IMPLEMENTATION AND ACCURACY EVALUATION OF FIXED CAMERA-BASED OBJECT POSITIONING SYSTEM EMPLOYING CNN-DETECTOR

Tero Partanen, Philipp Müller, Jussi Collin, Jani Björklund

Tampere University, Finland

ABSTRACT

Today, different positioning applications such as location-based services and autonomous navigation are requiring more and more precision. Especially fully autonomous navigation requires accurate positioning solution, not only for the vehicle but also for the surrounding objects. Thus, many new positioning techniques, algorithms and fusion schemes have been developed. One essential technique is visual positioning. Thanks to intensive research in neural networks and deep learning, Convolutional Neural Network-based (CNN) object detectors have evolved greatly in recent years. This paper proposes a widely deployable scheme of fixed camera-based (e.g. surveillance camera) object positioning utilizing the CNN-detector. The accuracy of the implemented positioning solution is evaluated with precise Real-Time Kinematic (RTK) satellite positioning receiver. The implemented system can be used in indoors and outdoors, and it can estimate simultaneously positions from multiple camera views for multiple objects in real-time. When positioning a person, the measured mean positioning error was 10.7–15.6 cm with a simple bias correction and a standard deviation was 6.7-8.7 cm. Thus, the accuracy is excellent and would be sufficient to wide variety of applications.

Index Terms *— Camera-based positioning, visual localization, convolutional neural networks (CNN), object detection, RTK GNSS*

1. INTRODUCTION

Nowadays vast amount of positioning applications exists, such as several types of *Location Based Services* (*LBS*) [1] and autonomous navigation. Especially, position-based automation has recently rapidly increased in the form of autonomously navigating vehicles. The position information can be important for the object itself in such applications as social networking, personal navigation, robot navigation, and *Autonomous Vehicle* (*AV*) navigation, but also third parties can utilize the object's position for, e.g. automated entrance and surveillance control, emergency applications, and vehicle remote control. In addition, the position information can be shared between the object and external side, for example to enable fusion or assisted navigation.

Global Navigation Satellite System (*GNSS*) [2] is the primary positioning system that allows to determine a user's position on Earth. However, GNSS performance in many areas is compromised. Especially indoors GNSS signals can be totally blocked. Also, outdoors GNSS signals can be weak or absent in many areas such as tunnels, multi-story parking lots, urban canyons, and any other areas covered by a roof or blocked by structures. In good conditions low-cost GNSS positioning solution may achieve average position errors of less than 10 meters, which is usually sufficient for personal navigation, but e.g. for AV the accuracy even in the ideal conditions may be insufficient. Multiple positioning algorithms and techniques have been developed to assist GNSS outdoors or replace GNSS indoors (see e.g. [3]) to improve the position accuracy. For outdoors the most commonly used technique is a fusion of GNSS and *Inertial Navigation System* (*INS*) [4],[5]. The INS is also widely used indoors since most smartphones nowadays include an *Inertial Measurement Unit* (*IMU*). However, the IMUs used in smartphones have, in general, low accuracy, wherefore errors accumulate quickly [6]. Together with the need for an initial position this leads to unreliable solutions. Therefore, INS usually has to be fused with other techniques.

Positioning systems can be categorized as active (object mounted) systems, passive (external) systems or a fusion of both. Examples of active systems include INS, ultrasound, *Ultra-Wideband* (*UWB*), LIDAR, Wi-Fi, Bluetooth, RFID, and visual positioning with mobile camera [3]. Several of the above techniques are based also on external infrastructures in addition to an object mounted device. An example of passive systems is fixed camera-based positioning [7], which can be implemented on surveillance cameras. Thus, it is also infrastructure-based. A major drawback of infrastructure systems is the additional cost associated to building the infrastructure. However, in the case of a surveillance camera system, such infrastructures are very common and utilizing an existing system for positioning can be potentially implemented with zero additional infrastructure cost. All the mentioned systems can be used also outdoors, but the range (distance between infrastructure and object) is usually limited to 1-50 m [8].

Computer vision has been intensively researched in recent years. Developments in neural networks and deep learning approaches have greatly advanced the performance of the visual recognition systems [9]. A very popular and effective architecture for computer vision is the *Convolutional Neural Network* (*CNN*) [10]. It is nowadays used for computer vision tasks such as image classification, object detection, object localization, object tracking, and semantic segmentation. Object detection and object localization can be utilized also for a camera-based positioning system. Pre-defined objects are searched from image or video (object detection) and the spatial object coordinates in the image are extracted (object localization). CNN-based detector can be trained to detect almost any object type, like a person, animal, robot, vehicle etc. Camera based positioning may be used for many similar applications as other positioning systems, for example for context aware location-based marketing [11],[12], in health service to localize patients and medical staff, in disaster management and recovery to find objects, for AV positioning aiding, and especially for security and surveillance purposes. It is also possible to fuse camera positioning with other positioning technologies, for example with INS [13],[14] or GNSS.

This work studies fixed camera-based visual positioning systems, which are underrepresented in previous papers related to indoor or outdoor positioning [3],[8],[15]-[20]. We review several previous works and implement our own solution. Instead of previously common background subtraction [21] method to detect moving objects, we use deep learning in the form of CNN-based

detector to detect and localize the objects in video frames. Unlike background subtraction, CNN will also detect non-moving objects and reliability in dynamic conditions is expected to be better. By using a planar homography [22] the discovered object coordinates are mapped from camera image coordinates to local area coordinates and further to geographic coordinates. The implemented visual positioning system is tested outdoors to examine the absolute positioning accuracy while a *Real-time Kinematic* (*RTK*) GNSS receiver is used as ground truth.

The purpose of this study is to describe how to implement real-time surveillance camera-based object positioning system using CNN detector and what exactly is the expected positioning accuracy. To the best of our knowledge, this is the first work that employs precise RTK GNSS to determine the accuracy of CNN-employed surveillance camera-based positioning system. The present paper is organized as follows. Section 2 represents the previous works and compares the methods used in this study. Section 3 describes the adopted methods used in this work. Section 4 introduces the implemented system and test setup. Section 5 presents the results and analysis. Section 6 concludes this work and results.

2. RELATED WORK

Extensive literature on visual object detection and tracking algorithms with fixed cameras exist, but only few papers transform the image coordinates to real world coordinates and analyse the position accuracy in relation to distance between camera and object. Also, almost all the papers that reported accuracies, lack accurate ground truths.

Object detection and object localization are the most fundamental components of video-based positioning. Before the recent rapid development in deep learning, background subtraction [21] has been a very popular method for object detection and tracking. However, it has several problems, especially in dynamic conditions. Its reliability decreases e.g. with lighting effects, partly overlapping objects, slow-moving objects, shadows, and effects from a moving scene such as swaying trees. There has been intensive work to diminish these drawbacks, e.g. adaptive Gaussian mixture models in [23] and a multi camera approach in [24], but deep learning is still more promising for object detection and tracking. For example, deep learning-based object detection can perform well in very dynamic conditions, thus it is commonly used with mounted camera applications (see [25] for an extensive review on the topic).

Next, we introduce papers which are only closely related to our work: In [24] authors presented a people tracking scheme using multiple cameras. They used a combination of multiple cameras and a planar homography constraint idea to tackle the drawbacks of the used foreground segmentation method and to improve the tracking challenges caused by occlusions. Their outdoor tests showed that by increasing camera views for the same scene and applying the homography constraint the detection yields significantly less false positives. While we localize objects using their root position on ground plane, in [24] locations of people on the ground plane are detected but no information on the localization is given.

In [26], authors studied a similar concept as in [24], but extended the ground plane detection to multiple layers. In addition, they used *Scale Invariant Feature Transform* (*SHIFT*) to identify people and also applied Kalman Filter [27] to predict an object's position in the next video frame. The authors reported an average localization error of 13 cm for indoor dataset PETS2007 [28], which have ground truth positions manually labeled from video images. Thus, there is obviously some discrepancy and the real accuracy of the proposed system remains uncertain.

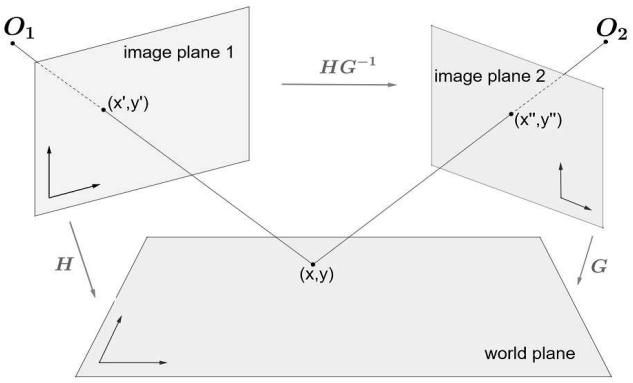

Figure 1. Example of planar homography transforms.

Einsiedler et al. in [29] developed a surveillance camera-based positioning system to detect and localize vehicles and pedestrians in a parking garage. The authors used a cascade of boosted Haar-like feature classifiers [30] for the vehicle detection and *Histogram of Oriented Gradients* (*HOG*) features to detect pedestrians [31]. They applied the GrabCut algorithm [32] to separate an object from the background to determine the object's root point. The root point was mapped to real world using a homography transform. Detection error was estimated by manually annotating the correct root points and comparing with detector results after transforming the points to real world floor plane. They report that on single camera view with 20m coverage and in 95 percentile the overall positioning error was below 0.9m.

Authors in [33] also developed a surveillance camera-based positioning system for parking garage. Their system uses multi-camera for the same scene to handle occlusions and to improve the localization accuracy. They used background subtraction method with several additional image processing steps to detect and segment moving objects. In addition, object tracking was implemented and the Alpha-beta-filter [34] was utilized to smoothen the movement trajectory. In the experiments they used LIDAR system [35] for reference positions. The LIDAR system has been determined to produce 0.19m mean Euclidean distance compared to Differential-GPS. The authors report 0.24m mean position error for the developed system.

In [36], authors introduced a camera-based pedestrian tracking system with IMU fusion. IMU was utilized to assist vision-based localization during visual occlusions. They implemented a scene-specific trained Support *Vector Machine* (*SVM*) and HOG pedestrian detector. The system was tested outdoors and the ground truths were created by hand annotating from each video frame. The average localization error when only visual tracking was used was about 0.9m. The distance from camera to object was not defined, but based on screenshots, the distance could be approximately up to 50m.

In [14] and [37] Yan et al. introduced an IMU-based indoor positioning system aided by camera positioning. In camera positioning they utilized Faster R-CNN [38] deep learning architecture for object detection and object localization. Authors used the middle point of the bottom boundary of extracted bounding box as the object root point. This is a similar principle to the one used in this paper. However, authors used a simple pinhole model to calculate a distance and exploited the camera positioning only to estimate distances between camera and object, and tested their setup only in narrow corridors. In contrast, we use CNN for 2D positioning. Authors reported in [37] average accuracy of 0.16m with fused positioning, but did not define the used reference position system.

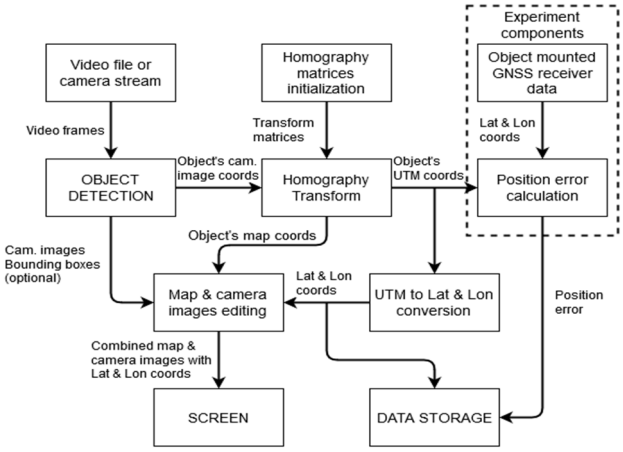

Figure 2. Components of the implementation.

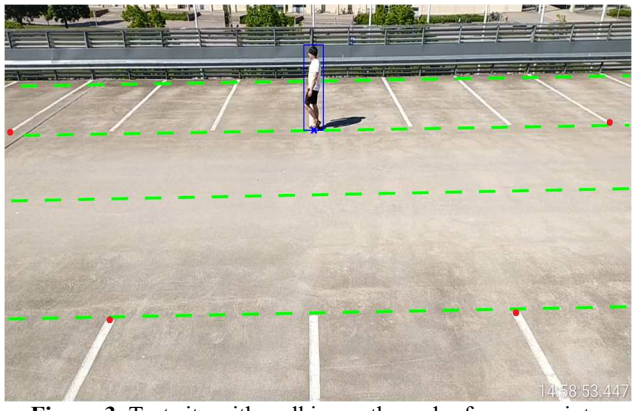

Figure 3. Test site with walking paths and reference points.

3. PROPOSED METHODS

In this paper, we propose several methods to be used for implementing a surveillance camera-based positioning system. The first part is the object detector, which detects the objects and locations in the camera image. Next, by using a homography transform the objects image coordinates are transformed to UTM-coordinates and further to more generally used geodetic latitude and longitude coordinates.

3.1 CNN detector

As discussed earlier, significant progress has recently been made for deep learning-based detectors, but those architectures have also become large and expensive because of the deep network structure. The deep learning-based detector is potentially more suitable for dynamic environment surveillance, but it comes with increased computational demand. Although, the surveillance camera-based positioning system could run on a dedicated computer, increasing the number of camera views will increase the computing power demand proportionally. Thus, to employ a deep learning-based detector for multi-view, one must carefully select the convenient architecture, with possible hardware acceleration (GPU) support, for the available computing power.

Detectors are mainly divided into one-stage and two-stage implementations. Two-stage detectors have generally more advantages in accuracy and precision, but in detection speed, the one-stage detectors have better performance. An example of one-stage detectors is YOLO architecture series [39]. In this work, we employed the recent YOLOv4 architecture [40]. YOLOv4 is mainly designed for fast operating speed, but still can provide good accuracy, and it is reported to outperform prior deep learning models in terms of its real-time performance [40]. YOLOv4 network model includes three main parts: backbone, neck, and head. The backbone is taking the input images and extracts feature maps. In the neck part, the extracted features are specifically enhanced to improve robustness and discriminatory aspects. Finally, the head is used to predict object classes and bounding box coordinates based on the enhanced features. In our implementation the middle point of the bottom boundary of the extracted bounding box is defined as the object root point (ground surface position) in the camera image.

3.2 Homography transform

Homography is a planar projective transformation, which is a linear transformation into homogeneous coordinates [41]. Planar homography maps points from one plane to another. The transformation is represented by a non-singular 3 x 3 matrix and the point projection from plane to another can be defined by

$$w\begin{bmatrix}x\\y\\1\end{bmatrix} = H\begin{bmatrix}x'\\y'\\1\end{bmatrix} = \begin{bmatrix}h_{11} & h_{12} & h_{13}\\h_{21} & h_{22} & h_{23}\\h_{31} & h_{32} & h_{33}\end{bmatrix}\begin{bmatrix}x'\\y'\\1\end{bmatrix},$$

where w is a scaling factor. H is usually called homography matrix or projective matrix. The matrix, has 8 degrees of freedom as it is generally normalized by $h_{11}^2 + h_{12}^2 + h_{13}^2 + h_{21}^2 + h_{22}^2 + h_{23}^2 + h_{31}^2 + h_{32}^2 + h_{33}^2 = 1$. To obtain the homography matrix, only four point pairs are needed from the two planes (total of eight points). Then a Direct Linear Transformation (DLT) algorithm can be used for solving the H [41]. Examples of planar homography transforms are given in Fig. 1, where image plane 1 and image plane 2 can be considered as camera views, O_1 and O_2 are camera centers and world plane is a ground surface. Point (x',y') in image plane 1 is transformed to a point (x,y) in world plane by using homography matrix H. Similarly, image plane 2 point (x",y") is transformed to the same world plane point (x,y) by using a homography matrix G. Homography transforms are reversable (e.g. H^{-1}) and can be concatenated (e.g. the HG^{-1} in figure).

Thus, a point in the camera image can be projected to real world coordinates by the homography. Generally, in planar homography, the target coordinate system can be any planar and linear coordinate system. For example, local area coordinates, map image coordinates or projected coordinate system.

3.3 Geographic coordinate conversion

Nowadays, in most GNSS applications and with common digital maps the position is defined by geodetic coordinates, latitude and longitude (and sometimes altitude) [42]. Latitude and longitude are angular coordinates measured in degrees, which is not suitable as target coordinates in the planar homography. In order to transform image coordinates to global coordinates using the homography, a *Projected Coordinate System* (*PCS*) can be used [43]. The PCS is generally based on some *Geographic Coordinates System* (*GCS*), but instead of sphere or spheroid the PCS is defined on a flat two-dimensional surface, i.e. a plane.

One example of PCS is *Universal Transverse Mercator* (*UTM*) [43], which is an implementation of transverse mercator projection. UTM is one of the most common conformal mappings in geodesy today. It is a multiple map projection in which the Earth is divided into 60 zones (numbered 1-60) in the north-south direction. Each zone is 6 degrees wide (in longitude) segment of the earth. The central meridian is in the middle of each zone and is used as a northing coordinate-axis and as zero for easting coordinates. Coordinates conversion between UTM and geodetic coordinates can be accurately defined [44]. The accuracy depends on algorithm

Table 1. Mean positioning error of the proposed system and a standard deviation of the error.

Camera distance	Mean position error	Standard deviation	Mean error direction [0,360°]	Camera direction [0,360°]	Corrected mean position error (vertical camera pixel offset)
5m	0.162m	0.067m	117°	117°	0.109m (-14)
8m	0.145m	0.072m	113°	118°	0.107m (-5)
11m	0.160m	0.069m	114°	117°	0.109m (-3)
15m	0.189m	0.087m	101°	118°	0.156m (-2)

and number of used terms in calculations, and even nanometer-level accuracy can be achieved [45],[46].

4. IMPLEMENTATION

4.1 Implemented positioning system

The implemented surveillance camera-based object positioning system supports video files or alternatively live IP-camera streams as inputs; thus, the system can run online or offline. The implementation was coded with Python programming language. The architecture of the implementation is shown in Fig. 2. The main component of the system is the object detection that supports multiple camera views and can detect multiple objects from each view. Core of the object detection is the object detector. A publicly available pretrained realization of YOLOv4 detector was used for this implementation [40]. The object detector provides detected object's camera image coordinates for the homography transform component and simultaneously forwards camera images with bounding boxes (if requested) to the image editing component. For homography transform, a map image is given as an input and when initializing the system, four corresponding points are defined between the map image and each camera view. In addition, latitude and longitude coordinates are given for all the four defined camera image reference points. Internally, the given geodetic coordinates are then converted to UTM coordinates to allow homography transform. The homography matrices are solved between each camera view and the UTM coordinate system, and between the views and the map image. The homography transform component transforms each received camera image coordinate to map image coordinates and to UTM coordinates. The resulting UTM coordinates are then converted to latitude and longitude for visualization and recording. The image editing component combines camera views and the map image into one image and draws detected object's position markers with corresponding geodetic coordinates on to the map image to visualize the positions. Finally, the edited images are shown for the user as a video stream.

Lenovo Thinkpad P1 Gen2 laptop equipped with Intel i7-9750H and low-end Nvidia Quadro T1000 3GB GPU was able to run the implemented one camera test system with a frame rate of about 10 fps.

4.2 Experimental setup

The implemented positioning system was tested also indoors, but to determine the positioning accuracy unambiguously, the main experiments were carried out in outdoors to enable RTK GNSS ground truth positions. A multi-frequency U-blox ZED-F9P RTK GNSS receiver [48] was utilized to provide very accurate ground truth positions. ANN-MB-00 active GNSS antenna was used, which allows centimeter-level accuracy with the ZED-F9P module [49]. Top floor of a parking garage was selected as the test site to allow open sky for the GNSS receiver and a planar ground surface. A mobile phone was used to feed RTCM correction data stream to

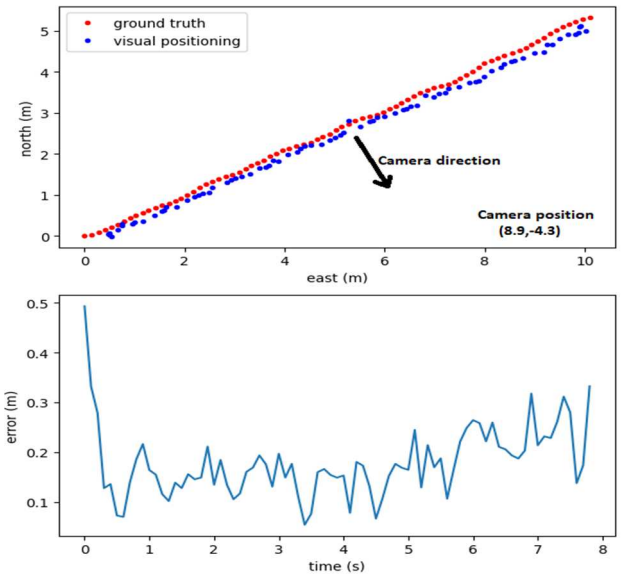

Figure 4. Positioning trajectories, and absolute positioning error.

the RTK receiver, in addition the phone also recorded the geodetic positions from the RTK receiver via Bluetooth connection using 10 Hz sample rate. Xiaomi Redmi Note 8T mobile phone with an IP-camera application was imitating the IP-surveillance camera. Only one camera was used to evaluate the positioning accuracy of a single camera. Recorded video resolution was 1280×720, however the image size for the object detector was resized to 512×512. The camera was attached at 3.5 meters above floor level. The test site is shown in Fig. 3. A person was acting as the target object to be positioned and the GNSS antenna was placed on the top of the person's head. The person walked back and forth at four different distances across the camera view. The perpendicular distances between camera and walking paths were: 5m, 8m, 11m, and 15m. The walking paths are visualized as (green) dashed lines in Fig. 3. The four reference geodetic coordinates for homography are marked with (red) dots. The object's position in the camera image is defined by the center coordinate of the bottom horizontal line of a bounding box.

5. EXPERIMENTAL RESULTS AND ANALYSIS

Accuracy results of the implemented positioning system are shown in Table 1, where position error is the Euclidean distance between the ground truth (RTK GNSS) geodetic position and the position estimation by proposed system. Results show that mean positioning error was only 14.5-18.9 cm and standard deviation of the error was 6.7-8.7 cm. Thus, the accuracy is significantly better than e.g. in general GNSS solution. The position error was systematically biased towards the camera. This is expressed in Table 1 by mean error direction in relation to camera direction. The camera direction is the direction from the center coordinate of the walking path towards to camera, this is visualized also in Fig. 4. Here direction is defined as 0° is to North, 90° to East, 180° to South and 270° to West.

In Fig. 4, ground truth and estimated position trajectories of one-way walk at 8m distance are presented as an example. In the bottom diagram of the Fig. 4 the absolute error of the estimated positions is shown. In the top diagram the positions are normalized from UTM-coordinates so that ground truth starting point is the origin. From the trajectories it is clearly visible that position estimates are biased to camera direction. The position bias is caused by the fact that the bottom bounding box limit (vertical coordinate) from the object detector is not the correct horizontal

root or center point of the person at ground level. The horizontal center point should be the middle point between person's feet, but the CNN detector extracts the lower bounding box limit from the bottom edge of the foot whichever is closer to camera. This error is intrinsically 15-20cm. If the person is walking towards or away from the camera, the bias error is expected to be quite similar. To correct that bias, we iteratively searched the best integer pixel offset for object's vertical camera coordinate to minimize the mean error; the results are shown in the last column of Table 1, mean position error was reduced to 10.7-15.6 cm, which is less than any related work. The pixel offset is in parenthesis, where minus values implies that coordinate of some higher pixel in the image is used in homography instead of the original provided by the object detector (in a camera image the origin is at top-left).

In the absolute error diagram at the bottom of Fig 4, a periodicity can be detected. The periodicity is caused by person's gait. The upwards and downwards moving feet causes fluctuation on the bottom bounding box limit. This fluctuation could be smoothened if some filter, e.g. Kalman filter, is employed for the trajectory. Another observation is that the position error is greatest in the path ends. The object detector can detect also partial objects correctly at camera image edges, which causes the object's root point to be detected with some error if the object is only partially visible. A simple solution would be to discard the detections that are close to image edges. Alternatively, some position error correction functionality for image edge areas could be implemented.

6. CONCLUSIONS

This paper presented an implementation of surveillance camera-based positioning system employing CNN-detector. The implemented system can provide geodetic coordinates simultaneously for all the visible predefined object types in multiple camera views. The absolute accuracy of CNN-based fixed camera positioning has remained unclear; thus, we tested the accuracy precisely using a RTK GNSS receiver to provide unambiguous ground truth positions. The tests were repeated in several distances, while a person was used as a target object. The mean positioning error was in range of 14.5-18.9 cm and a standard deviation of the error was 6.7-8.7 cm. It was determined that the mean error was systematically biased towards the camera. To correct the bias, we added offset to the detected vertical camera coordinate, which reduced the mean positioning error to 10.7-15.6 cm. The evaluated positioning error is less than any related work we found. Other error sources were also discussed and some suggestions for error reduction were given. In the future, we aim to extend this scheme to moving platforms.

ACKNOWLEDGMENT

This Research was financially supported by Business Finland as part of the AUTOPORT project. More details on the project and partners are available at [50].

REFERENCES

[1] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: a survey," GeoInformatica, vol. 19, pp. 525-565, 2015.

[2] B. Hofmann-Wellenhof, H. Lichtengger, and E. Wasle, GNSS: GPS, GLONASS, Galileo & more, Springer Wien New York, 2007.

[3] F. Alkhawaja, M. Jaradat, and L. Romdhane, "Techniques of indoor positioning systems (IPS): A survey," in Proc. Adv. Sci. Eng. Technol. Int. Conferences (ASET), Mar. 2019, pp. 1–8.

[4] J. Collin, P. Davidson, M. Kirkko-Jaakkola, and H. Leppäkoski, "Inertial Sensors and Their Applications," 2013, Handbook of Signal Processing Systems. Bhattacharyya, S., Deprettere, E., Leupers, R. & Takala, J. (eds.). 2 ed. New York, NY, USA: Springer, pp. 69-96.

[5] P. Srinivas and A. Kumar, "Overview of architecture for GPS-INS integration," in Proc. Recent Develop. Control, Automat. Power Eng. (RDCAPE), Oct. 2017, pp. 433–438.

[6] P. Davidson and R. Piché, "A survey of selected indoor positioning methods for smartphones," IEEE Communications Surveys and Tutorials, 19:2, 1347-1370, 2017.

[7] R. Mautz and S. Tilch, "Survey of optical indoor positioning systems," in Proc. Int. Conf. Indoor Positioning Indoor Navigat., Mar. 2011, pp. 1–7.

[8] L. Mainetti, L. Patrono, and I. Sergi, "A survey on indoor positioning systems," 22nd Int. Conf. on Software, Telecommunications and Computer Networks, Split, Croatia, 17–19 Sept. 2014, pp. 111–120.

[9] L. Jiao and J. Zhao, "A Survey on the New Generation of Deep Learning in Image Processing," IEEE Access, vol. 7, pp. 172231-172263, 2019.

[10] Elhassouny and F. Smarandache, "Trends in deep convolutional neural Networks architectures: a review," 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), pp. 1-8, 2019.

[11] J. Liu, Y. Gu, and S. Kamijo, "Customer Behavior Recognition in Retail Store from Surveillance Camera," in IEEE International Symposium on Multimedia (ISM), pp. 154–159, 2015.

[12] M. Popa, L. Rothkrantz, Z. Yang, P. Wiggers, M. Popa, R. Braspenning, et al., "Analysis of shopping behavior based on surveillance system," Proc. of IEEE International Conference on Systems Man and Cybernetics, pp. 2512-2519, 2010.

[13] W. Jiang and Z. Yin, "Combining passive visual cameras and active IMU sensors for persistent pedestrian tracking," J. Vis. Commun. Image Represent., vol. 48, pp. 419-431, Oct. 2017.

[14] J. Yan, G. He, A. Basiri, and C. Hancock, "3-D passive-vision-aided pedestrian dead reckoning for indoor positioning," IEEE Trans. Instrum. Meas., vol. 69, no. 4, pp. 1370-1386, Apr. 2020.

[15] S. Adler, S. Schmitt, K. Wolter, and M. Kyas, "A survey of experimental evaluation in indoor localization research," in Proc. Int. Conf. Indoor Positioning Indoor Navigat., Oct. 2015, pp. 1–10.

[16] M. Asaduzzaman, T. K. Geok, S. Sayeed, Md A. Bari, F. Hossain, and T. C. Peng, "A Comparative Survey on Indoor Object Location Tracking Techniques and Technologies," IEEE 10th Int. Conf. on System Engineering and Technology (ICSET), Shah Alam, Malaysia, Nov. 2020.

[17] D. B. Ahmed, L. E. Díez, E. M. Diaz and J. J. G. Domínguez, "A Survey on Test and Evaluation Methodologies of Pedestrian Localization Systems," IEEE Sensors Journal, 20(1), pp.479-491, Sept. 2019.

[18] F. Zafari, A. Gkelias, and K. K. Leung, ''A survey of indoor localization systems and technologies,'' IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2568–2599, 3rd Quart., 2019.

[19] X. Guo, N. Ansari, F. Hu, Y. Shao, N. R. Elikplim and L. Li, "A survey on fusion-based indoor positioning," IEEE Commun. Surveys Tuts., vol. 22, pp. 566-594, 1st Quart. 2020.

[20] L. Mainetti, L. Patrono, and I. Sergi, "A survey on indoor positioning systems," Proc. 22nd Int. Conf. Softw. Telecommun. Comput. Netw. (SoftCOM), pp. 111-120, 2014.

[21] Sen-Ching S. Cheung, and Chandrika Kamath, "Robust Techniques for Background Subtraction in Urban Traffic Video," EURASIP

Journal on Applied Signal Processing, vol. 2005, pp. 2330-2340, Jan. 2005.

[22] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3-D machine vision based metrology using off-the-shelf TV cameras and lenses," IEEE Journal on Robotics and Automation, 3(4): 323-344, 1987.

[23] C. Stauffer, W. Eric, and L. Grimson, "Adaptive background mixture models for real-time tracking," Computer Society Conf. on Computer Vision and Pattern Recognition 1999.IEEE., vol. 2, 1999.

[24] S.M. Khan and M. Shah. "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in ECCV, pages IV: 133–146, 2006.

[25] L. Aziz, Md. Salam, U. Sheikh, and S. Ayub, "Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review," IEEE Access, Vol. 8, pp. 170461 - 170495, 2020.

[26] D. Arsié, B. Schuller, and G. Rigoll, "Multiple Camera Person Tracking in Multiple Layers Combining 2D and 3D Information," in Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008, Oct 2008, Marseille, France.

[27] Kalman, R.E., "A new approach to linear filtering and prediction problems," in transactions of the ASME Journal of Basic Engineering (1960) 35–45

[28] J. Ferryman and D. Tweed, "An overview of the pets 2007 dataset," in Proceeding Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2007, IEEE, Rio de Janeiro, Brazil.

[29] J. Einsiedler, D. Becker, and I. Radusch, "External visual positioning system for enclosed carparks," in Positioning, Navigation and Communication (WPNC), 2014 11th Workshop on, pp. 1–6. IEEE, 2014.

[30] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, pages 511-518, 2001.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005., volume 1, pp. 886-893 vol. 1, 2005. (Pubitemid 43897286).

[32] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in ACM Transactions on Graphics (TOG), volume 23, pages 309-314, 2004.

[33] A. Ibisch, S. Houben, M. Michael,R. Kesten, and F. Schuller, "Arbitrary object localization and tracking via multiple-camera surveillance system embedded in a parking garage," 9407. DOI: 10.1117/12.2075528, 2015.

[34] R. Penoyer, "The alpha-beta filter," The C Users Journal 11, pp. 73–86, July 1993.

[35] A. Ibisch, S. Stumper, H.Altinger, M. Neuhausen, M. Tschentscher, M. Schlipsing, J. Salmen, and A. Knolls, "Autonomous driving in a parking garage: Vehicle-localization and tracking using environment-embedded lidar sensors," in Proceedings of the IEEE Intelligent Vehicles Symposium, pp. 829 – 834, 2013.

[36] W. Jiang and Z. Yin, "Combining passive visual cameras and active IMU sensors for persistent pedestrian tracking," Journal of Visual Communication and Image Representation, Vol. 48, pp. 419-431, ISSN 1047-3203, 2017.

[37] J. Yan, G. He, A. Basiri, and C. Hancock, "Indoor pedestrian dead reckoning calibration by visual tracking and map information," in Proc. Ubiquitous Positioning Indoor Navigat. Location-Based Services (UPINLBS), pp. 1-10, May 2018.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91-99.

[39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified real-time object detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.

[40] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.

[41] Hartley, Richard and A. Zisserman, Multiple View Geometry in Computer Vision, Second edition. Cambridge, UK, Cambridge University Press, 2003. Print.

[42] B. Hofmann-Wellenhof, H. Lichtenegger, and E. Wasle, GNSS – Global Navigation Satellite Systems. Wien: Springer, (2008).

[43] L. Zhiping, Q. Yunying, and Q. Shubo, Geodesy – Introduction to Geodetic Datum and Geodetic Systems. Springer, 2014.

[44] The Universal Grids and the Transverse Mercator and Polar Stereographic Map Projections. NGA.SIG.0012_2.0.0_UTMUPS (2014). Available online: https://earth-info.nga.mil/GandG/update/coordsys/resources/ NGA.SIG.0012_2.0.0_UTMUPS.pdf [accessed on Mar 2021].

[45] C. Karney, "Transverse Mercator with an Accuracy of a Few Nanometers," J. Geodesy, vol. 85, no. 8, pp. 475-485, 2011. E-print: arXiv:1002.1417.

[46] C. Enríquez, "Accuracy of the coefficient expansion of the Transverse Mercator Projection", International Journal of Geographical Information Science, 18:6, 559-576, 2004.

[47] https://github.com/hunglc007/tensorflow-yolov4-tflite [Mar 2021].

[48] Multi-band RTK receiver: Ublox C099-F9P application board, https://www.u-blox.com/en/product/c099-f9p-application-board.

[49] A. Krietemeyer, H. van der Marel, N. van de Giesen, and M.-C. ten Veldhuis, "High Quality Zenith Tropospheric Delay Estimation Using a Low-Cost Dual-Frequency Receiver and Relative Antenna Calibration," Remote Sens. 2020, 12, 1393. https://doi.org/10.3390/rs12091393.

[50] https://autoport.fi