Active learning for image preparation of automatic vending machine (AVM) employing transfer learning method

Fang Li\(^{1a}\), Min Zeng\(^{2b}\), Jia Xiao\(^{3c}\), Xiaojun Li\(^{4d}\), Yuanyuan Li\(^{5e}\), Guosheng Hu\(^{6f}\)

\(^{1,2,3,4,5,6}\)Communications and Information Dept. Shanghai Technical Institute of Electronics & Information Shanghai, P. R. China

\(^{a}\)lfang@stiei.edu.cn, \(^{b}\)zengmin@stiei.edu.cn, \(^{c}\)xiaojia@stiei.edu.cn, \(^{d}\)lixiaojun@stiei.edu.cn, \(^{e}\)liyuanyuan@stiei.edu.cn, \(^{f}\)huguosheng@stiei.edu.cn

Abstract— In this paper, we employed active learning methods to prepare annotated images for our training of automatic vending machine (AVM) system, in order to minimize human annotation cost. Due to the tiny data of our system, transfer learning approach is used by implementing the already trained Yolov3-tiny model for COCO dataset as our training start. Also, we evaluated the effectiveness of 3 annotation strategies: smallest annotation area (SAA), largest annotation area (LAA) and moderate annotation area (MAA), for photos of top views from above. The results show that the idea of employing active learning methods to prepare annotated data is feasible. Also, the annotation strategy of MAA demonstrates the superior performance, for its enough object area and the least background area.

1. INTRODUCTION

There are many types of payments for automatic vending machine (AVM), such as non-mobile involved methods, like the traditional paper money or coin payment [1], credit card payment, and mobile-involved methods, like the short message payment [2], quick response (QR) code mobile payment, infrared (IR) connection payment [3], and price push payment. The mobile-involved payments are replacing the non-mobile payments, for the rapid development of multiple mobile functions.

Active learning approach has been used to reduce the effort of manual labeling for annotation [4][5][6][7]. It has been applied to image/video classification [8][9], text/web classification [10][11], and image/video retrieval [12][13]. Researchers proposed a high-order label correlation driven active learning approach to select the informative example-label pairs from which it learns, so as to learn an accurate classifier with less annotation efforts [14]. Ref [15] presented an approach for live learning of object detectors, in which the system autonomously refines its models by actively requesting crowdsourced annotations on images crawled from the Web.

Transfer learning is the method aiming to solve the problem that the training and data are not in the same feature space and don’t have the same distribution in machine learning [16][17]. Researchers proposed a transfer learning framework to transfer cross-domain image knowledge for the new computer vision tasks [18]. Ref [19] presented a dimensionality reduction method to find a latent space, in which the distance between distributions of the data in different domains in a latent space is minimized.

In this paper, we presented the process of data preparation for our AVM system, using active learning and transfer learning methods. To reduce human labor of labelling, the active learning
approach was employed by repeating the process of training 11 labeled photos and 33 unlabeled photos, checking the obviously wrongly labeled photos, for 3 times. Due to the limited item photos, we employed the already trained Yolov3-tiny mode for CO-CO data set as the training start to train our data. We evaluated the performance of 3 types of annotation strategies: smallest annotation area (SAA), largest annotation area (LAA) and moderate annotation area (MAA), respectively. The results show that the method of employing active learning and transfer learning to prepare labeled data is feasible, and the annotation strategy of MAA demonstrates best performance.

In the next section, the AVM system is introduced. Then, the data preparation is demonstrated. There are four parts: image composition, annotation strategy, active learning methods and transfer learning methods. In the end, the results and conclusion are given.

2. THE AVM SYSTEM

2.1. AVM system flowchart
The flowchart of our AVM systems is shown in Fig. 1: A client sent the door opening request on an APP, then the request is sent to the AVM, so the door can be opened. After picking items and closing the door, the photos are taken and sent to the clouds. The picked items are recognized by comparing photos before and after the door opening. In the end, the priced are pushed and charged to the original cell.

2.2. Structure of AVM
To minimize the cost, we employed the general freezer without traditional mechanical motion equipment in our AVM, as shown in Fig. 1. The cooling space is divided into 5 parts vertically, and there is a camera at each ceiling center. Each cooling space can hold up to 42 can drinks (7 columns, 6 rows). The photo captured by the cameras record real time item changes, and all the on-cloud transaction data make the refined management and analysis possible. With the transparent glass door, it’s easy for consumers and goods staffs to see all the items, make pick and place items easy.

2.3. Selling items of AVM
Our AVM is designed to be deployed in the office building, so the categories of the items are limited to drinks and snakes. Considering the compact packing and uniformed volume of can drinks, we choose types of can drinks as our items in our earlier experiments, such as Cola, Fanta, Sprite and etc, in order to make sure that the fixed AVM inner space can be efficiently used and is occupied with maximal items.

3. DATA PREPARATIONS FLOWCHART
The learning flowchart is composed of two parts: image preparation and model training, as shown in Fig. 2. To reduce the cost of annotation, we employed active learning method to enlarge the quantity of the labeled photos. Also, we used the transfer learning approach to train model for recognizing items.
3.1. Image compositions
To evaluate the feasibility of the learning system, we simplify the item categories to only 3 in the learning. In order to obtain the maximum learning performance, we should provide the photos as many as possible. Therefore, we investigate the possible photos that the camera could take in real situations for the proposed 3. We arrange the can drinks in the following steps: Firstly, we located the cans in various orientations, leaving the ring pulls pointing to up, right, down and left alternatively, as shown in Fig. 3, in which the arrows denoting the orientations. The four orientations are chosen to represent possible location orientations. Secondly, we placed the cans of one item on different locations, such as full placement and not full placement alternatively, to simulate possible situations, as shown in Fig. 4. Thirdly, all three types of cans are presented, as shown in Fig. 5. Lastly, we presently possible placements with three types of cans altogether, as shown in Fig. 6.

With the considerations of four aspects mentioned above, we took 44 photos for 3 types of cans, in total.

3.2. Different annotation strategy
For the AVM system, items often block the view of each other when they are fully placed. Blocking, makes our object annotation containing background. When more the blocking, the more background will be annotated to the object, which will reduce the learning performance. To investigate the influence of annotation area to the system, we evaluated the performance of 3 types of annotations strategies. They are smallest annotation area (SAA), largest annotation area (LAA) and moderate annotation area (MAA), respectively.

SAA means the bounding box only contains the top silver area of the can and not any background at all, as shown in Fig. 7 (a). LAA is the annotation method that the object is fully contained within the bounding box, which induces the most background area, as shown in Fig. 7 (b). MAA contains not only the top silver area but also more of its top, especially the colorful upper area, as shown in Fig. 7 (c). In fact, MAA is implemented by the principle that taking the maximum area of the object, in the meanwhile, the minimum area of the background.

3.3. Active learning
To reduce annotation cost, we employed active learning approach to provide enough labeled data for our final comparison and recognition, as depicted in Fig. 2. That is, we labeled a part of our data, then sent them to the model to label the rest of the data. By selecting and checking the results of the learning manually, we re-annotated the wrongly labeled and unlabeled cans by human annotator, and put them into the train data, then resent them back to the model again. After repeating the process for about 3 times, the output 44 photos are all labeled, which can be used as labeled train data for latter recognitions.

We only labeled 11 photos, among the total 44 photos. Be noted that, although only 11 photos are labeled, there are multiple labeled tags in one photo, and there are up to 42 labeled tags in one labeled photo when cans are fully placed.
3.4. Transfer learning

As each camera is set on the middle of the ceiling for each vertically separated space and the positions of the camera remain unchanged, so the photos taken are views from above of each individual cooling space, as shown in Fig. 1. Also, in order to minimize the empty area, we installed several thin and short division plates and then each item shelf contains 7 columns, so the cans are placed on the fixed positions of shelf. The identical visual angle of the camera and the stable position of the cans, makes the limited picture numbers of the selling items top views, and thus induce the limited dataset for our learning model.

![Figure 3: one item placements for various orientations: (a) up, (b) right, (c) down, (d) left](image1)

![Figure 4: one item placements for various situations: (a) fully, (b) not fully, (c) not fully, (d) not fully](image2)

![Figure 5: three items placements: (a) Fanta, (b) Soda, (c) Cola](image3)

![Figure 6: three items mixed placements: (a) placement_1, (b) placement_2, (c) placement_3, (d) placement_4](image4)

![Figure 7: Different annotation strategies: (a) SAA (smallest annotation area), (b) LAA (largest annotation area), (c) MAA (moderate annotation area)](image5)
Since the dataset for our AVM is little, we employed transfer learning approach in our system. Transfer learning uses a model which is already trained for one task, to another task. It makes the learning model possible when the training data are little. For example, in the area of computer vision, transfer learning only re-trains the later layers which are relative to task-specific feathers, in the meanwhile leaves the earlier and middle layer untrained which are edges and shapes feathers. It is realized by transferring the weights that trained for one problem, as the train beginning for another problem.

We employed the pretrained YOLOv3-tiny model for COCO dataset as our training beginning. The COCO (common objects in context) dataset contains 80 categories with over 80 thousand training images and over 40 thousand validation images. It is a large-scale dataset for object detection, segmentation, person detection, stuff segmentation, and caption. Yolov3 is a real-time object detection model, which uses convolutional neural networks to classify and locate objects. Since the GPU capacity of our system, we chose the Yolov3-tiny version in our training.

4. EXPERIMENTAL RESULTS
We employed the Yolov3-tiny as the learning model, with the uniform parameters. we re-trained the Cloud service of Tencent company to implement the learning. The hardware is GPU computing type of GN7, and it has capacity of 4 cores, 20GB RAM and 4GB VRAM.

The learning results of MAA, SAA, and LAA are depicted as follows. The F1 parameter of MAA exhibits better performance than the SAA and LAA situation for all the epochs, as illustrated in Fig. 8 (a), and F1 of MAA reaches to 0.85 for the final 300 epochs. Also, the mAP_0.5 parameter of MAA demonstrates superior performance than the SAA and MAA case for epoch lower than 200, and the gap between them gets smaller for epoch larger than 250, as shown in Fig. 8 (b). The precision shows similar characteristics with the F1: the MAA demonstrates better performance than SAA and LAA for all epochs, as shown in Fig. 8 (c). The recall of MAA exhibits superior performance than the SAA and LAA situation when epoch is lower than the 150, and they get close when epoch is larger than 200, as depicted in Fig. 8 (d).

In all, the MAA demonstrates overall superior characteristics than SAA and LAA, for the reason that MAA presents enough object, in the meanwhile avoids background as much as possible.

5. CONCLUSION
The learning results show that the idea of employing active learning for data preparation is practicable. By training Yolov3-tiny model with only 11 photos are manually annotated, 44 labeled photos are available after the learning. Also, the annotation strategy of MAA demonstrates superior performance than SAA and LAA. The parameters of F1, mAP_0.5, precision and recall show that the training model are effective for epoch larger than 200.

ACKNOWLEDGMENT (HEADING 5)
The paper is sponsored by Shanghai Municipal Education Commission, the funds No. is E-002-20-GQ0501-27.
REFERENCES
[1] V. Kanagasabapathi, K. Naveenraj, V. Neelavarnan and S. Naveen Raj, "Automatic chocolate vending machine," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, is India, 2019, pp. 584-587, doi: 10.1109/ICACCS.2019.8728319.
[2] W. Zhang and X. L. Zhang, "Design and Implementation of Automatic Vending Machine Based on the Short Massage Payment," 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), Chengdu, 2010, pp. 1-4, doi: /10.1109/WICOM.2010.5600192.
[3] S. B. Z. Azami and M. Tanabian, "Automatic mobile payment on a non-connected vending machine," Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No.04CH37513), Niagara Falls, Ontario, Canada, 2004, pp. 731-734 Vol.2, doi: 10.1109/CCECE.2004.1345218.
[4] Y. Deng et al., "A New Framework to Reduce Doctor’s Workload for Medical Image Annotation," in IEEE Access, vol. 7, pp. 107097-107104, 2019, doi: 10.1109/ACCESS.2019.2917932.
[5] K. Wang, D. Zhang, Y. Li, R. Zhang and L. Lin, "Cost-Effective Active Learning for Deep Image Classification," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 12, pp. 2591-2600, Dec. 2017, doi: 10.1109/TCSVT.2016.2589879.
[6] Y. Huang, Z. Liu, M. Jiang, X. Yu and X. Ding, "Cost-Effective Vehicle Type Recognition in Surveillance Images With Deep Active Learning and Web Data," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 1, pp. 79-86, Jan. 2020, doi: 10.1109/TITS.2018.2888698.
[7] K. Wang, D. Zhang, Y. Li, R. Zhang and L. Lin, "Cost-Effective Active Learning for Deep Image Classification," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 12, pp. 2591-2600, Dec. 2017, doi: 10.1109/TCSVT.2016.2589879.
[8] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang and D. Xu, "Image Classification by Cross-Media Active Learning With Privileged Information," in IEEE Transactions on Multimedia, vol. 18, no. 12, pp. 2494-2502, Dec. 2016, doi: 10.1109/TMM.2016.2602938.
[9] D. Tuia, F. Ratle, F. Paciﬁci, M. F. Kanevski and W. J. Emery, "Correction to ‘active learning methods for remote sensing image classiﬁcation’, IEEE Trans. Geosci. Remote Sens., vol. 48, no. 6, pp. 2767, Jun. 2010.
[10] D. Hakkani-Tur, G. Tur, M. Rahim and G. Riccardi, "Unsupervised and active learning in automatic speech recognition for call classiﬁcation," 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Que., 2004, pp. I-429, doi: 10.1109/ICASSP.2004.1326014.
[11] D. Cai and X. He, "Manifold Adaptive Experimental Design for Text Categorization," in IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 4, pp. 707-719, April 2012, doi: 10.1109/TKDE.2011.104.
[12] B. Demir and L. Bruzzone, "A Novel Active Learning Method in Relevance Feedback for Content-Based Remote Sensing Image Retrieval," in IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 5, pp. 2323-2334, May 2015, doi: 10.1109/TGRS.2014.2358804.
[13] S. C. H. Hoi, Rong Jin, Jianke Zhu and M. R. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-7, doi: 10.1109/CVPR.2008.4587350.
[14] B. Zhang, Y. Wang and F. Chen, "Multilabel Image Classification Via High-Order Label Correlation Driven Active Learning," in IEEE Transactions on Image Processing, vol. 23, no. 3, pp. 1430-1441, March 2014, doi: 10.1109/TIP.2014.2302675.
[15] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," CVPR 2011, Providence, RI, 2011, pp. 1449-1456, doi: 10.1109/CVPR.2011.5995430.
[16] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[17] L. Shao, F. Zhu and X. Li, "Transfer Learning for Visual Categorization: A Survey," in IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 5, pp. 1019-1034, May 2015, doi: 10.1109/TNNLS.2014.2330900.

[18] H. Wang, Feiping Nie, H. Huang and C. Ding, "Dyadic transfer learning for cross-domain image classification," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 551-556, doi: 10.1109/ICCV.2011.6126287.

[19] S. J. Pan, J. T. Kwok and Q. Yang, "Transfer learning via dimensionality reduction", Proc. 23rd Nat. Conf. Artif. Intell. (AAAI), pp. 677-682, Jul. 2008.