Customer Activity Recognition System using Image Processing

Maria Waqas[1]†, Mauizah Nasir[1], Adeel Hussain Samdani[2], Habiba Naz[1], Maheen Tanveer[1]

[1] Department of Computer and Information Systems Engineering, NED University of Engineering and Technology,
Karachi 75270, Pakistan.
[2] Department of Computer Engineering, Sir Syed University of Engineering and Technology, Karachi 75300, Pakistan†
corresponding author [mariaw@neduet.edu.pk]

Summary
The technological advancement in computer vision has made system like grab-and-go grocery a reality. Now all the shoppers have to do now is to walk in grab the items and go out without having to wait in the long queues. This paper presents an intelligent retail environment system that is capable of monitoring and tracking customer's activity during shopping based on their interaction with the shelf. It aims to develop a system that is low cost, easy to mount and exhibit adequate performance in real environment.

Key words: *Image processing; Convolutional neural network; Customers' tracking; Deep learning; Computer vision*

1. Introduction

The study of human behavior has become one of the most interesting topic for researchers and in recent years a lot of work has been done in computer vision, from smart cars to smart cities many milestones have been achieved. Monitoring customers' behavior during shopping has proved to be very useful idea considering the busy schedule of people who don't like standing in long queues to pay the bill after shopping. This paper aims to develop a system that uses image processing techniques using camera to detect the activity of customers during shopping about whether or whether not the item has been picked from the shelf. If the item has been picked the system is trained to add this item to that customers' cart and generate the bill after the customer has left the shelf area. The project also aims to propose a database server that maintains the record of its customers along with the count of items they have bought.

The image and video processing techniques to recognize human actions have become more successful and accurate by the use of neural networks. One common problem faced in every retail environment is of check-out process. Since life has become fast and people have become more conscious about their time management so in such scenario standing waiting for their turn behind the long queues of customers is so tiring and time consuming. This results in lack of customer satisfaction. Hence in such scenario a system that identifies its regular customer and not only keep track of their activity but also automatically generate the billing details that can later be paid by customer through card or on spot payment can prove to be of high value.

We are interested in applying image processing techniques along with deep neural network application for customer identification and tracking. A lot work has already been done in this era. Two related research directions can be identified, one in which similar task is accomplished by using both weight sensors and computer vision that is successfully running in smart retail environment of Amazon Go [1]. The other focus on only carrying this out through computer vision with neural network to help customers shop conveniently. The second approach is simple to implement and also cost effective because sensors are usually complex to install and also need maintenance that will add up to overall project cost. Here convolution neural network (CNN) has been applied for feature extraction from videos [2], as it is a good approach to identify visual patterns from videos and it takes very little pre-training work. CNN is the most used network in the era of digital image processing for human action recognition.

Other similar significant research in the area includes human hand detection for grab-and-go grocery [1], [3]. Projects like smart cars uses computer vision techniques very efficiently [4]. Image processing has proved to be very creative and useful solution for many types of application. The process involve extracting some specific features from images and using this information for segmentation or manipulation. This project is based on research techniques taken from these papers and modifying them according to the system requirements.

2. Methodology

As mentioned before the system is purely based on computer vision technology relying on image processing and having a back-end server as database. The project aims to correctly recognize the activity of customer present near

Manuscript received September 5, 2021
Manuscript revised September 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.9.7

the shelf and based on this prediction generate the bill as the customer leaves the window. It is able to keep track of the exact count of the things picked from the shelf. For simplicity, we have fixed the type of items on the shelf. An applicable description of the system is given in Fig. 1. The system first captures images via video strean from the wifi camera. Each image is assigned a unique ID and is stores in the backend database. The saved images are then used for testing and validation of the CNN system. The process is further described in detail in the following subsections.

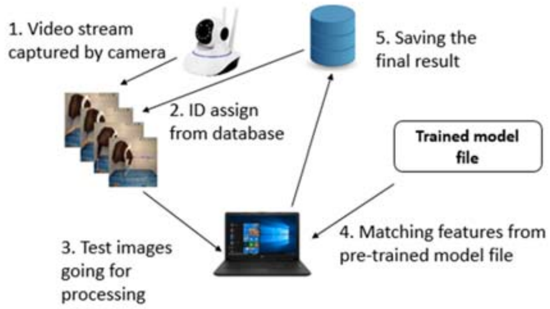

1. Video stream captured by camera
5. Saving the final result
2. ID assign from database
Trained model file
3. Test images going for processing
4. Matching features from pre-trained model file

Fig. 1 An overview of the systems

2.1. Image capturing

The system has been trained over custom environment using deep learning model, hence the first task is to collect a set of images to feed into CNN. For this purpose, a WiFi IP camera has been deployed over the top view. This view has been chosen because it is able to cover the whole scenario including customers and the shelf. Two types of models have been trained, one in identify whether a person is present near the shelf or not and the second model detect the activity performed by purchaser. Each model consists of two classes and each class has been trained over 2000 trained images and 200 validation image sets. Fig. 2 shows the view captured by these images.

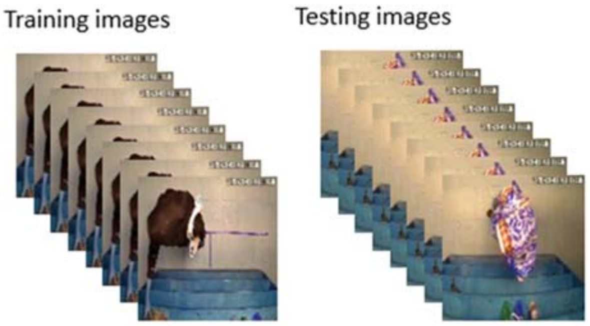


Fig. 2 Training and testing images showing the top view of the shelf and the customer.

2.2. Model Training

The collected images are then fed into a CNN which consist of several layers. Each model is trained over 15 epochs with learning rate of 0.003. A brief description of various layers and their operating functions is given below.

According to [5] the first layer, which is the convolutional layer, plays the most vital role in CNN. To understand this let's consider an image of size 6x6 and a kernel of size 3x3. This kernel consists of numerical values and it glides all over the image and scalar product is calculated for each value in kernel. The final output is also called activation map. Convolution layer also tends to reduce the complexity for further processing. Next layer that we used is the ReLU Layer. This layer is generally used as an activation function. There are other activation functions as well but ReLU stands out to be the faster and reliable function in comparison with sigmoid and softmax. Then comes the pooling layer. The primary purpose of this layer is to reduce the computations required for training [6]. It typically operates over the activation maps obtained previously and apply the "max" function. It has been estimated that pooling layer can reduce the activation maps by 25% [5] while preserving the depth to its typical size. Lastly, comes the fully connected layer. As the name suggests this layer implies that every neuron in the previous layer must be fully connected to the neurons in next layer. The high-level output obtained from convolutional and pooling layer is passed through fully-connected layer which further classify them into various classes based on training datasets. Use of this layer increases the probability of better prediction and classification. Brief summary of the training model used is shown in Table 1.

Table 1: The training model.

Layer (type)	Output Shape	Number of params
conv2d_1 (Conv2D)	(None, 126, 126, 32)	896
max_pooling2d_1 (MaxPooling2)	(None, 63, 63, 32)	0
conv2d_2 (Conv2D)	(None, 61, 61, 64)	18496
max_pooling2d_2 (MaxPooling2)	(None, 30, 30, 64)	0
conv2d_3 (Conv2D)	(None, 28, 28, 128)	73856
max_pooling2d_3 (MaxPooling2)	(None, 14, 14, 128)	0
conv2d_4 (Conv2D)	(None, 12, 12, 128)	147584
max_pooling2d_4 (MaxPooling2)	(None, 6, 6, 128)	0
flatten_1 (Flatten)	(None, 4608)	0
dropout_1 (Dropout)	(None, 4608)	0
dense_1 (Dense)	(None, 512)	2359808
dense_2 (Dense)	(None, 1)	513
Total params: 2, 601, 153 Trainable params: 2, 601, 153 Non-trainable params: 0		

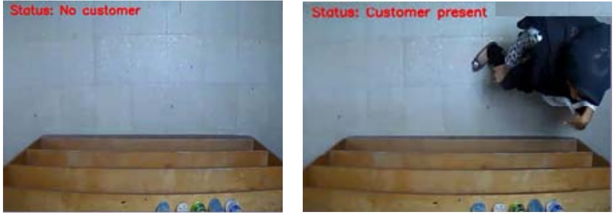
2.3. Saving the Records

This module is linked with back-end server where all the record files are maintained. For this purpose, Django framework has been used as it goes the best aside with python. The records obtained from previous phase will be saved automatically that can later be viewed through web-based application. The system is also able to generate bill for a particular customer ID showing the product name, quantity and total billed amount.

3. Testing and Results

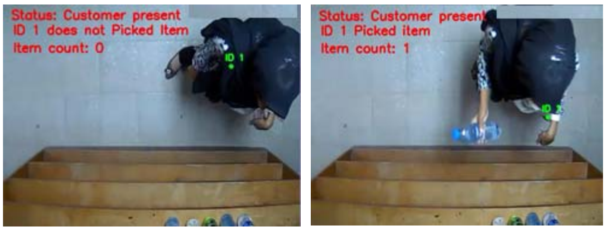
Since the system is low cost and mainly based on computer vision technology so as for hardware a single smart WiFi IP cameras has been integrated, so that it can work over recorded environment as well as on real-time scenario. For software requirements the code has been deployed on python3.5 using Pycharm IDE. Some important pre built-in libraries has been used for code compilation include Opencv, Dlib, Tensorflow and Keras. As for application purpose Django framework is used for front-end designing and for back-end Mysql server is integrated using libraries pymysql and mysqlclient.

After completing the training procedure a model file is obtained. When the system is executed, the file is loaded and the results saved in this file are used for testing and prediction purpose. After the system detects the presence of customer in the frame as can be seen in Fig. 3, it first prompts the admin through a GUI to enter the ID of corresponding shopper, if the entry exists in the database only then the system proceeds. The ID associated with the customer is used for tracking. Now the program detects the activity of customer about how many items has been picked by purchaser as shown in Fig. 4. After testing the system many times, both on recorded video and on real time it has been concluded that if all the desired conditions meet the system works with 75% accuracy on real-time and with 85% accuracy on recorded video.



(a) (b)

Fig. 3 Detecting customer presence in the frame. (a) No customer is present. (b) A customer is present



(a) (b)

Fig. 4 Detecting customer activity. (a) No item picked. (b) 1 item picked

Fig. 5 shows the results produced by the system for progressive epochs. The training accuracy obtained for the system is 94%, whereas that for validation is 89%. The training and validation losses decrease as the epochs progress. The system is able to perform satisfactorily under constrained environment.

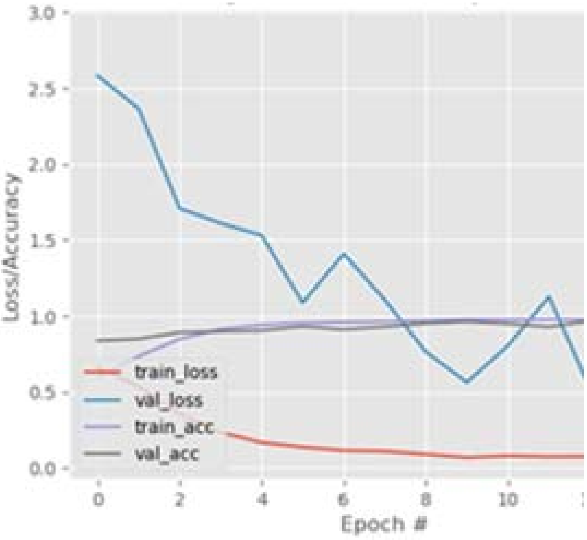


Fig. 5 Training loss and accuracy on dataset

4. Conclusion

An intelligent retail environment system that detects the shoppers and recognize their activities has been proposed in this paper. The system is tested to provide the count of items picked by purchasers and generate their invoice without human interaction with satisfactory results. The system is also able to generate customer's bill. The system can be improved by integrating weighting sensors which would help detect scenarios if customer places the item back on the shelf. Further it can be extended to work over multiple people standing simultaneously near the shelf.

References

[1] Kong, L., Fan, X., Lussier, J., *Item Removal Detection in Retail Environments with Neural Networks.* cs231n.stanford.edu.

[2] Zhao, C., Han, J.G. and Xu, X., *CNN and RNN Based Neural Networks for Action Recognition*. In Journal of Physics: Conference Series, vol. 1087 (6), pp. 062013 (2018).

[3] Qiu, X. and Zhang, S., *Hand Detection for Grab-and-Go Groceries*. Technical report (2017). cs231n.stanford.edu

[4] Chen, Y., *A new design on image processing scheme for smart car*. In Journal of Physics: Conference Series, vol. 1087 (6), p. 062012 (2018).

[5] O'Shea, K., Nash, R., *An introduction to convolutional neural networks*. arXiv preprint arXiv:1511.08458 (2015).

[6] Murphy, J., *An overview of convolutional neural network architectures for deep learning*. Microway Inc (2016).