An RFID and Computer Vision Fusion System for Book Inventory using Mobile Robot

Jiuwu Zhang*†, Xiulong Liu*†, Tao Gu‡, Bojun Zhang*†, Dongdong Liu*†, Zijuan Liu*†, Keqiu Li*†
*College of Intelligence and Computing, Tianjin University, China
†Tianjin Key Laboratory of Advanced Networking (TANK)
‡Macquarie University, Australia

IEEE INFOCOM 2022 - IEEE Conference on Computer Communications | 978-1-6654-5822-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INFOCOM48880.2022.9796711

***Abstract*—Mobile robot-assisted book inventory such as book identification and book order detection has become increasingly popular in smart library, replacing the manual book inventory which is time-consuming and error-prone. The existing systems are either computer vision (CV)-based or RFID-based, however several limitations are inevitable. CV-based systems may not be able to identify books effectively due to low accuracy of detecting texts on book spine. RFID tags attached to books can be used to identify a book uniquely. However, in high tag density scenarios such as library, tag coupling effects of adjacent tags may seriously affect the accuracy of tag reading. To overcome these limitations, this paper presents a novel RFID and CV fusion system for Book Inventory using mobile robot (RC-BI). RFID and CV are first used individually to obtain book order, then the information will be fused by the sequence based matching algorithm to remove ambiguity and improve overall accuracy. Specifically, we address three technical challenges. We design a deep neural network (DNN) model with multiple inputs and mixed data to filter out interference of RFID tags on other tiers, and propose a video information extracting schema to extract book spine information accurately, and use strong link to align and match RFID- and CV-based timestamp *vs.* book-name sequences to avoid errors during fusion. Extensive experiments indicate that our system achieves an average accuracy of 98.4% for tier filtering and an average accuracy of 98.9% for book order, significantly outperforming the state-of-the-arts.**

***Index Terms*—Computer Vision, RFID, Book Inventory, Multi-modal Fusion.**

I. INTRODUCTION

A. Motivation and Problem Statement

Mobile robot-assisted book inventory has become increasingly popular in smart library to replace the manual book inventory which is both time-consuming and error prone especially when dealing with millions of books [1], [2]. This paper studies mobile robot-assisted book inventory including book identification and book order detection, we formally describe the problem as follows. Given a typically library setting where books are placed on bookshelves with multiple tiers, a mobile robot equipped with an RFID reader and a camera moves between bookshelves and scans books tier by tier. Two specific tasks need to accomplish for book inventory. *The first task is to identify all the books on a particular tier of bookshelf given that tags are read from all tiers within the range of antenna, e.g., determine if a tag belongs to the tier currently in scanning.* Given the total number of tags n read by antenna and the number of tags m that have been correctly determined if they belong to the tier in scanning or not, we define the accuracy of tier filtering as $\psi{=}\frac{m}{n} \in [0, 1]$. *The second task is to detect book order for each tier of bookshelf.* To thoroughly evaluate the book ordering accuracy, we adopt two metrics as follows. The first metric is Normalized Kendall Tau Distance (NKTD), defined as $\kappa{=}K(\eta, \eta')/\binom{n}{2}$, where η is the ground-truth sequence, η' is the sequence measured, and K is the Kendall tau distance [3] that indicates the total number of disordered pairs. Divided by $\binom{n}{2}$, the Kendall Tau distance is normalized into a range of $[0, 1]$. The second metric is Normalized Value Deviation (NVD), defined as $\varepsilon{=}D(\eta, \eta')/\binom{n}{2}$, where D is obtained by summing the absolute value of the difference between elements in the corresponding positions of the two sequences. We also normalize it by using $\binom{n}{2}$ to divide it, thus the accuracy of book order ε is also within $[0, 1]$.

Correspondence author: Xiulong Liu.

B. Limitations of Prior Art

The state-of-the-art systems for book inventory can be classified into two categories: RFID-based and CV-based. RFID-based systems such as STPP [2] and RF-Scanner [4] use the order of the perpendicular point (*i.e.*, the point where antenna passes by tag) for tag order. These systems perform well in most cases given an advantage of quick identification of RFID. However, dense deployment of RFID tags in a library setting will cause serious tag-coupling effect and tag collisions, dramatically degrading the performance of inventory (i.e., missing tags or wrong order detected). Different from RFID, CV-based systems [5]–[7] utilize rich texts and graphic patterns on book spine to differentiate and locate books, thereby detecting book order. However, detecting texts and patterns with CV is not perfect, particularly in various lighting environments and different book orientations. Even many books have no text on their spine. In summary, neither of these systems can achieve a satisfactory accuracy for book inventory.

C. Our Approach

In this paper, we propose an RFID and CV fusion system for Book Inventory using mobile robot named RC-BI, leveraging on the advantages of both RFID and CV for improving accuracy and robustness of book order detection. As illustrated in Fig. 1, RC-BI uses a mobile robot equipped with a camera and an RFID reader connected with an antenna to scan bookshelves. During scanning, the RFID reader continuously

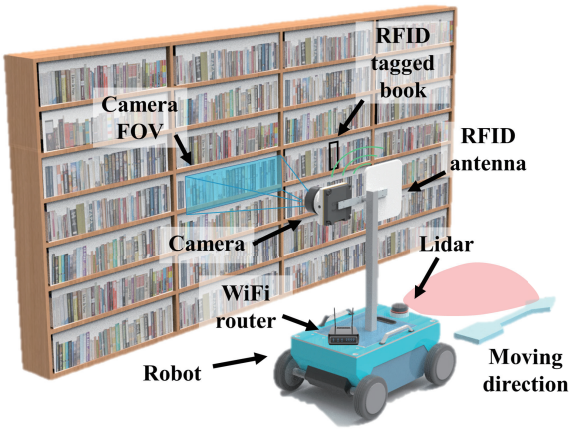

Fig. 1. RC-BI hardware.

collects data from tags attached to books, and the camera records the video of the facing tier of bookshelf. We design our system in three modules. In the RFID module, we design a multiple inputs and mixed data deep neural network (DNN) to filter out tags from other tiers. Then, we sort tags by the timestamp of the maximum RSSI, and thus generate an RFID-based timestamp *vs.* book-name sequence. In the CV module, RC-BI first stitches frame images of the video into a panorama. To extract book texts accurately, we use Optical Character Recognition (OCR) on frame image and sort the results on panorama with timestamps. To deal with overlapped and fragmented OCR results, we propose a cluster algorithm to cluster the fragmented texts into book-names, thus obtain the CV-based timestamp *vs.* book-name sequence. In the RFID-CV fusion module, we propose the concept of strong link to align and find one-to-one mapping between sequences from two modals. Finally, according to the mapping, RC-BI updates the RFID-based sequence, and obtains the final book order.

D. Challenges and Solutions

To extract and fuse the information from RFID and CV, we identify three technical challenges.

The first challenge is how to filter out tags on other tiers of bookshelf. Since RFID antenna typically has a wide coverage, tags located on other tiers (*i.e.*, not the current tier in scanning) will be read, causing serious interference to order detection. In a normal case, the changing rate of phase profile indicates the tag-antenna distance, *i.e.*, the distance between tag and antenna trajectory. However, in a library, books are densely placed on bookshelves with a typical distance of $2cm$ away from each other, which brings heavy couple effect and channel competition. Therefore, the sample count of most tags can be extremely low. Using the sparse data, the unwrapping method often fails and fitting results on these data may result in large errors. In this paper, we design a multiple inputs and mixed data DNN model. We combine features from unwrapped phase and RSSI profile images and numerical sequences. The quadratic coefficient of phase fitting function and robot-shelf distance, *i.e.*, the distance between trajectory of robot and shelf, are also taken into account. This design can effectively find features of each input and their relationship, potentially achieving high accuracy.

The second challenge is how to extract highly credible and orderly information from the video. Although the video contains much information of book spine, how to extract and sort them is a critical problem. A common idea is to conduct OCR on each frame image, but the recognized text blocks are fragmented. Furthermore, since the speed of robot is unstable in most cases, it is difficult to determine the sample interval to extract the image, which causes the recognition results contain many unordered and repeated information. Another method is to conduct OCR on panorama, but the panorama usually contains many pixel artifacts, which seriously reduces recognition accuracy. In our paper, we combine the advantages of these two ideas. We conduct OCR on the image of each frame, and map the results into the panorama, hence avoiding the problem of pixel artifacts while keeping high recognition accuracy. To deal with the problem of overlap and fragmentization, we propose an de-overlap and clustering algorithm.

The third challenge is how to align the two sequences from RFID and CV. To avoid introducing extra errors to RFID sorting results, before fusing CV sequence into RFID sequence, we have to align these two sequence. Since RFID sequence usually has a large time offset, and its items are out of order, we cannot align two sequences directly using the average timestamp difference. Additionally, if we use the timestamp offset between the matched items between two sequences, how to avoid the effect of mismatching is also a tricky issue. In our paper, we propose the concept of strong link. Strong link can effectively avoid mismatching problems, and multiple rounds of strong link based adjusting ensure highly accurate alignment.

E. Contributions and Advantages over Prior Work

The contributions of this paper are summarized as follows.

- We propose a novel tier filtering method based on the specially designed multiple inputs and mixed data DNN model. The model can effectively filter out the data from other tiers in a densely deployed scenario with an average accuracy of 98.4%.
- We propose a novel video information extracting schema, which combines the OCR on frame image, the mapping with panorama, and the upper box based cluster algorithm. The schema can extract information from video orderly and accurately.
- We fully implement RC-BI, and conduct extensive experiments in a library setting to evaluate its performance. With our fusing method based on strong link, RC-BI achieves an average order accuracy of 98.9% (measured with normalized Kendall tau Distance), demonstrating the effectiveness of fusing CV and RFID.

We organize the remainder of this paper as follows. Section II introduce the brief architecture of RC-BI, and Section III presents the details of our approach. Section IV presents our implementation, and Section V shows the exper-

imental results. The related work is discussed in Section VI, and Section VII finally concludes the paper.

II. System Architecture

The RC-BI system is designed for the inventory of RFID tagged objects in the libraries, warehouses, *etc*. As illustrated in the Fig. 1, the hardware architecture of RC-BI contains an RFID reader connected with an antenna, an RGB camera (we employ a smart phone to play this role in experiments), and a mobile robot. When the robot drives the whole system moving in a straight path, the RFID reader keeps reading tags and the camera records the video of the shelf simultaneously. The processing mechanism of RC-BI can be briefly divided into three modules: RFID module, CV module and RFID-CV fusion module, as illustrated in the Fig. 2.

RFID module: Since all the tags in the read range of RFID antenna will be read, the recorded RFID data contain irrelevant data. We first use a multiple inputs and mixed data DNN to filter out the tags from other tiers and environmental noisy tags. Then, RC-BI gets the timestamps when the robot just passes by the tags using the vertex of RSSI profile. By querying the database using RFID Electronic Product Code (EPC), the RFID module eventually acquire the RFID-based timestamp *vs.* book-name sequences.

CV module: In the CV module, we extract RGB images from video frames. For each image, we conduct OCR to acquire the texts and the according coordinate of text boxes in that image (local coordinate). Then, we stitch the images into a panorama and mapping all text boxes into the corresponding position (global coordinate) of this panorama. Since the text boxes are fragments with overlap, we conduct de-overlap and clustering algorithms to get tidy book-name sequence. Finally, we acquire the CV-based timestamp *vs.* book-name sequences by calculating the timestamp of text boxes according to the timestamp of its original frame image.

RFID-CV fusion module: In this module, we propose strong link to mine the most feasible correspondence of book-names between RFID-based and CV-based sequences. To align two sequences, RC-BI uses the average time difference between both ends of the strong link to adjust the timestamps. For each book-name in the RFID-based sequence, RC-BI matches it with the most similar and closest book-name in the CV-based sequence. Since the timestamps have been aligned, RC-BI replaces the each book-name in the RFID-based sequence with the timestamp of matched book-name in the CV-based sequence. Finally, the system re-sorts the RFID-based sequence according to the updated timestamps.

III. Details of the RC-BI System

In this section, we first introduce some preliminary knowledge of RFID. Then, we introduce details about the operation of RC-BI following the data flow in the Fig. 2.

A. Preliminary knowledge of RFID

The RFID system consists of reader, antenna and RFID tags attached on objects. The antenna is connected with the reader, and the reader is controlled by the PC. The RFID data we utilize in this paper is EPC, phase and Received Signal Strength Indicator (RSSI). EPC is the unique identifier of each RFID tag. The phase can be calculated as follows [8].

$$\theta = \left[\frac{2 \times d}{\lambda} \times 2\pi + \mu \right] \mod 2\pi, \tag{1}$$

where d is the distance between antenna and tag, λ is the wavelength, and μ is hardware offset. The relationship between RSSI and distance can be expressed as follows [8], [9].

$$\frac{P_R}{P_T} \propto G_T G_R \left(\frac{\lambda}{4\pi d} \right)^2, \tag{2}$$

where the power attenuation is described as the power ratio between receiver and transmitter $\frac{P_R}{P_T}$. G_T and G_R are respectively the antenna gains from the transmitter and receiver.

B. RFID Processing

As illustrated in the Fig. 1, RC-BI makes mobile robot move straightly along the bookshelf carrying the RFID reader and camera. During the scanning, RFID reader keeps collecting the data from the RFID tags, which are attached on the back side of the books and near the spine of books.

1) Filtering out tags on other tiers: We use the circular polarization antenna to scan the shelf. Considering most of tags are blocked by books, we set the transmitting power to maximum, *i.e.*, $32.5dBm$, to ensure the activation energy of tags. However, many unrelated tags (*e.g.*, tags on the other tiers and tags in the environment) will be read, which brings great interference to the ordering. To distinguish the tags on the current tier, the previous works mainly focus on shape of phase profile [2] [4]. According to the Eq. (1), we can infer that the unwrapped phase is linearly related to the distance between the antenna and the tag, *i.e.*, tag-antenna distance, as illustrated in Fig. 4(c). Assuming the robot moves along the shelf in a straight path, theoretically the larger tag-antenna distance is, the smaller the correspondence radial velocity is. According to the curvature at the vertex of the phase profile or the phase changing rate, the tiers of tags can be distinguished. However, the average distance between tags on the shelf is around $1cm \sim 2cm$. Such a dense distribution will cause strong coupling effect, and due to the occlusion of tags, the noise in phase profile is intense. Plenty of experiments show that even in the same tier, the curvature of the profile is different. What's worse is that, the phase profile from the further tier sometimes has nearly same curvature as the phase profile from the facing tier.

To solve this problem, we develop a multiple inputs and mixed data DNN to filter out tags on other tiers. The net has 7 inputs: the unwrapped phase profile image, RSSI image, unwrapped phase sequence, RSSI sequence, sample count, quadratic coefficient of fitted phase curve, and the distance between the robot and the shelf. The unwrapped phase profile image and unwrapped phase sequence are two representations of the same data, as are the RSSI image and RSSI sequence. But they are not redundant, since we choose these inputs for two reasons: mining as many distance-related features as

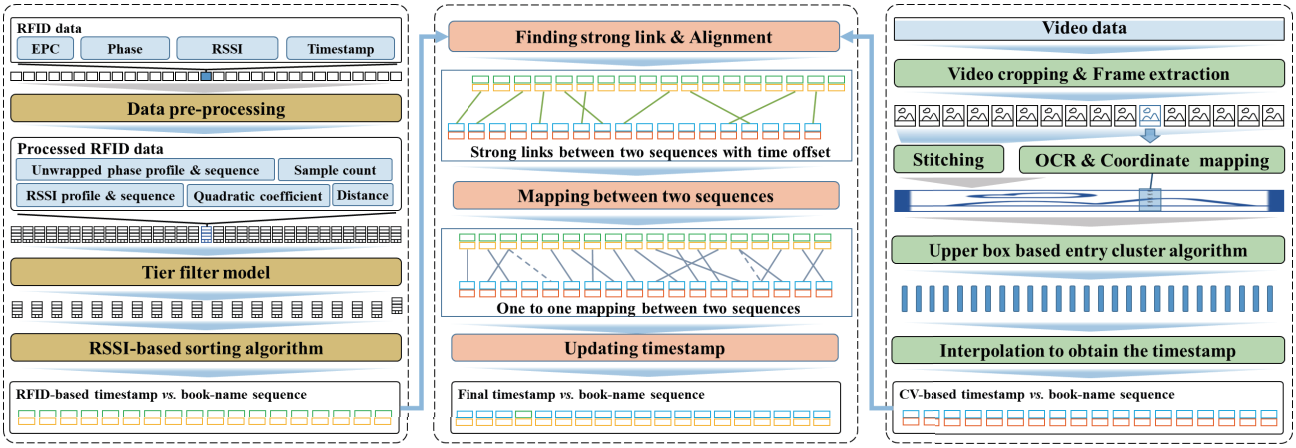

Fig. 2. System overview.

possible and avoiding the instability of a single parameter. As mentioned above, the phase profile contains information of the tag-antenna distance, but some tags on other tiers can get the similar phase profile. According to Eq. (2), the changing rate of RSSI profile can also reflect the tag-antenna distance. However, compared to the phase, RSSI is an unstable parameter, which is more vulnerable to the multipath effect. Besides, it is of higher probability for tags with smaller tag-antenna distance to get a large sample count, since RFID tag needs to gain power from the antenna. Theoretically, the time *vs.* unwrapped phase curve is a branch of the hyperbola. Considering the complexity of hyperbolic fitting, we simplify the calculation and fit the curve approximately with the quadratic function. The quadratic coefficient is positively correlated with phase changing rate, thus, negatively correlated with tag-antenna distance. Therefore, the quadratic coefficient is an important parameter to accelerate the convergence of the model in the training. Furthermore, with different distances between the robot and shelf, the distribution of the features acquired from the other six parameters will be different, so the distance between the robot and shelf is also a necessary parameter. In summary, except for distance between the robot and the shelf, these inputs all reflect the tag-antenna distance to a certain extent. Since each single parameter has many outliers, it is rational and necessary to combine these variables together for calculation. However, it is difficult to determine the functional relationships, weights between them if these parameters are combined. Therefore, we use a neural network-based method. Theoretically, the DNN can not only learn the relationships, weights, and thresholds automatically, but also discover other features beyond the tag-antenna distance.

As illustrated in Fig. 3, we adopt a DNN with multiple inputs and mixed data. Among the inputs, the phase and RSSI profiles are two image data. Besides the RSSI curve itself, we also draw a baseline at $0dbm$ and limit the Y-axis from $-80dBm$ to $10dBm$ in the RSSI profile image, in order to utilize the offset between the baseline and the curve to make the net more sensitive to learn the value of RSSI. The image data are first inputted into a ResNet50 [10], which is a commonly used network to extract features from the image. After ResNet50, we link a multi-head attention block, which helps the net captures richer features. Through multi-head attention block and one fully connected layer, we get a feature vector with length 16. With the same operation on RSSI profile image, we also get a feature vector of length 16. The unwrapped phase and RSSI sequence are exactly two 2×256 arrays. The first row of the array is the robot moving distance (acquired using timestamp$\times$speed) sequence. The second row of the array is the unwrapped phase sequence or RSSI sequence. To match the input size, we use linear interpolation to make both sequences with length 256. To extract the features as fully as possible and mine more relations between moving distance and phase/RSSI, we use a network structure based on FPN [11], [12], which is adept at capturing multi-level features. We also link a multi-head attention block and fully connected layer after FPN-based network and get another two feature vectors with length 16. Then, the sample count, quadratic coefficient and robot-shelf distance are concatenated with the above four feature vectors to form a feature vector with length 67. This feature vector is sent to multiple fully connected layers to find relations between different features of inputs. Finally, it outputs a vector with length 2, and the index of the maximum is the class number.

2) Acquiring RFID-based timestamp vs. *book-name sequence:* When the robot moves along the shelf, the real-time distance between antenna and the tag decreases first, and increases after the robot arrive at the perpendicular point, *i.e.*, the point at which the robot just passes by the tag. The raw phase profile looks like Fig. 4(a), which has many segments due to the periodicity. To acquire the timestamp sequence, STPP [2] utilizes a Dynamic Time Warping (DTW) based matching algorithm to find the V-zone in the center of the profile, as shown in Fig. 4(b). Then it conducts quadratic fitting on the V-zone, and takes the symmetry axis of the conic as the timestamp of the tag. Different from STPP, RF-scanner [4] uses unwrapping method to splice these segments into an unwrapped phase profile, as Fig. 4(c) shows, and use the whole unwrapped phase profile for quadratic fitting. For most cases

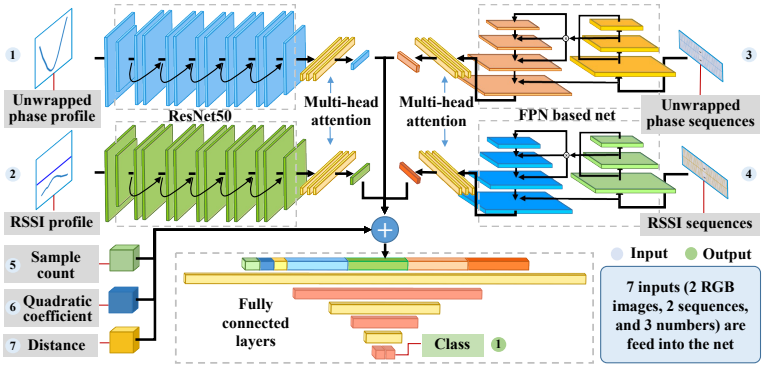

Fig. 3. The multiple inputs and mixed data net for filtering out books on other tiers.

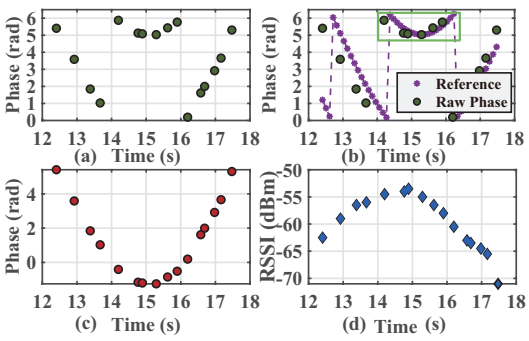

Fig. 4. Examples of RFID data.

where the phase profile is legible with enough sample points, these phase based methods perform well. However, when the sample count of the profile is sparse, even in some cases sample count is no more than five, which brings great error to the DTW matching and unwrapping. In summary, these phase based methods is lack of robustness for sparse data.

On the contrary, although the RSSI is more sensitive to multipath effect than the phase, it does not contain periodicity. Thus, the trend of RSSI changing with distance is obvious, as shown in Fig. 4(d). Even when the data is very sparse, RSSI is still meaningful to measure the distance. But in that case, the phase data becomes less reliable. Because sparse phase is difficult to unwrap, separated phase points with periodicity are ambiguous. Therefore, RSSI is a more robust quantity than the phase for solving our problem. Furthermore, since we have the CV module to correct errors in the end, we don't request the RFID-based timestamp sequence to be very precise. Finally, we adopt the timestamp of the peak in the RSSI profile as the timestamp for the corresponding book. After querying the book-name for each tag in the database, we get the RFID-based timestamp *vs.* book-name sequence, *i.e.*, α=<<t_1, b_1>, $\cdots$, <t_i, b_i>, $\cdots$, <t_n, b_n>>, where b represents book-name.

C. CV processing

The camera is installed horizontally close to the antenna, and its field of view exactly covers the upper and lower edges of one tier in the bookshelf. During the scanning, the camera records the video of the facing tier, which contains the image information of the book spine.

1) Extracting and stitching frame images: Compared with the video itself, the panorama can reflect the positional relationship of the books more clearly and orderly. To get the panorama, we first extract frames from the video, then stitch them into one image. RC-BI extracts frames with a pre-defined interval, *i.e.*, $\tau = \frac{l \cdot f}{v}$, where l is the excepted moving distance interval between two adjacent extracted frame, f is the frame rate of the video and v is the moving speed of the robot. In our system l is set as $0.05m$, the frame rate of the camera is set as $30fps$. To reduce the consumption of frame extracting and image stitching, we develop a classifier, which is simple DNN consisting of a ResNet34 and multiple fully connected layers, to detect whether there are books in the frame image. If three consecutive frames are judged to be with books, we infer that the bookshelf has entered the camera's field of view, and the system start to store the extracted images. When three consecutive frames are judged to be without books, we infer that the bookshelf has left the camera's field of view, then RC-BI stops extracting images. After extracting, we acquire a set of frame images, then we use an open-source program called OpenPano [13] to stitch the frame images. To improve the stitching efficiency, we divide the images extracted from one tier into six groups, and use multiple threads to process these six groups in parallel. The stitched images from these groups are eventually stitched again to get the panorama of the tier.

2) Conducting OCR and coordinate mapping: The OCR model we use is PaddleOCR [14], which is an open-source framework with excellent recognition performance. However, when we carry out OCR on the panorama, the recognition result is not satisfactory. The reason is that the panorama contains a large range of pixel artifacts caused by lens distortion and jitter. To solve this problem, we conduct OCR on frame images instead of the panorama. The frame image is sharper than the panorama, and the same book shows many times in frame images due to the necessary overlap for stitching. Therefore, the spine of the same book will be recognized many times, thus we can select the recognition result with the highest confidence as final result. OCR program gives set of recognition entries formatted as $<fid, box, text, score>$, where fid represents which frame image the recognition entry comes from, box represents four vertex coordinates of rectangular area where the text is located, $text$ and $score$ represents respectively the recognized text and its confidence.

The result produced by OCR is a set of disordered recognition entries. The box of each entries is according to the coordinate system of corresponding frame image,*i.e.*, local coordinate, and we need to map them into the coordinate system of panorama, *i.e.*, global coordinate. In RC-BI, we utilize the theory of image stitching to realize the coordinate transform. RC-BI takes Scale-invariant Feature Transform (SIFT) algorithm to extract keys points from the phase image and the panorama, then it uses knnMatch algorithm to find the mapping between them, finally uses Random Sample Consensus (RANSAC) algorithm to calculate the homography matrix. The homography matrix represents mapping between local

Algorithm 1: Upper Box Based Entry Cluster (UBEC)

Input: Entry sequence E, image shape (w, h), Distance threshold D
Output: Cluster-Entry mapping M

$M \leftarrow \{\}$; $I \leftarrow 0$; $Sort(E)$;
for i *in range(*0, *E.length)* **do**
  $pre\_entry \leftarrow E[i]$;
  if $pre\_entry.cluster \neq null$ **then** $continue$;
  else
    $pre\_entry.cluster \leftarrow I$; $pre\_entry.upperbox \leftarrow GetUpperBox(pre\_entry.box, w, h)$;
    $M[I].add(pre\_entry)$; $I \leftarrow I + 1$;
  for j *in* $range(i + 1, E.length)$ **do**
    $post\_entry \leftarrow E[j]$;
    if $Distance(post\_entry, pre\_entry) > D$ **then**
      $continue$
    if *Cover(pre_entry.box, pre_entry.upperbox)* **then**
      $post\_entry.cluster \leftarrow pre\_entry.cluster$;
      $M[pre\_entry.cluster].add(post\_entry)$
return M;

coordinate and global coordinate, using which we transform local coordinates in each entry to global coordinates. After transform, the entry can be simplified as $<box, text, score>$, where box is according to global coordinate system, and the set of entries can be represented as E.

3) Acquiring CV-based timestamp vs. *book-name sequence:* The recognized entries are repeated and fragmented, *e.g.*, the same texts are stored in multiple overlap entries; or the text of one book's title is stored in one entry, but the corresponding text of its authors is stored in another entry. To remove the overlap between entries, we group the entries whose distance of the box center is less than $50pixel$, *i.e.*, an empirical threshold, and only reserve the one with the highest confidence. To concatenate the fragments as the whole book-name, we develop an Upper Box Based Entry Cluster (UBEC) algorithm to cluster these entries. In our algorithm, we sort the entries by the area of the box from large to small, then traverse the sorted sequence. For each entry, if this entry doesn't belong to any cluster, we give it a new cluster number, and use it as the representative of this cluster, then calculate its upper-box. The upper-box is acquired by extending the height of the box until it hits the border of the panorama, thus forming a large box expected to cover the whole book spine. The detailed processing can be seen in the pseudo-code in Algorithm 1. Then, we stitch these texts in same cluster together as the book-name, and take upper-box of the representative as the location of the book-name, formatted as $<class\_no, upper\_box, book\_name>$.

While the book-name has been obtained, how to acquire the timestamp is still a problem. Fortunately, the fixed frame rate of video makes the frame number a nature timer. Using the frame rate and frame index of the video, each coordinate in the center of corresponding image can be bounded with a timestamp. Next, utilizing the homography matrix in the previous subsection, the center points of the frame images can be mapped to the panorama. Thus, as we only care about the horizontal coordinate, we get a timestamp *vs.* coordinate sequence, *i.e.*, $\beta = <<t_1, x_1>, \cdots, <t_i, x_i>, \cdots, <t_n, x_n>>$, where t represents timestamp, x is the horizontal coordinate. We take the horizontal coordinate of the upper-box's center as the location of each cluster. Based on the sequence β, taking the locations of clusters as interpolation points, we can get the CV-based timestamp *vs.* book-name sequence, *i.e.*, $\gamma = <<t_1, b_1>, \cdots, <t_i, b_i>, \cdots, <t_n, b_n>>$, where b represents book-name.

D. RFID-CV fusion

The RFID-CV fusion module utilizes the relative order of CV-based timestamp *vs.* book-name sequence to update the timestamp of RFID-based timestamp *vs.* book-name sequence.

1) Timestamp alignment: It is a common phenomenon that timestamps of different modal data are not synchronized. For our system, the timestamp offset is mainly caused by two reasons. First, the clock times of RFID reader and camera are not fully calibrated, and their startup and communication delay are different. Second, since the camera and RFID antenna are deployed horizontally with around $15cm$ distance along the robot moving direction, there are naturally offset in timestamp. To solve this problem, we propose the concept called strong link, which represents the most feasible matching pairs of the elements from RFID-based and CV-based sequences. We utilize the FuzzyWuzzy [15], which is a fuzzy string matching algorithm based on the Levenshtein Distance, to match the book-names from two sequences and give a matching score in the interval of $[0, 100]$. The strong link should meet three conditions: (a) The timestamp offset between both ends of the link should within a threshold τ_1, which is an empirical value set as half of the range of the RFID timestamp sequence. (b) The matching score between the book-names of both ends should exceed a threshold σ_1, which is an empirical value set as 50. (c) The matching score of the strong link should be the highest in the searching range τ_1, and exceed the second highest with a threshold $\Delta\sigma$, which is an empirical value set as 5. Before the alignment, RFID-based sequence α and CV-based sequence γ are sorted by timestamp from small to large. Next, RC-BI traverses items in α to find the strong links with items in γ, and calculate the average timestamp offset between ends of strong link. If the average timestamp offset is smaller than a threshold $0.3s$, we consider two sequences have been aligned. Otherwise, RC-BI repeats the procedure above again until the average timestamp offset meets the threshold.

2) Updating RFID timestamp sequence: RC-BI matches each item in RFID-based sequence α with a corresponding item in CV-based sequence γ. The matching method is just like the method to find strong link, but with the threshold modified. We use extreme thresholds to find strong link, because we only need few strong link with highest confidence, which is unsuitable for the matching for most items. In the matching

procedure, we determine the threshold considering the mess degree of both α and γ. However, how to define the mess degree without the knowledge of ground-truth. We define the length of α and γ respectively as u and v, the range of timestamps in α and γ respectively as S and T. The mess degree of RFID-based sequence can be defined as $a=\frac{S}{T}$, because the timestamps of γ is concentrated with the limit of the panorama. The mess degree of CV-based sequence can be defined as $b=\frac{v}{u}$, because assuming the DNN model for filtering out tags on other tiers is pretty accurate (we reveal this point in section V), the length of α should be very close to the number of book-names recognized, except for some books without text. Then we set $\tau_1=\frac{abS}{2}$, σ as 20 and $\Delta\sigma$ as 10. With the searching range τ_1 enlarged and lower bound of matching score decreased, increasing $\Delta\sigma$ is necessary to avoid the extra error caused by mismatching books with similar name. After matching, we get the mapping form α to γ, then we utilize the matching score as metric to transform one-to-many mapping into one-to-one mapping. According to the mapping, RC-BI updates timestamps in α, and get the final ordering sequence.

IV. Implementation

In this section, we introduce the hardware and software deployment of RC-BI system.

A. Hardware

The hardware includes fours parts: mobile robot, RFID reader, RGB Camera, and RFID tagged books. The mobile robot is based on a self-designed automated guided vehicle, which mainly consists of a motor, lidar, lifting arm, communication module, and a central control unit. The robot communicates with the server by the WiFi module. Since many books have a certain tilt and the antenna is always moving, we select a circular polarization antenna, JT-628 RFID antenna, to achieve a more stable reading rate. The antenna is installed on the lifting arm connected with an Impinj Speedway R420 RFID reader. Horizontally aligned with the RFID antenna, we also install a Oneplus 7 Phone servicing as an RGB camera to record the video. The lifting arm can adjust the camera and antenna to different heights to read different tiers of the bookshelf. Each book is tagged with the Alien AZ-9640 RFID tag on its back cover and near the middle of its spine.

B. Software

The robot is driven by the Robot Operating System (ROS) system, and the data collection program is developed based on the Impinj Octane Java SDK [16]. The image stitching and OCR program are developed based on OpenPano [13] and PaddleOCR [14]. The tier filtering DNN program is developed on the TensorFlow framework, and the classifier to detect the start and end frame is developed on PyTorch. Both of the two models are trained on Dell Precision 7920 server. We develop the RFID CV fusion algorithm in Python, running in a Lenovo X1 Carbon 2018 laptop with 8G RAM and Intel Core i5 8250U CPU. During the operation, the RFID system is set with a fixed frequency of $920.625MHz$ and $32.5dBm$ transmitting power, and the frame rate of camera is set as $30fps$.

V. Evaluation

In this section, we conduct extensive experiments to evaluate the effectiveness of the models, the accuracy of the RFID-CV fusion algorithm and compare the performance with the state-of-the-art. The metrics to measure the accuracy have been defined in the part of problem statement in section I.

A. Effectiveness of the models

We use two indispensable models in the system to deal with some detailed but critical issues.

1) Accuracy of the model for tier filtering: The first model is the multiple inputs and mixed data model used for filtering out the RFID signal from other tiers. We collect sets of data with different conditions, including different speeds, tiers, distances, and slopes of books (the details about condition setting will be explained in the evaluation of order accuracy). As explained in Section III-B1, for each tag data, we generate 7 inputs: two image data of unwrapped phase profile and RSSI profile, two sequences of moving distance *vs.* unwrapped phase and moving distance *vs.* RSSI, sample count, quadratic coefficient and the robot-shelf distance. The RFID reader collects all data from the tags in the reading range and an average of 263 tags can be read during one scanning, of which around 70% is the RFID signal from other tiers. We build the training data set using the 24612 data collected from 60 times scanning with different conditions, insisting of 17035 positive samples and 7577 negative samples. The model is trained with a batch size of 64 and 80 epochs, and the loss function is cross-entropy. We collect the test data set with 77750 items from 296 times scanning with different conditions, including 24569 positive samples and 54727 negative samples. As shown in Table I, we test the model under different conditions. The minimum accuracy is 94.2%, and most of the accuracy exceeds 99.1%, and the average accuracy is 98.4%. which is enough for RC-BI to filter out the tags from other tiers.

2) Accuracy of the model for video cropping: The second model is the book classifier used to determine the start and end time of the video. To train this model, we collect 6452 images, of which 4884 images are the positive samples with book spines, 1568 images are negative samples without book spine. The model is trained with batch size of 8 and 20 epochs. The trained model is tested in a data set with 1870 positive samples and 1226 negative samples. The accuracy, precision and recall of the test result are respectively 99.42%, 99.95% and 99.09%. In actual use, we judge the scanning starts when three consecutive frame images are all tested positive, and ends when three consecutive frame images are all tested negative. Using this mechanism, the start and end scanning points of more than 300 of the videos we collect are all correctly judged.

B. Performance under different conditions

In this section, we first evaluate the performance of RC-BI with different speeds, tiers, distances, and slopes, then we compare RC-BI with the state-of-the-art. Since the high accuracy of tier filtering model, in order to compare order accuracy more conveniently, we set a reasonable assumption

TABLE I
PERFORMANCE OF THE MODEL FOR TIER FILTERING

Speed (m/s)	Tier	Distance (m)	Slope (degree)	Number	True Negative	False Positive	False Negative	True Positive	Accuracy	Precision	Recall
0.15	2	0.3	0	8376	5760	24	6	2586	0.9964	0.9908	0.9977
0.2	2	0.3	0	8507	5797	37	38	2635	0.9912	0.9862	0.9858
0.1	2	0.3	0	8371	5825	35	37	2474	0.9914	0.9861	0.9853
0.1	2	0.3	30	8989	6293	10	14	2672	0.9973	0.9963	0.9948
0.1	2	0.4	0	8696	5995	32	166	2503	0.9772	0.9874	0.9378
0.1	3	0.3	0	8811	5592	119	393	2707	0.9419	0.9579	0.8732
0.1	2	0.2	0	8877	6104	19	9	2745	0.9968	0.9931	0.9967
0.1	1	0.3	0	8178	5683	115	178	2202	0.9642	0.9504	0.9252
0.1	2	0.3	15	8946	6201	25	22	2698	0.9947	0.9908	0.9919
Total				77751	53250	416	863	23222	0.9836	0.9824	0.9642

that all the books on the facing tier are correctly classified. In all experiments of our evaluation, we deploy three tiers in a wooden shelf, with the number of the book being 78, 82 and 93, respectively. The default value of speed, tier, robot-shelf distance, and slope of book are $0.1m/s$, 2, $0.3m$, and $0degree$, respectively.

1) Performance with different speeds: We set the speed of the robot at three values: $0.1m/s$, $0.15m/s$, and $0.2m/s$. The other parameters keep the default. At each speed, we collect 30 groups of data. As shown in Fig. 5(a), the upper figure shows the distribution of order accuracy of all scanning. The result with the speed of $0.1m/s$ is the most stable, all accuracy exceeds 98% and 97%, respectively in NKTD and NVD. As shown in the lower figure in Fig. 5(a), with the speed growing, the average order accuracy both in NKTD and NVD drops slightly, but still keeps the average accuracy more than 97%. The effect of speed on RC-BI is two-folds. First, since the reading rate of the RFID reader is limited, as the speed increases, the time of each tag within the range of antenna drops, so as the sample count. Second, the faster speed makes the video more blurry, which reduces the performance of OCR.

2) Performance with different tiers: To scan each tier, we adjust the lifting arm of the robot to make the antenna and camera face the middle of that tier. As the accuracy of the tier filtering model is evaluated in the Section V-A1, we only evaluate the order accuracy in this section. As we can see from the upper figure in Fig. 5(b), the order accuracy of the different tiers in NKTD all exceeds 97%, and the accuracy in NVD is all above 95%. Through the lower figure in Fig. 5(b), we find the average order accuracy with different tiers fluctuates slightly, but the overall average accuracy remains more than 98% and 96% in NKTD and NVD, respectively.

3) Performance with different distances: We set three robot-shelf distances, *i.e.*, $0.2m$, $0.3m$, and $0.4m$, with other parameters keep the default. We collect 30 groups of data with each distance. Considering that the distance between bookshelves is limited, these distances are representative for robots on operation. As we can see from the upper figure of Fig. 5(c), the order accuracy with all distances exceeds 96%. As shown in the lower figure in Fig. 5(c), the average accuracy decreases slightly with the robot-shelf distance increasing. That's because with the distance increasing, the power each tag received decreases, and the resolution of the text drops.

4) Performance with different slopes: In this set of experiments, we deploy 40% of the books in the tier with the slope of 0, 15, and 30 degrees. The rate of tilted book and the angles we selected are reasonable in the actual library. For each setting, we still collect 30 groups of data. From results shown in the upper figure in Fig. 5(d), the distribution of accuracy with the slope of 15 degrees is relatively scattered, but all exceed 92%. From the lower figure in Fig. 5(d), we find no obvious correlation between the average accuracy and the slope.

C. Comparison with the state-of-the-arts

We compare RC-BI with STPP and RF-Scanner on 296 groups of data covering three tiers with the speed varying from $0.1m/s$ to $0.2m/s$, robot-shelf distance varying from $0.2m$ to $0.4m$, the slope varying from 0 to 30 degree. As shown in Fig. 6(a), both in NKTD and NVD, the distribution of accuracy in RC-BI is more concentrated. And nearly 100% of the order accuracy of RC-BI is more than 90%. Fig. 6(b) shows that, the average order accuracy of RC-BI is 98.89% in NKTD, and 98.08% in NVD. While the average order accuracy of STPP in NKTD and NVD are 86.30% and 90.08%, and the average order accuracy of RF-Scanner in NKTD and NVD are 90.06% and 90.15%. That proves that RC-BI achieves excellent performance over STPP and RF-Scanner.

VI. RELATED WORK

We summarize and classify the related work into three categories as follows.

A. RFID-based approaches

RFID has been used in many aspects such as human-computer interaction [17], [18], localization, troubleshooting [19], etc. For localization, many RFID-based work focuses on absolute localization. Wang *et al.* [20] devised PinIt, which estimates the location of the target according to reference tags whose multipath profiles are most similar. Tagoram [21] uses

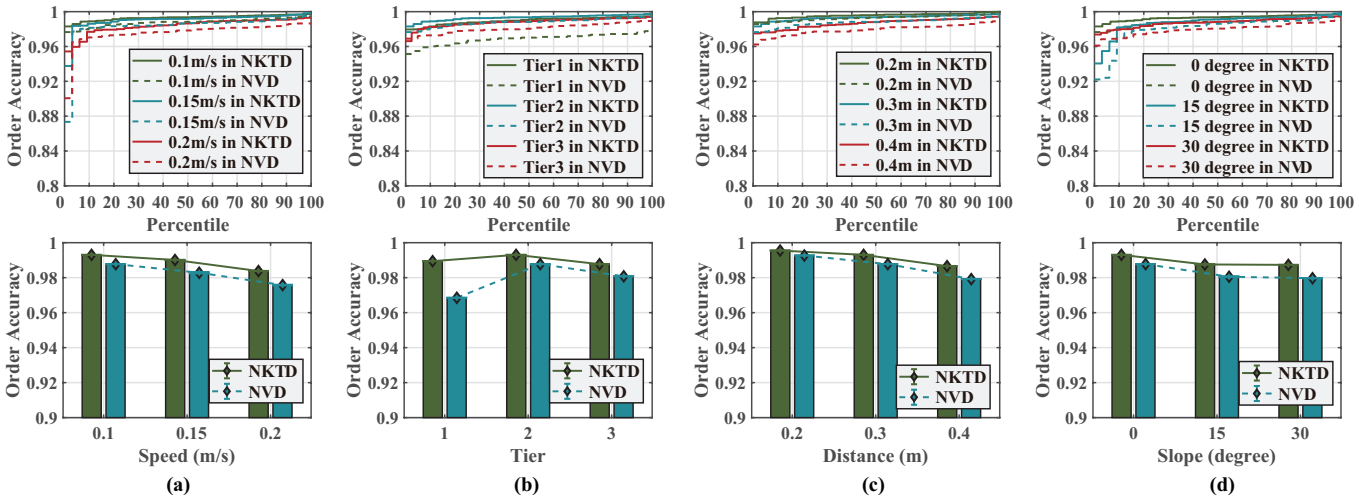

Fig. 5. The performance of RC-BI under different conditions. (a) Different speeds. (b) Different tiers. (c) Different distances. (d) Different slopes.

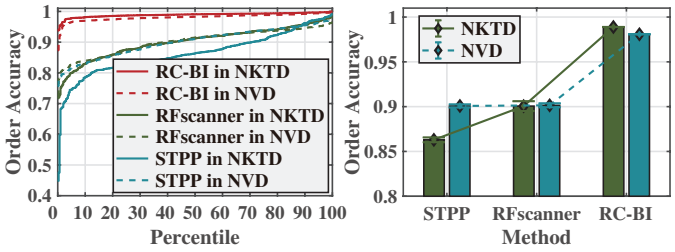

Fig. 6. Comparison with the STPP and RF-Scanner.

Differential Augmented Hologram (DAH) for high-precision real-time localization and tracking. OmniTrack [22] utilizes the phenomenon of tag polarization to simultaneously calculate location and orientation. MRL [1] and FILM [23] locate the tag using three points in the phase profile captured by mobile scanning. For some work like book inventory, relative localization plays a more important role. STPP [2] calculates relative localization by detecting the V-zone of RFID-based phase profile. In [4], through the modeling and fitting of phase profile, a fine-grained book relative localization is realized. Xu *et al.* [24] designed an inventory system towards HF RFID-based library, which uses fuzzy logic to estimate the location of books. TagSort [25] uses hybrid features to detect the peak of phase profile and utilizes a dynamic template selection algorithm to improve the robustness to tag-antenna distance. However, the accuracy of the above works is still limited, and they cannot avoid environmental interference on RFID signals.

B. Computer Vision-based approaches

Computer vision based methods are mainly focused on retrieving and recognizing objects or texts in the image. [5] and [6] first detect the book boundaries on the bookshelf image and extract the spine images, then input them into the OCR engine to get book texts. Yang *et al.* [7] located and recognized the text on the images of bookshelves with convolutional neural nets (CNN) and recurrent neural nets (RNN) to build a digital database. Considering the texts in the actual scene are often in irregular, Cheng *et al.* [26] proposed arbitrary orientation network (AON) to recognize arbitrarily oriented texts. Seo *et al.* [27] designed an OCR-based inventory algorithm and proposed deform-and-recover (DAR) learning technique to make the system robust against damaged images. However, the computer vision methods are sensitive to light conditions and incapable to distinguish objects with the same appearance.

C. Multi-modal-based approaches

There are also some multi-modal fusion methods for localization and tracking. TagVision [28] uses optical flow to detect moving objects and transform target location in the pixel image to the real world position. Then it matches each motion blob with an RFID phase sequence to achieve fusion localization. TagAttention [29] adopts the visual attention mechanism, in which the RFID signal assists the computer vision module to keep tracking the target with unknown appearances. Wang *et al.* [30] proposed an RF-Focus system, which uses computer vision to make up problems caused by RF phase periodicity, and match RFID and CV data to recognize and locate moving objects within range of interest on belt conveyor. Although these works present many solutions for multi-modal fusion, they are not suitable for the scenarios in this paper.

VII. Conclusion

In this paper, we proposed an RFID and CV fusion system for mobile robot assisted book inventory in the smart library. By designing a DNN network with multiple inputs and mixed data, we solved the challenge of filtering out tags from other tiers on bookshelf. To accurately acquire the book-name sequence from video, we proposed a novel video information extracting schema including frame capturing, image stitching, and coordinate transform. We proposed the concept of strong link to solve the challenge in aligning the sequences of two modals. RC-BI combines the robustness of RFID and the precision of CV. We conducted extensive experiments under different conditions in a library setting and proved that RC-BI achieves a tier filtering accuracy of 98.4% and an accuracy of 98.9% for book order, significantly outperforming the state-of-the-art. For our future work, we plan to deploy our system in other scenarios including smart warehouse and smart factory.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 62002259, 62072221, 62032017, 61772251.

REFERENCES

[1] X. Liu, J. Zhang, S. Jiang, Y. Yang, K. Li, J. Cao, and J. Liu, "Accurate Localization of Tagged Objects Using Mobile RFID-augmented Robots," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1273 – 1284, 2021.

[2] L. Shangguan, Z. Yang, A. X. Liu, Z. Zhou, and Y. Liu, "STPP: Spatial-Temporal Phase Profiling-Based Method for Relative RFID Tag Localization," *IEEE/ACM Transactions on Networking*, vol. 25, no. 1, pp. 596–609, 2016.

[3] R. Kumar and S. Vassilvitskii, "Generalized Distances between Rankings," in *Proc. of WWW*, 2010, pp. 571–580.

[4] J. Liu, F. Zhu, Y. Wang, X. Wang, Q. Pan, and L. Chen, "RF-Scanner: Shelf Scanning with Robot-assisted RFID Systems," in *Proc. of IEEE INFOCOM*, 2017, pp. 1–9.

[5] S. S. Tsai, D. Chen, H. Chen, C.-H. Hsu, K.-H. Kim, J. P. Singh, and B. Girod, "Combining Image and Text Features: A Hybrid Approach to Mobile Book Spine Recognition," in *Proc. of ACM MM*, 2011, pp. 1029–1032.

[6] M. Nevetha and A. Baskar, "Automatic Book Spine Extraction and Recognition for Library Inventory Management," in *Proc. of ACM WCI*, 2015, pp. 44–48.

[7] X. Yang, D. He, W. Huang, A. Ororbia, Z. Zhou, D. Kifer, and C. L. Giles, "Smart Library: Identifying Books on Library Shelves Using Supervised Deep Learning for Scene Text Reading," in *Proc. of ACM/IEEE JCDL*. IEEE, 2017, pp. 1–4.

[8] K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication*. John wiley & sons, 2010.

[9] K. Chawla, C. McFarland, G. Robins, and C. Shope, "Real-time RFID Localization Using RSS," in *Proc. of IEEE ICL-GNSS*, 2013, pp. 1–6.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of IEEE CVPR*, 2016, pp. 770–778.

[11] R. Girshick, "Fast R-CNN," in *Proc. of IEEE ICCV*, 2015.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. of IEEE CVPR*, 2017, pp. 2117–2125.

[13] Y. Wu, "Openpano," https://github.com/ppwwyyxx/OpenPano, 2020.

[14] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang *et al.*, "PP-OCR: A Practical Ultra Lightweight OCR System," *arXiv preprint arXiv:2009.09941*, 2020.

[15] seatgeek, "Fuzzywuzzy," https://github.com/seatgeek/fuzzywuzzy, 2020.

[16] M. Lenehan, "Octane sdk," https://support.impinj.com/hc/en-us/articles/202755268-Octane-SDK?_ga=2.162007420.1362618380.1627288515-675985672.1627288515, accessed July 26, 2021.

[17] Z. Zhou, L. Shangguan, X. Zheng, L. Yang, and Y. Liu, "Design and Implementation of an RFID-Based Customer Shopping Behavior Mining System," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2405–2418, 2017.

[18] H. Wang and W. Gong, "RF-Pen: Practical Real-Time RFID Tracking in the Air," *IEEE Transactions on Mobile Computing*, vol. 20, no. 11, pp. 3227–3238, 2020.

[19] Y. He, Y. Zheng, M. Jin, S. Yang, X. Zheng, and Y. Liu, "RED: RFID-Based Eccentricity Detection for High-Speed Rotating Machinery," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1590–1601, 2021.

[20] J. Wang and D. Katabi, "Dude, Where's My Card? RFID Positioning That Works with Multipath and Non-Line of Sight," in *Proc. of the ACM SIGCOMM*, 2013, p. 51–62.

[21] L. Yang, Y. Chen, X.-Y. Li, C. Xiao, M. Li, and Y. Liu, "Tagoram: Real-Time Tracking of Mobile RFID Tags to High Precision Using COTS Devices," in *Proc. of ACM MobiCom*, 2014, p. 237–248.

[22] C. Jiang, Y. He, X. Zheng, and Y. Liu, "Orientation-Aware RFID Tracking with Centimeter-Level Accuracy," in *Proc. of ACM/IEEE IPSN*, 2018, pp. 290–301.

[23] X. Bian, X. Wang, W. Cheng, X. Chen, J. Liu, and L. Chen, "FILM: Fine-Grained Book Localization with Mobile RFID Scanning," in *Proc. of IEEE ICPADS*, 2019, pp. 578–585.

[24] L. Xu, J. Liu, X. Wang, H. Gong, Y. Wang, and L. Chen, "HF RFID-based Book Localization via Mobile Scanning," in *Proc. of IEEE SECON*, 2020, pp. 1–9.

[25] J. Lai, C. Luo, J. Wu, J. Li, J. Wang, J. Chen, G. Feng, and H. Song, "TagSort: Accurate Relative Localization Exploring RFID Phase Spectrum Matching for Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 389–399, 2019.

[26] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: Towards Arbitrarily-oriented Text Recognition," in *Proc. of IEEE CVPR*, 2018, pp. 5571–5579.

[27] M. Seo, D. Kim, H. Kang, D. Cho, and D.-G. Choi, "OCR-based Inventory Management Algorithms Robust to Damaged Images," in *Proc. of IEEE ICRA*, 2021, pp. 12 889–12 895.

[28] C. Duan, X. Rao, L. Yang, and Y. Liu, "Fusing RFID and Computer Vision for Fine-grained Object Tracking," in *Proc. of IEEE INFOCOM*, 2017.

[29] X. Shi, M. Wang, G. Wang, B. Huang, H. Cai, J. Xie, and C. Qian, "TagAttention: Mobile Object Tracing without Object Appearance Information by Vision-RFID Fusion," in *Proc. of IEEE ICNP*, 2019.

[30] Z. Wang, M. Xu, N. Ye, R. Wang, and H. Huang, "RF-Focus: Computer Vision-assisted Region-of-interest RFID Tag Recognition and Localization in Multipath-prevalent Environments," in *Proc. of ACM IMWUT*, vol. 3, no. 1, 2019, pp. 1–30.