orientation", *ISPRS Journal of Photogrammetry and Remote Sensing,* 60, 2006, pp. 284-294, DOI: http://dx.doi.org/10.1016/j.isprsjprs.2006.03.005

[26] B. Wrobel, D. Klemm, "Über die Vermeidung singulärer Fälle bei der Berechnung allgemeiner räumlicher Drehungen," *International Archives of Photogrammetry and Rmote Sensing,* 25, 1984, pp. 1153-1163.

[27] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions." JOSA A, 4.4, 1987, pp. 629-642, DOI: http://dx.doi.org/10.1364/JOSAA.4.000629