AUGMENTED REALITY ASSET TRACKING USING HOLOLENS

J.I.Fan, K.Khoshelham

Dept. of Infrastructure Engineering, The University of Melbourne, Parkville 3010 Australia

KEY WORDS: Asset Tracking, Asset Mapping, Asset Detection, Augmented Reality, Indoor Mapping, Smart Glasses, Photogrammetry.

ABSTRACT:

Asset Tracking is an essential component of building management process. It involves creating and maintaining a database of detailed information of assets such as location, condition, brand, and type. This information can help building owners make informed decisions for cost-effective maintenance of building assets. Existing approaches to perform asset tracking require a manual process of measuring and recording the asset condition and location, which is labour-intensive and costly. The typical approach usually includes a human operator with pen and paper inspecting the site and manually recording the information about the asset. In this paper, we propose an augmented reality asset tracking system using HoloLens to reduce the manual labour involved in this process. The system can automatically detect the asset, record and update its related information by visual inspection. Assets are detected by feeding images captured by the HoloLens built-in camera to a pre-trained object detection network. Using a combination of various sensor readings from the HoloLens, the system can estimate the location of the asset using visual simultaneous localization and mapping (vSLAM). This information is then viewed and verified by the user using the augmented reality user interface. Upon the user confirmation, this information will be uploaded to a database. As a case study, we demonstrate a vending machine tracking system which is able to detect and localise the vending machines in an indoor environment and create a database of vending machine information. The system can detect vending machines with a mean average precision of 94.8% and a localization accuracy of 2.3 meters without pre-screening or user input.

1. INTRODUCTION

Asset management is essential for building owners and facility managers to perform efficient and cost-effective maintenance operations. Whenever an asset is installed, relocated, or removed, this information needs to be collected and stored to ensure the area continues to function within design parameters. The system used to accurately track these large sets of assets is called asset tracking system. The traditional approach to asset tracking involves a human operator with pen and paper manually localizing assets and recording its information on paper. This process is labour-intensive and error-prone due to its manual nature. Automated approaches using barcode and RFID (radiofrequency identifier) tags are expensive and the total cost of setting up such systems limits their practical application (Angeles 2005). In consideration of these disadvantages, we propose a more efficient and low-cost asset tracking system based on HoloLens.

To overcome the limitations of the RFID based asset tracking systems, most research efforts focus on using smartphones to develop alternative low-cost asset tracking solutions (Chen et al., 2010; Kostoeva et al., 2019). The successful use of deep learning method allows the asset tracking system to understand the scene and identify assets without pre-installed tags or barcode. This leads to a reduction in the hardware cost of these asset tracking systems. However, smartphone-based asset tracking systems face challenges in accurately localizing the asset. Due to the limitation of smartphone sensors, indoor localization methods used by the smartphone-based asset tracking systems require an additional manual process or a complete survey of the environment beforehand. For example, the WiFi-fingerprint based localization method used by Chen et al. (2010) can only provide room-level accuracy where the system is able to locate the room that object is located but cannot accurately locate the object in the room.

In consideration of these disadvantages, we propose a more efficient and low-cost asset tracking system based on HoloLens. HoloLens is a wearable device that embeds electronic components such as a display screen, an integrated processor, and a suite of sensors. HoloLens allows visualization and interaction with an interface that is displayed through the lens. It provides the user with a simultaneous view of both physical and digital worlds (Belkacem et al, 2019), where information in the digital world is overlaid on the view of the physical world. The augment reality technology can improve the process for the user to interact with the asset tracking system. Compared to other devices which can provide AR experience such as smart phones, HoloLens is equipped with a powerful processor and the spatial mapping capability. This allows HoloLens to utilize imagery captured by both RGB camera and depth camera mounted on the device to build a spatial map of the nearby physical environment. This information can be used to calculate the location of each recognized asset. The user can see the result of the asset tracking overlaid on their view to have a clear understanding of the location of the detected asset. Despite the advantaged of HoloLens for asset tracking, there are still questions around the feasibility of this approach. For asset detection, we need to carefully consider an object detection method that can run on a resource limited device. For asset localization, since GNSS cannot be used in indoor environments an alternative method is needed to estimate the location of the asset.

The proposed system takes advantage of the spatial mapping capability of HoloLens to accurately localize the assets in an unknown environment automatically. The system is designed based on HoloLens due to the fact that it is one of the most powerful smart glasses available in the market. To truly leverage the low-cost aspect of this asset tracking system design, other low-cost smart glasses with depth camera, RGB camera and ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume V-4-2021 XXIV ISPRS Congress (2021 edition)



Figure 1. Conceptual framework of the vending machine asset tracking system

access to internet can be used to reduce the setup cost of asset tracking.

In this paper, we describe the asset detection, localization, and verification methods, and perform experiments to evaluate the proposed asset tracking system. We provide evaluation results for a vending machine tracking system implemented based on the design in terms of detection accuracy, localization accuracy, power consumption, processing time and system usability.

2. ASSET TRACKING SYSTEM DESIGN

In our asset tracking system, the HoloLens runs a mixed reality application which can automatically detect and localize the assets and provide a graphical user interface for user to verify, modify, and confirm the result of the detection. The application can be divided into three different modules. After user starts the application, the HoloLens will start recording using the front camera. The asset tracking system will pull the latest frame from the video recording and sends it to the asset detection module. The result from the asset detection module will then feed into the asset localization module to calculate the coordinates of the asset within the spatial map of the environment. Finally, the user interaction module will let the user modify, accept, or decline the result before it can be sent to the database. A graphical representation of the design of the asset tracking system can be seen in Figure 1.

2.1 Asset Detection

The asset detection module is responsible for finding the asset in the image captured by the front camera of HoloLens which represents the user's current field of view. The asset related information that we are interested to extract from the RGB image includes the asset's type and its location in the image. The asset detection module will computer the bounding box and determine the type for each asset in the image using a pre-trained deep convolutional network. The asset detection module simulates a human inspector's behaviour: when the inspector comes to the building, they will start looking for assets within their sight. After the processing, the asset detection module will feed the resulting assets' type and position in the image to the asset localization module. The task of finding assets' type and position is called object detection. The current state of the art in object detection is the deep learning-based approach. The robust training algorithms allow deep convolutional networks to learn informative object representations without having to manually design features as compared to traditional approaches (Goodfellow et al., 2016).

The major challenge in deploying deep learning based methods on HoloLens is that HoloLens has limited computational resources as compared to a desktop computer with dedicated graphic processing unit. Achieving good performance using deep learning requires a large amount of computational resources for larger and deeper networks (Huang et al., 2018). Thus, an important trade-off between the accuracy and the computational complexity of the deep network needs to be made. The detection network used in this research is a Yolo-v2 tiny (Redmon et al., 2017) with some modification to the network structure. It is a reduced size version of Yolo-v2 which can run on mobile devices. The final version of the network is only 42 MB in size, making it suitable for running on HoloLens. The architecture of the detection network is shown in Figure 2.



Figure 2. Architecture of the detection network

To train the network, we used the few-shot learning technique. (Kostoeva R. et al., 2019). Due to the limitation of the number of vending machine samples available for training, we used multiple views of the same instance for training. Images taken from different angles and pixel size were used. We manually labeled the type and ground truth bounding box for each asset that appeared in each image.

2.2 Asset Localization

The asset localization module consists of two computation steps. First, it takes the bounding box coordinate as input from the asset detection module and computes the asset's location in the 3D world relative to the camera. Then, it takes the coordinates from the previous step and transforms these to a universal coordinate system defined by the spatial map constructed by Hololens. An intuitive visualization of this transformation can be seen in Figure 3. These two computation steps will be explained separately in the following sections.



Figure 3. Transformation between world and camera coordinate system (Tola 2005)

The first computation is a reverse of camera projection. It is done based on Equation 1 where image coordinates are used to calculate the object coordinates in the camera space:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix},$$
 (1)

where u,v = image coordinates

X,Y,Z = object coordinates in camera space $f_x, f_y, c_x, c_y =$ camera intrinsic parameters

The asset detection module draws a bounding box around the detected asset. The coordinates of the bounding box will be fed into asset localization module as input. The module will project a ray from each corner of the bounding box through the perspective centre of the camera to compute the location of the bounding box in the camera's coordinate system. To find the exact place of the asset in camera's coordinate system, we need to estimate the scale factor s which depends on the distance between the asset and the camera. This distance is estimated by the width of the vending machine in the image. Due to the short range of the depth camera, the depth information is not used. As a workaround, we use the prior knowledge on the size of the vending machine and estimate the average width of a vending machine in the image as 940 pixels when it is 2 meters away from the camera. Using this information, we can calculate the position of the vending machine in camera's coordinate system:

$$depth = \frac{averagePixelForVendingMachineAt2M}{boundingBoxWidth} \times 2, \qquad (2)$$

The next step is to compute the asset's location in a universal coordinate system. In this project the world coordinate system has its origin at the camera position when the user starts the application. The camera coordinate system for the first image taken by the front camera will be used as world coordinate system for the current asset tracking session. The spatial mapping algorithm in Hololens estimate the camera extrinsic parameters for all subsequent images and constructs a map of the environment. The detailed process of computing the extrinsic parameters which provides the transformation between the camera coordinate system and the world coordinate system is Microsoft's proprietary software and its algorithms are unpublished (Khoshelham et al., 2019). But based on publications by Microsoft researcher (Nießner et al., 2013; Glocker et al., 2014), the spatial mapping technology is based on RGB-D SLAM algorithm, where RGB-D data captured by the depth sensor onboard Hololens are used to construct a map of the environment while simultaneously keeping track of the device's location within it in real time (Mur-Artal et al., 2017).

Based on above computations, the position of the asset is calculated which can be used to localize the asset. Recording the position and other relevant information about the asset is done through an interface which facilitates the storage of the recorded information in an asset database.

2.3 User Interaction

The main design idea behind the augmented reality interface for this asset tracking system is to make it easy to use and easy to understand. We expect that the design elements can explain themselves during the experience. When the user uses the asset tracking system, they will perform intuitive actions to get their desired outcome. We try to give minimum instruction to the user via a tutorial but using intuitive hints by colour and simple words. Figure 4 shows a view of the user interface for a detected vending machine. The information provided to the user can be seen in the bottom blue square, and include capture timestamp, object category, confidence level and object coordinate in meters.



Figure 4. User interface of the asset tracking system

3. EVALUATION METHOD

We designed an experiment where a user performed the complete asset tracking process for all vending machines in Baillieu Library at the University of Melbourne. The user was asked to complete the asset tracking process in a traditional approach using pen and paper. This enabled us to make a comparison between the two different approaches. The asset tracking system was also deployed, and the results were recorded for evaluation. Unfortunately, due to COVID-19 restrictions, access to the University campus was severely limited. Therefore, the evaluations were partly done at home without further access to the vending machines or the test environment.

3.1 Power Consumption

Energy consumption is one of the most constraining limitations of any wireless devices and their applications. Most modern devices are built with irreplaceable batteries which means if the device run out of the power, it must wait until it is charged again. Thus, power consumption directly associated with the amount of time that an application can be run for.

To estimate the power consumption of the application, we measure the power consumed by the application for 5 minutes then use the following formula to calculate the estimated run time in minutes for the application. The experiment was repeated for multiple times and the average run time was obtained to avoid bias.

estimated run time in minute
$$=\frac{t}{p} \times 100$$
, (3)

where t = time spent in minute p = power consumed in percentage

3.2 Asset Detection Accuracy

This section will focus on the evaluation of the object detection network used in the asset tracking application. Due to the pandemic situation, the evaluation was based on previously captured data instead of vending machines in the test environment. Precision, recall, mean average precision and mean intersection over union were calculated using test data.

$$Precision = \frac{True \ Positive}{True \ Positive + False \ Positive}, \quad (4)$$

$$Recall = \frac{True \ Positive}{True \ Positive+False \ Negative}, \tag{5}$$

The average precision score summarizes a precision recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight in Equation 6.

$$AP = \sum (R_n - R_{n-1})P_n, \qquad (6)$$

where

AP = average precision

$$R_n$$
 = recall score at nth threshold
 P_n = precision score at nth threshold

Intersection over union describes the localization accuracy of the predicted bounding box compared to the ground truth bounding box. In the Figure 5, we can see a visualization of the ground truth bounding box and predicted bounding box overlaid on the original image. The ground truth bounding box is drawn in white and the predicted bounding box is drawn in red. The intersection



Figure 5. An example of vending machine detection in an image. The ground truth bounding box is drawn in white and the predicted bounding box is drawn in red.

over union is calculated using the following formula, based on the intersection and the union of the predicted bounding box and the ground truth.

$$IoU = \frac{Intersection}{Union},\tag{7}$$

Equation 7 demonstrates how to calculate IOU in the simplest scenario. In some test images there might be multiple bounding boxes and multiple objects appearing in the image. The number of predicted bounding boxes may not necessary match the number of ground truth bounding boxes. Figure 6 shows an example on object detection result where multiple objects are present in an image. The predicted bounding box is drawn in red while the ground truth bounding box is drawn in white. The number of vending machines present in the image is 5 while the



Figure 6. An example of detecting multiple vending machines in an image. The predicted bounding boxes are drawn in red while the ground truth bounding boxes are drawn in white.

detection network predicted 7 vending machines in the image. From a human's perspective, we can see that the object detection network mistakenly identifies the wall and the content of vending machine as vending machines. Computing the intersection over union in this case is difficult since one ground truth bounding box may have multiple predicted bounding boxes intersecting with it. The solution we use for this problem is based on a greedy approach. For each ground truth bounding box, we compute the intersection over union for every predicted bounding box and take the maximum value as the intersection over union for that ground truth bounding box. The intersection over union is calculated on a separate test set.

3.3 Asset Localization Accuracy

To evaluate the asset localization accuracy, we ask a user wearing HoloLens to perform asset tracking in the environment. To measure the accuracy, we map the result of asset tracking system on to the floor map and calculate the difference in the coordinates from the ground truth. Unfortunately, due to the pandemic situation, access to the university campus was restricted and therefore we established ground truth by manually locating the camera and vending machines on the floor map.

3.4 Processing Time per Frame

To calculate the processing time for each frame, we measure the time taken to process one frame and then repeat the process for a large number of frames. We take the average as the processing time required per frame. The asset tracking system aims to perform in near real-time. If the processing time is too long the user will experience a lag in the detection results which can render the system unusable in practical application.

3.5 System Usability Scale

To evaluate the usability of the system we invited several volunteers to try on the asset tracking application and complete a survey. Again due to the pandemic restrictions, we had to do the survey online. Each volunteer was shown a demo video of the application of the system and the user interface and was asked to complete a survey based on the video. The questionnaire was designed based on an industry standard system usability scale (Brooke J. 1996). Since the volunteers did not have access to the system, we replaced the learning related questions with more specific system usability questions. The questionnaire designed to be focus on the user's ability to easily understand the various functions and the information displayed for the user. The responses were collected and used to calculate a score to represent the usability of the asset tracking system.

4. RESULTS

We implemented the vending machine tracking system as a mixed reality app for HoloLens. The app was then used to detect, locate, and record vending machines in a virtual setup. Each time that user sees an asset in front of him, the asset tracking system will automatically recognize the type and location of the asset without user's interaction. Then, the recognition result is displayed on the user's view by HoloLens. Using augmented reality, the predicted location and size can be visualized in 3D world. With our augmented reality interface, the user can freely move the predicted bounding box to modify the location if the user thinks it is not accurate. The interface will allow the user to accept or decline the detection result. If the user accepts the detection result, the location information will be stored in the database.

4.1 Power Consumption

The maximum running time we measure of our asset tracking system is 142 min. The acceptable running time range for fully charged HoloLens on active use specified by Microsoft is 2-3 hours (Mattzmsft 2019). We observe that our asset tracking system can run within this time range. It means that our asset tracking system only consumes a moderate amount of power while operating. Since we did not have access to the environment, the relationship between the maximum running time of the asset tracking system and the time required to perform asset tracking in one building remains unknown. Further investigation on this relationship is required.

4.2 Asset Detection Accuracy

The evaluation results of our asset detection module are shown in Table 1. The results show that our object detection network is able to provide a reliable result to be recorded in database. The mean intersection over union shows that the predicted bounding box from our asset detection module has an average 85.77% overlap with the ground truth bounding box. It indicates that the bounding box result can be used to make depth estimation in asset localization module since the bounding box represents the width of the vending machine. Misclassification occurs when the object detection module confuses other similar objects, such as a cabinet, with vending machines. In Figure 7, we show an example of misclassification result from our object detection network. The predicted result with a confidence level of 0.5166 is drawn in red in the figure. As we can see the cabinet in the figure has a similar appearance with vending machine. This similar appearance confuses the object detection network to consider the cabinet as a vending machine. This problem can be resolved by introducing more diverse examples from different sources to the training process.

Precision	88.9%
Recall	82.1%
mAP	94.8%
mIOU	85.77%

 Table 1. Asset detection accuracy evaluation results



Figure 7. A cabinet misclassified as vending machine

4.3 Asset Localization Accuracy

The average localization accuracy after removing outliers (distances larger than 5 meters) was found to be 2.311 meters. In Figure 8, we can see the distribution of localization errors with respect to the ground truth vending machine locations. The

majority of the predicted locations are shifted in the west direction. One of the reasons for having this result is that we have manual error in establishing the ground truth locations. The ground truth location is estimated manually based on captured photos of vending machines. In the manual process, errors in determining the camera location and orientation on the floor map are unavoidable.



Figure 8. Distribution of predicted location compare to ground truth location

4.4 Processing Time by Frame

The processing time we measured as average over 100 frames is 1611 ms per frame. It is around 0.62 frame per second. In a practical application, the user will experience a small lag during the asset detection, however this does not seriously affect the operation of the asset tracking system.

4.5 System Usability Scale

For the augmented reality user interface evaluation, 10 volunteers were asked to complete the questionnaire. The results showed that most of the participants were able to read and understand the information we presented in the user interface. The participants believed that the information was well organized. Most of the participants found the system easy to use and not too complex. One issue that was identified from the responses was that the system has some inconsistency and education is needed before one can start using the asset tracking system. The system usability scale score that our asset tracking system obtained was 75.6/100. This score is above the average system usability scale score of 68. It demonstrates that the augmented reality user interface of our asset tracking system can help the user understand the information easily and interact with the system naturally. It also shows that there is still room for improvement to provide a better user experience. To reduce the knowledge needed for operating the system, a step-by-step tutorial can be added when the user starts the application for the first time. More informative visualisations can also be introduced, e.g. a map of the current environment to help the user understand the current position of the system.

5. CONCLUSION

We have proposed a new design of asset tracking system based on HoloLens which uses machine learning to detect assets in near-real time. We demonstrated a vending machine tracking application using HoloLens and perform an evaluation on the application's capability. The result has shown high accuracy in asset detection and acceptable accuracy in asset localization. The system achieves 94.8% mean average precision in asset detection and 2.311 meters in asset localization accuracy. The system is demonstrated to be easy to use.

The major limitation of this research is the lack of the evaluation in a real environment. Due to the pandemic situation, access to the University campus was restricted. Therefore, all evaluations were done in a minimum setup. Another limitation of this research is that the computer vision-based asset tracking system cannot uniquely identify an asset. Therefore, the location information is only reliable if the assets are not likely to move. For moving assets, additional approaches such as barcode or serial number recognition needs to be used to uniquely identify an asset. In that case, we can use the location of the asset as the unique identifier. The proprietary SLAM software of HoloLens is another limitation. This restricts our ability to modify or improve the performance of the SLAM algorithm.

Overall, the result of evaluation shows the capability of the asset tracking system in increasing the efficiency of asset tracking and reducing the manual process required in asset tracking. Further research can be focused on evaluation in real environments, reducing the detection lag and improving the localization accuracy.

REFERENCES

Angeles, R., 2005. RFID technologies: supply-chain applications and implementation issues. Information systems management, 22(1), pp.51-65.

Belkacem I, Pecci I, Martin B., 2019. Pointing task on HoloLens: Comparison of four interaction techniques. ArXiv, abs/1905.05810

Brooke, J., 1996. SUS: a "quick and dirty'usability." Usability evaluation in industry, p.189.

Chen, D., Tsai, S., Kim, K.H., Hsu, C.H., Singh, J.P. and Girod, B., 2010, September. Low-cost asset tracking using location-aware camera phones. In Applications of Digital Image Processing XXXIII (Vol. 7798, p. 77980R). International Society for Optics and Photonics.

Glocker, B., Shotton, J., Criminisi, A., Izadi, S., 2014. Real-Time RGB-D Camera Relocalization via Randomized Ferns for Keyframe Encoding. IEEE Transactions on Visualization and Computer Graphics 21, 571-583.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., 2016. Deep learning (Vol. 1, p. 2). Cambridge: MIT press.

Huang, R., Pedoeem, J. and Chen, C., 2018, December. YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 2503-2510). IEEE.

Khoshelham, K., Tran, H. and Acharya, D., 2019. Indoor mapping eyewear: Geometric evaluation of spatial mapping capability of Hololens. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences

Kostoeva, R., Upadhyay, R., Sapar, Y. and Zakhor, A., 2019. INDOOR 3D INTERACTIVE ASSET DETECTION USING A SMARTPHONE. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences. Mattzmsft (2019). HoloLens (1st gen) hardware. [online] docs.microsoft.com. Available at: https://docs.microsoft.com/en-us/hololens/hololens1-hardware#device-specifications [Accessed 21 Oct. 2020].

Mur-Artal, R. and Tardós, J.D., 2017. Orb-slam2: An opensource slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics, 33(5), pp.1255-1262.

Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M., 2013. Real-time 3D Reconstruction at Scale using Voxel Hashing. ACM Transactions on Graphics 32, 1-11.

Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

Tola, E., 2005. Multiview 3D Reconstruction of a scene containing independently moving objects (Master's thesis).