DERIVING PEDESTRIAN POSITIONS FROM UNCALIBRATED VIDEOS

Zoltan Koppanyi, Post-Doctoral Researcher
Charles K. Toth, Research Professor
The Ohio State University
2046 Neil Ave Mall, Bolz Hall
Columbus, OH, USA, 43210
koppanyi.1@osu.edu
toth.2@osu.edu

Tamas Soltesz
Assistant Lecturer
Budapest University of Technology and Economics
Stoczek utca 2, Building ST
Budapest, Hungary, 1111
soltesz.tamas@mail.bme.hu

ABSTRACT

Uncalibrated camera videos, such as recordings by surveillance cameras, offer large datasets to analyze pedestrian behavior in traffic situations. This paper addresses the problem to derive pedestrian trajectories from these types of data sources. Processing of uncalibrated images is challenging due to the lack of accurate ground control points, intrinsic camera and distortion parameters. During test data acquisition, GoPro cameras are installed around intersections and record the traffic. The applied workflow for video processing includes the estimation of the camera poses (1) and the pedestrians' trajectories (2). GoPro cameras are not calibrated earlier, yet general calibration model, including the linear and lens distortion parameters is assumed and applied to rectify the video images for (1). The ground control points are measured in Google Earth to obtain the extrinsic parameters of the cameras. The collinearity equations are solved by applying the Levenberg-Marquardt method to provide robust solution. For the trajectory estimation (2), the pedestrians' image coordinates are digitized frame by frame in all video recordings. The 3D coordinates can be derived with triangulation using the collinearity equations. The initial guess for this algorithm is estimated by a 2D interpolation function, basically using a DEM, modeled by a plane of the road surface, to obtain the 3D coordinates of an image pixel. The results suggest that 0.5-1 m relative accuracy between the derived trajectories can be achieved, suitable for many traffic engineering investigations.

KEYWORDS: uncalibrated video, pedestrian tracking, traffic tracking, fish-eye lens, photogrammetric workflow, close-range photogrammetry

INTRODUCTION

Pedestrians are the most vulnerable road users (VRU) and require special protection. Currently, many car manufacturers offer warning systems that alert drivers of nearby pedestrians. These warning systems use camera and radar technology, and rely on proximity sensing. One of the key issues is that pedestrian movement patterns are more unpredictable compared to vehicles, and thus, they can appear in the sensor's field of view only at the last moment when the accident can no longer be avoided. Consequently, a more robust alerting system needs to be situation aware. This means that it has to be able to predict the most likely scenarios of pedestrian movements in a particular traffic situation, based on their trajectory and body gestures. The goal of the European Union's Proactive Safety for Pedestrian and Cyclists (PROSPECT) project is to understand these pedestrian scenarios to improve situational analysis. Recordings of existing surveillance videos, or camera systems that can be easily installed, are essential data sources for this behavior analysis. Clearly, the existing recordings are mass data sources for machine learning algorithms that can provide model for the traffic scenarios. In this study, photogrammetry is used to derive the positions of pedestrians' actions, the locations of their gestures, and, in overall, their trajectory relative to other road users, from multiple video recordings.

There are several issues to be addressed in the conventional photogrammetric workflow in order to extract positions from existing video sources. These issues are mostly related to the lack of a prior photogrammetric planning. Clearly, these videos, such as surveillance camera recordings, are not primarily installed for deriving geometric data. Figure 1 compares the conventional photogrammetric approach to derive 3D coordinates in the object space from digitized image points with challenges of using uncalibrated video sources. These challenges are the followings:

- **No camera calibration:** In the conventional photogrammetric workflow, the first step is the principal point correction and then the removal of the radial and tangential distortion from the images. Removing distortion allows us to use the pin-hole camera model and the collinearity equations. In most cases, surveillance cameras have significant distortions, and unfortunately, these calibration parameters are not available. In this paper, we assume general distortion parameters for certain types of cameras.

- **No GCPs**: Ground control points are essential to accurately determine the extrinsic or exterior parameters of the camera and for quality control. The extrinsic parameters allow for spatially relating cameras, and thus, deriving 3D coordinates in the object space. For existing camera sources, it is unlikely to have surveyed GCPs in the camera field of view. In this paper, we use GCPs digitized from road infrastructure elements, such as road curbs or traffic signs. Then, the 3D coordinates of these points are measured in Google Earth.

- **Numerical instability:** The solution of the collinearity equations without well-calibrated camera and good GCPs suffers numerical instability. Furthermore, it is difficult to find good initial values that are important to provide convergence. In addition, surveillance cameras are not likely to point to the same spot of the surveilled area, and thus, images may not overlap, and there are no tie points between the images. The lack of tie points also increases the uncertainty and instability of the computation to resolve the extrinsic parameters and to derive trajectories. This paper presents a processing workflow to increase the stability.

The proposed solutions for these challenges are presented in the "Methods" section. The "Data Acquisition" section provides information on the developed software that is used to digitize pedestrians in the videos. The "Results" section presents the solution of the camera pose estimation and the derived pedestrian trajectories. The paper ends with discussion and conclusion, where we summarize the main findings of this effort.

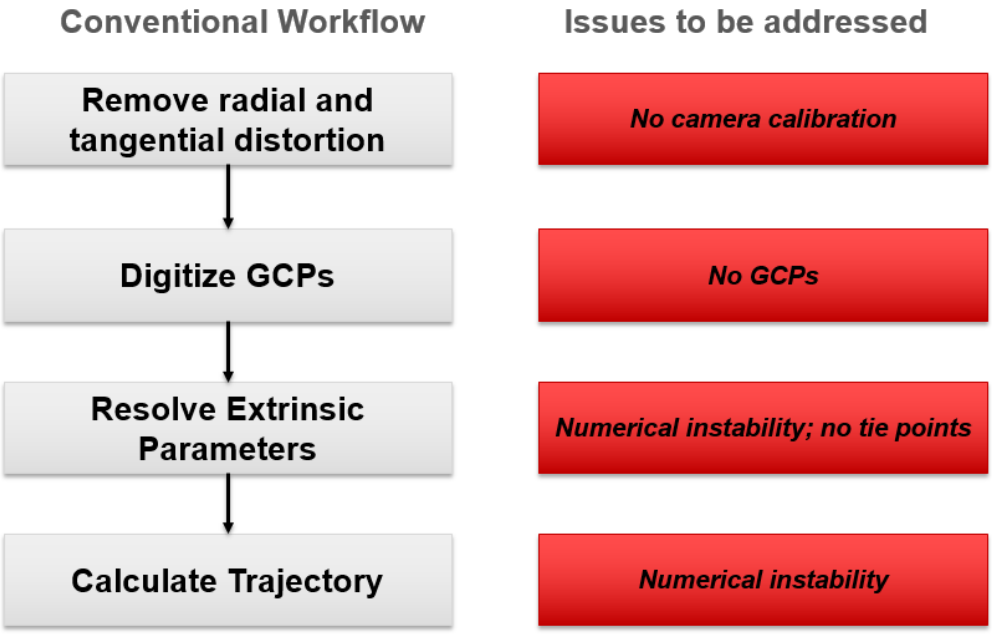


Figure 1 Conventional photogrammetric workflow and issues to be addressed for uncalibrated cameras

METHODS

Camera calibration

The camera calibration is solved for a typical camera, and then the derived parameters are used for all others. The intrinsic and distortion parameters are determined using OpenCV library (Bradski and Kaehler 2013). A calibration video is recorded a 7-by-10 chessboard at various location in the camera's field of view. Figure 2 shows one of the snapshot from the calibration video and the identified chessboard corners. Since the size of chessboard's squares are known, the intrinsic and distortion parameters can be determined from multiple images. Here, we used 23 images captured from various chessboard locations. Note that the chess pattern in the calibration images should cover the entire image area to obtain acceptable distortion model (this is not the case in Figure 2). Figure 3 shows the undistorted image. Note that the edges of the shelf on the left side is straight compared to Figure 2, which indicates that the distortion is effectively removed.

Figure 2 Calibration board captured by the GoPro camera.

Figure 3 Undistorted image, after applying distortion parameters.

Ground control points from Google Earth

To achieve the best performance, in photogrammetry, the conventional way is to use control points to determine the extrinsic parameters of the cameras (McGlone et al. 2004). The control points are points whose coordinates are known in the image plane (x_i, y_i) as well as in the global frame (X, Y, Z). If these coordinate pairs are known, the elements of the rotation matrix and the position of the principle point, i.e. extrinsic parameters, can be derived. The control points have to be measured with conventional surveying tools, such as total station or GNSS technique, to provide the most accurate coordinates. Because of the lack of field surveys, the global coordinates of the control points are obtained from Google Earth, which only provides 2D coordinates. The problem here is that the collinearity equations cannot be solved with 2D coordinates due to the fact that the nonlinear equation system becomes singular. To overcome this problem, we can choose a plane in the object space, and all points in the object space are assumed to be located on this selected plane. Here, the road surface was chosen as the dedicated plane. In this case all image points have $Z = 0$ coordinate in the object space, and thus, the collinearity equations are

$$x_i = -f_x \frac{R_{11}(X - X_0) + R_{21}(Y - Y_0) + R_{31}Z_0}{R_{13}(X - X_0) + R_{23}(Y - Y_0) + R_{33}Z_0} + c_x,$$
$$y_i = -f_y \frac{R_{12}(X - X_0) + R_{22}(Y - Y_0) + R_{32}Z_0}{R_{13}(X - X_0) + R_{23}(Y - Y_0) + R_{33}Z_0} + c_y. \quad (1)$$

Once the extrinsic parameters are determined, the 3D object coordinates can be calculated. Note that, for this problem, Equation 1 has two equations and two unknowns that means a single camera is enough to determine the 3D coordinates of an image point assuming that the point is located on the dedicated plane.

Numerical instability

In most cases, the nonlinear system of Equation 1 is solved with Gauss-Newton method (McGlone et al. 2004) in least squares sense for the extrinsic parameters and for resolving the 3D coordinates of a pixel point. The disadvantage of this approach is that it is sensitive for the initial values. In contrast, trust-region least-squares numerical optimizers, such as the dogleg-method, provide a more robust estimation of the parameters (Nocedal and Wright 2006). Thus, it is essential to choose a good optimizer due to the uncertainties of the weak calibration and inaccurate GCP points. In this research, we use the Levenberg-Marquardt method (Jianchao and Chern 2001). Although these optimizers are very robust, yet they still require initial values. The initial guess has an impact on the algorithm whether it terminates in a local minimum or not. To overcome this issue, an interpolation function is also determined during the estimation of the extrinsic parameters:

$$F: (x_i, y_i) \rightarrow (X, Y, Z), \quad (2)$$

where F is a 2D interpolation function. This function provides the initial guess for the Levenberg-Marquardt algorithm.

DATA ACQUISTION

For the PROSPECT project, data acquisition software is developed that enables for digitizing and tagging pedestrians and vehicles in videos. The entities are digitized with bounding boxes. The user interface is presented in Figure 4. The application enables to attach metadata to the digitized entities, i.e. pedestrians or vehicles. These metadata can be, for instance, a pedestrian looking around before stepping on the crosswalk or listening music or walking in group, essential information for understanding typical pedestrian scenarios. Recordings at 26 intersections are captured with GoPro cameras around Budapest, Hungary. 1022 pedestrians and 469 vehicles are digitized manually by 20 users. Thus, the developed software had to manage concurrent users who might work on the same videos parallel. Therefore, the software is able to connect to a PostgreSQL database management system server and the recorded data are stored there. Authentication, version and user management are also implemented to support team work; see the client-server architecture in Figure 5. For the test dataset, we choose a site, where three GoPro cameras were attached to separate traffic posts around an intersection, see Figure 6. The GoPro cameras were equipped with "fish-eye" lenses to provide large field of view (Kannala and Brandt 2006). Snapshots from the scene can be seen in Figure 7.

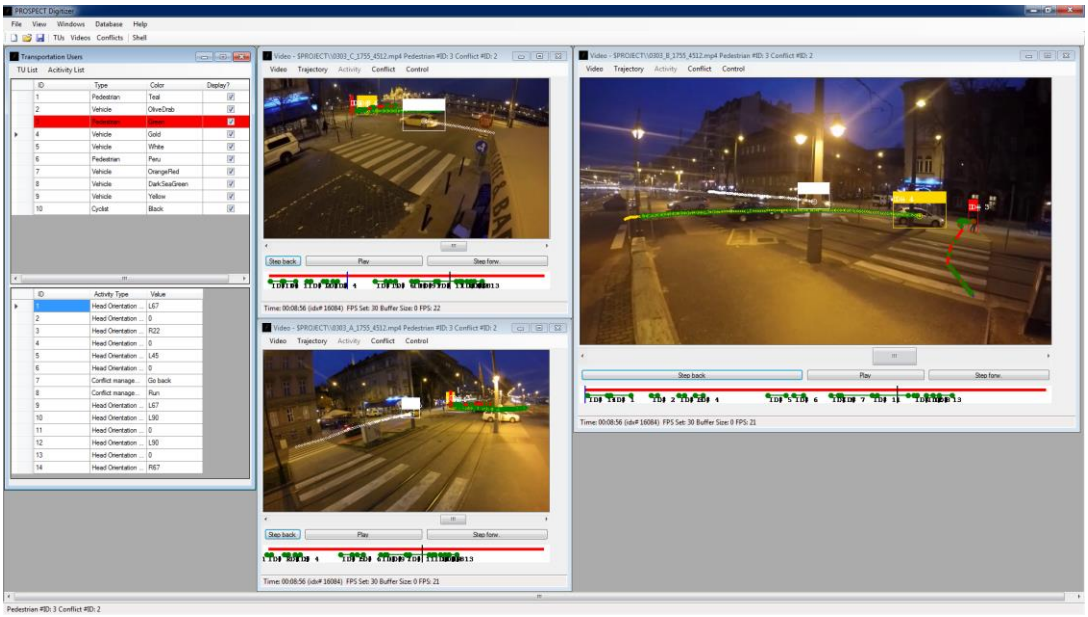


Figure 4 User interface of the digitizer software.

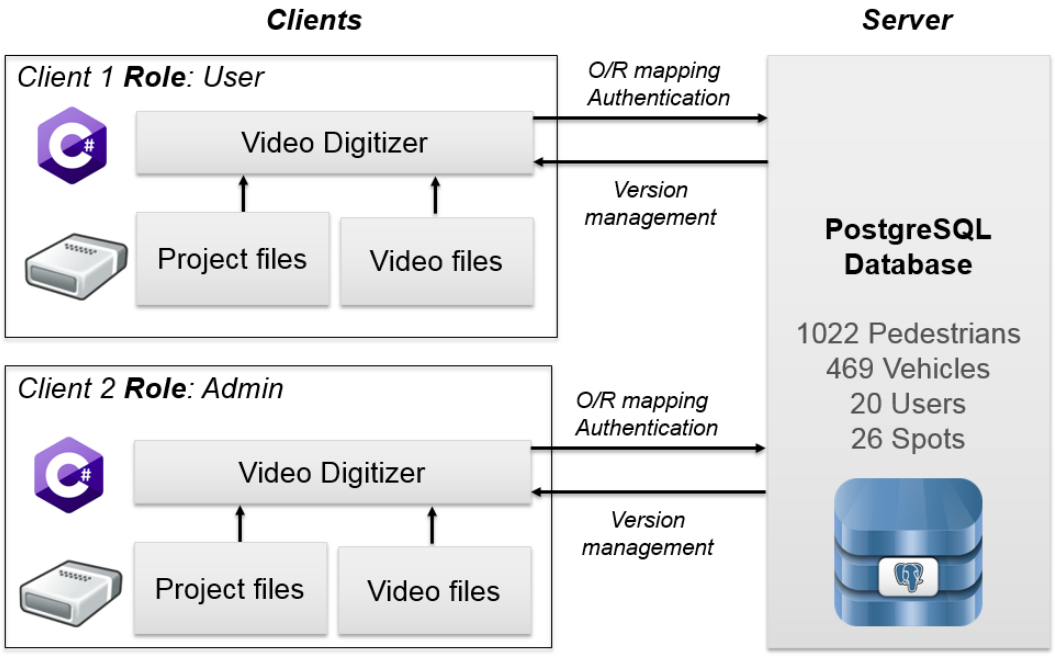


Figure 5 Software architecture of the digitizer software.

Figure 6 GoPro cameras attached to traffic signs around an intersection.

Figure 7 Camera snapshots from the three GoPro cameras, installed around an intersection in downtown Budapest, Hungary.

RESULTS

The following workflow is applied to resolve the extrinsic parameter estimation and to generate trajectories:

- **Measure GCPs in the images.** The image coordinates of the control points are measured in the images. The global UTM coordinates of these points are measured in Google Earth, see Figure 8. Typically, road features, such as curbs, corners, crosswalk paintings and traffic signs are used as GCPs, because they are easy to recognize and identify in Google Earth.

- **Solve the extrinsic parameters.** The extrinsic parameters can be estimated with the digitized GCPs using the approach presented in the "Methods" section. After this calculation, the camera position is known in the global frame. Figure 9a presents the GCP points depicted by red triangles and the camera position represented by blue triangle in the object space. These 3D coordinates are projected back onto the image plane. Thus, the reprojection error of each control point can be derived in pixels. These errors are represented as error ellipses in Figure 9b. Note that the error ellipses are larger around the edges of the image due to the inaccurate camera distortion parameters.

- **Derive the trajectory of the VRUs in the object space.** The pedestrians' image coordinates, epoch by epoch and by cameras, are known from the digitalization. The digitized image coordinates of the bounding boxes of the pedestrians are corrected by the distortion parameters. Note that we have to choose a reference point based on the bounding box to represent the trajectory. The bottom center point of the bounding box is selected as reference point, as we have to guarantee that this point is on the street surface, see Equation 1. Trajectories can be obtained from each camera resulting in three different trajectory solutions for each tracked person. The final trajectory is derived with calculating the median positions of the three trajectories. The trajectories of 10 pedestrians are presented in Figure 10 in top view, and in Figure 11 in Google Earth.

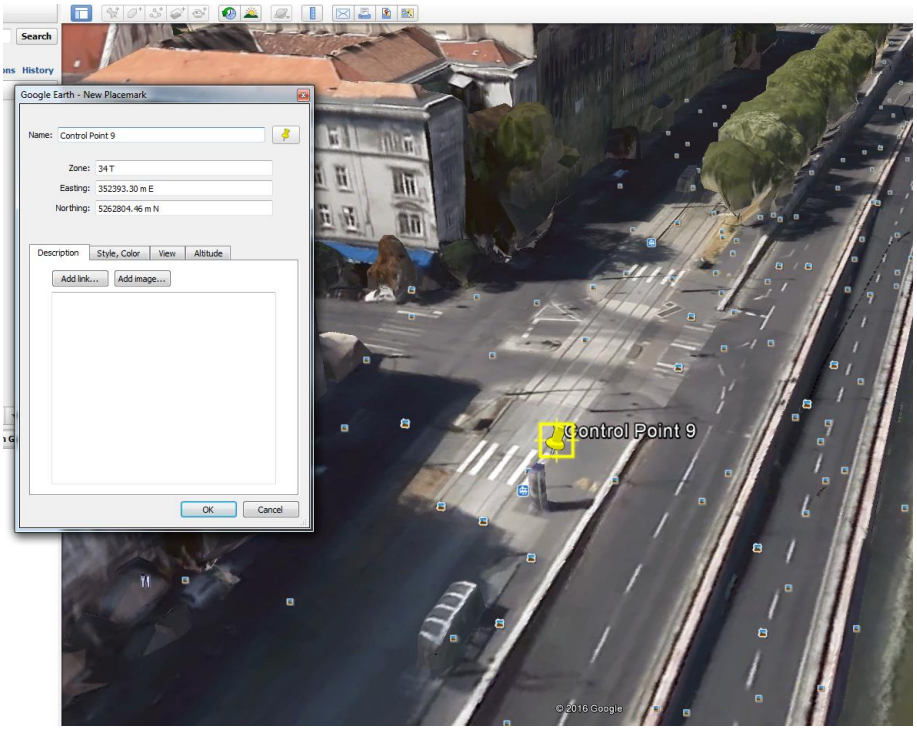
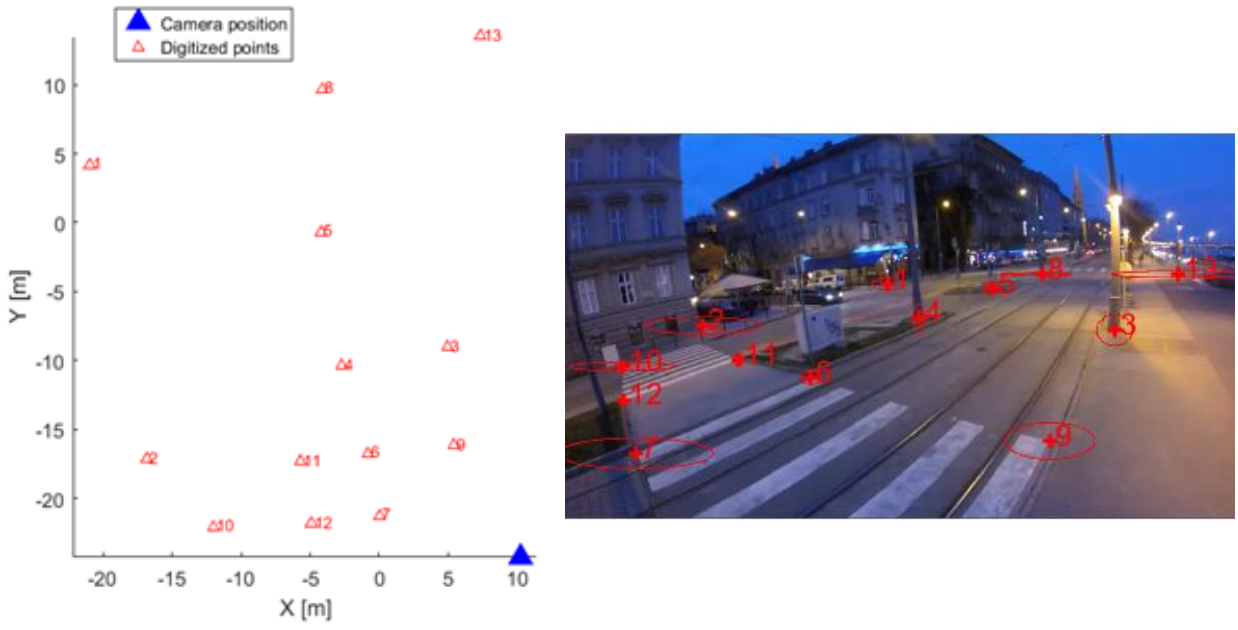


Figure 8 Measuring control point in Google Earth.



(a) (b)

Figure 9 Camera position (blue) and control points (red) in the local object space (a), and the reprojection error represented by error ellipses (b).

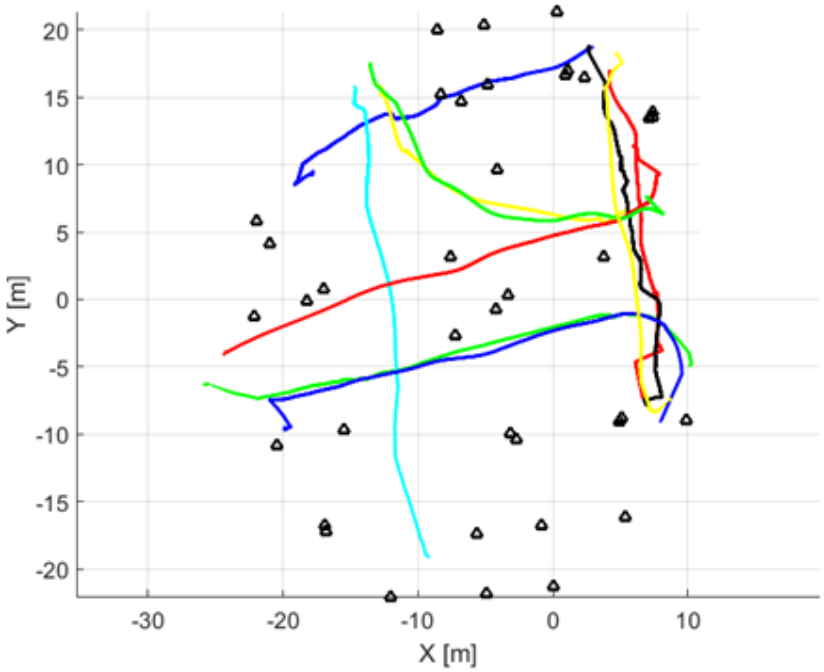


Figure 10 Trajectories of 10 pedestrians with different colors and GCPs depicted by black triangles.

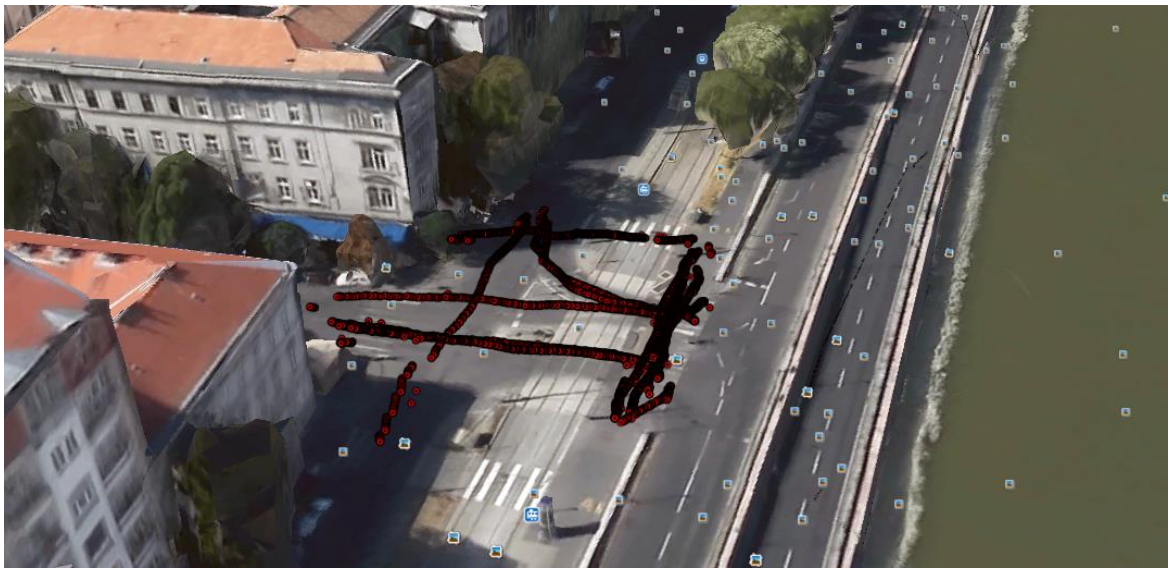


Figure 11 Trajectories in Google Earth.

DISCUSSION AND CONCLUSION

As no ground truth solution is available, the positioning accuracy can be evaluated only qualitatively. In general, the positioning error is larger at the edges of the images. The global accuracy may achieve 1-2 meters around the center of the covered area and 1-5 meters along the edges. The relative error is easier to assess, and it is around 0.5-1 m, based on Figures 9 and 10. Clearly, the relative error is the relevant accuracy metric in this study, as the distance between pedestrians and vehicles are of interest.

There are three main error sources that have relevant impact on the positioning error. First, the sensor model is assumed to be pin hole, despite the inaccurate distortion, the intrinsic parameters. Second, the accuracy of the control point coordinates determined from Google Earth is low. And third, as we assumed, all image points, of which object space coordinates are estimated, have to be on the dedicated plane, i.e. the road surface. But the pedestrians are digitized as bounding boxes in the images, and thus, a reference point has to be chosen to represent the image coordinates epoch by epoch. If the reference point derived from the bounding box is not on the road surface, then the derived 3D position is corrupted.

Overall, the proposed photogrammetric workflow is capable for deriving pedestrian trajectories from a single uncalibrated video. Using more videos improves the accuracy and reliability in the scene. The relative accuracy of the derived trajectories are suitable for investigating typical pedestrian scenarios in various traffic situations. From the derived trajectory solutions, traffic engineers are able to calculate the kinematic parameters of the pedestrians, such as acceleration, relative position and speed with respect to vehicles. Knowing trajectories allows them to find collision points; these are the points, where the pedestrian and vehicle trajectories intersect each other. At a fixed epoch, the time to reach this point is called Time To Collision (TTC), which is an important parameter, since it gives the amount of time to avoid an accident. As the road users converging to this point, TTC decreases, until one or both of them takes an action and change their speed or direction. The minimum value of TTC, the relative position and speed at the moment of collision as well as the estimated impact point on the vehicle well describe the severity of each situation. Clearly, the proposed photogrammetric workflow is able to provide this information with the desired accuracy from videos captured by inexpensive, easy-to-install and easy-to-use camera systems.

REFERENCES

Bradski, G, and Adrian K., 2013. *Learning OpenCV: Computer Vision in C++ with the OpenCV Library*. 2nd edition, O'Reilly Media, Inc.

Jianchao, Y., and Chern, C.T., 2001. Comparison of Network-Gauss with Levenberg-Marquardt Algorithm for Space Resection. *Proceedings of 22nd Asian Conference on Remote Sensing*. Singapore.

Kannala, J., and Brandt, S.S., 2006. A Generic Camera Model and Calibration Method for Conventional, Wide-Angle, and Fish-Eye Lenses, *IEEE Transactions on Pattern Analysis and Machine Intelligence 28* (8): 1335–40.

McGlone, J. C., Mikhail E.M., Bethel, J.S, 2004. *Manual of Photogrammetry*. American Society for Photogrammetry and Remote Sensing.

Nocedal, J., and Wright, S., 2006. *Numerical Optimization*. 2nd edition. Springer, New York, NY, USA