Optical Character Recognition System for Business Cheque in Insurance Claim Payment Process

Chinnawat Ngamsom[1] and Passakorn Phannachitta[2]

[1]Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand

chinnawat.ngamsom@gmail.com

Abstract. This independent study aims to develop a data pipeline system that is able to transform a printed standard of business cheque image into digital numeric data using the OCR technique. This system developed specifically to enhance the efficiency of the data input process of the insurance claim payment process. The evaluation of the system is in two folds. The first one is to evaluate the efficiency among different algorithms used in building the OCR system based on accuracy and runtime. The selected algorithms are k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Gradient Boosting Machine (GBM) respectively. GBM was found to be the most accurate and it demanded the least runtime among the three techniques. The second one is to appraise based on the result of evaluation survey from 10 experts who are either the developers or the person in charges of claiming process in the insurance industry. The survey result shows that both of the accuracy and speediness of the system developed is outstanding and satisfaction. Therefore, it can be concluded that the purposed system can increase capability of data input process of the insurance claim payment process.

Keywords: Optical Character Recognition, Histogram of Oriented Gradient, Image Classification, Business Cheque, Claim Payment Process.

1 Introduction

Nowadays, digital platforms have generated value of the insurance industry by connecting customers, agents and business partners together [1]. The insurance integration systems on digital platforms make business ready for competition and easy to use for customers and agents. One of the important processes is the payment process connecting the financial information with business agent partner's websites. Claim payment process is a part of the insurance financial process. When claim happens, an agent needs to send invoice to the insurance company, and records a business cheque information including routing number, check number and accounting number for payment transaction that paid to insurance company on web application. The issue can happen with human error when an agent inputs the wrong business cheque information into the system. If the issue happens, it makes delay the payment process because auditing in the accounting process needs to start at beginning again. The author sees the opportunity to integrate Optical Character Recognition (OCR) technique to convert text on business cheque into machine-encoded text by computer [2]. OCR is able to classify image character with machine learning concept and then send out information via Web API instead of input information by a human. The system will able to reduce defects that can happen during input information to insurance claim payment process. In the experimental setup, the author uses an image dataset from Institute for Development and Research in Banking Technology (IDRBT) [3], including 112 cheque images. The workflow of proposed system described in Section 2. The experimental results are shown in Section 3. Next, the discussion and conclusion has been made in Section 4 and Section 5.

2 The Proposed System

The system is separated into two parts: OCR and Web API, respectively. The OCR processes in this study is developed following the basic character recognition workflow proposed by Purohit et al. [4], in which the subprocesses are illustrated in Fig. 1

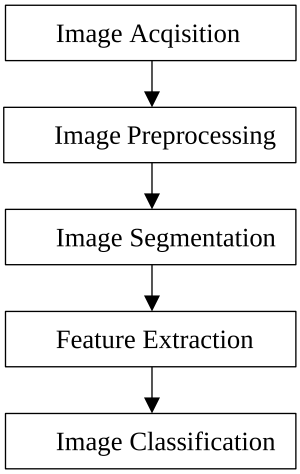

Fig. 1. Diagram of the character recognition development workflow. [4]

A. Image Acquisition

This process is getting a business cheque from the image and crop image remaining about consideration area. In this work, we consider only the number at the bottom of business cheques, including routing number, check number, and accounting number. A business cheque image will be cropped to be the bottom 18% of the entire cheque's height.

B. Preprocessing

In this work, the processing covers noise reduction, converting image to black and white image, aligning document, and resizing the image. For noise reduction, the author uses the method named fastNMeanDenoising [5] from the cv2 Python library to reduce noise from an image. Note that, this method requires an explicit mathematical function for the filter parameter and the author opts for Standard Deviation (SD). For converting an image to black and white, the author chooses the erode and dilate techniques, which is done by using the morphology and Blackhat methods of the cv2 library [6]. After that, the author further uses the Close technique in morphology to reduce small spots inside the foreground objects [6]. An example product up to this step is shown in Fig.2

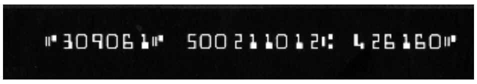

Fig. 2. An example results of morphology method

C. Segmentation

Segmentation is to separate each individual character in an image, which requires three working steps. The first step is to find the threshold between object and background. The second step applies the best threshold value to obtain the contour. For image thresholding, the author selects the Binary and Otsu technique, the for the contour, the author selected the findContours method of the cv2 library [7]. Finally, the unwanted background is removed by a technique named bounding box of the cv2 library [8]. An example of the segmentation product is region of interest (ROI), as shown in Fig.3.

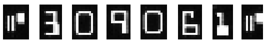

Fig. 3. An example results of segmentation product is region of interest (ROI)

D. Feature Extraction

The goal of feature extraction is to extract that pattern, which is most pertinent for Classification. As suggested by Moghneih [13], this study applies the Histograms of oriented Gradients (HOG) for extracting the essential features for each individual character. Applying the technique, all the images of each character are resized to 50x50 pixel then the resized image will be an input of the HOG method with following parameters: pixels per cell = 10 x 10 pixel, cells per block = 5 x 5 pixel and orientations

= 8. These are the same configuration as the work of Moghneih [13]. The result from HOG will return vectors of features that are later used to train the system.

E. Classification

After images are extracted and labelled as 0-9, the author found that labels are imbalance. Therefore, the author applies gaussian and speckle [15] to generate image of character by using font that appears on business cheque [14]. The generated character images are covered by random noise with gaussian and speckle techniques in order to make it similar with the real defect in business cheque number. Feature extraction also applies to these generated images. An example results of the generated characters are shown in Fig.4.

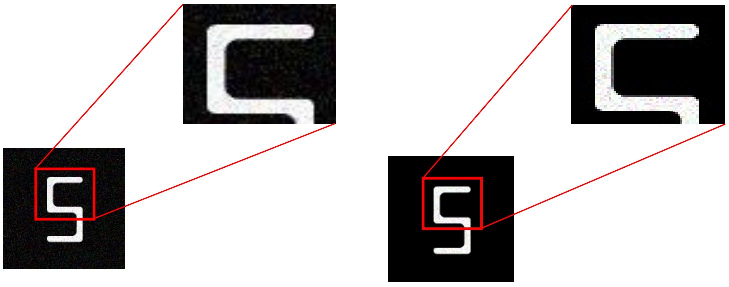

Fig. 4. Random noise with gaussian (Left), and speckle (Right)

All features are extracted and given as an input to the trained classifier including KNN, SVM and GBM. Data are separated into two parts. 70% of the image data will be reserved for train classifiers. The other 30% of the image data will use in blind test. Character images are generated from font appearing on the business cheque to train classifiers. The train dataset is rebalanced by adding the generated image until all the classes are balanced. Then, the train dataset is built by resampling 100 characters per class without replacement. Note that the Hyper-parameter Optimization technique (Hyperopt) [9] is applied for finding the best parameter for each classifier in the 10 fold-cross-validation process [10]. The parameter sets examined by Hyperopt is shown in Table 1.

Table 1. Parameter sets for Hyperopt [11].

Classifier	Hyperparameter Set
KNN	Number of neighbors = {1,3,5}, Metric = { euclidean, manhattan }
SVM	C = {1,10,100,1000}, Kernel = { linear, rbf }
GBM	Number of estimators = { 2,4,8,…,256 }, Learning rate = {0.001,0.01,0.1}

After the classifiers are completely trained, the blind test over the remaining 30% of the data will be undertaken for validation.

After the best classifier is determined by the previous process, the author develops the Web API for the real world application. This Web API will be connected the insurance payment process. The Web API receives cheque image as input and carry out image acquisition, preprocessing, segmentation, feature extraction and image classification. Classifier compares the input feature with stored pattern and find out the best matching class for a specific input. After that the digit character will form as group of json format. Group will separate from number of digits in each group [12]. For example, the group of routing number will contain 9-digit codes, excluding the character symbol surrounding the numbers. The workflow of OCR processes on Web API is shown in Fig.5.

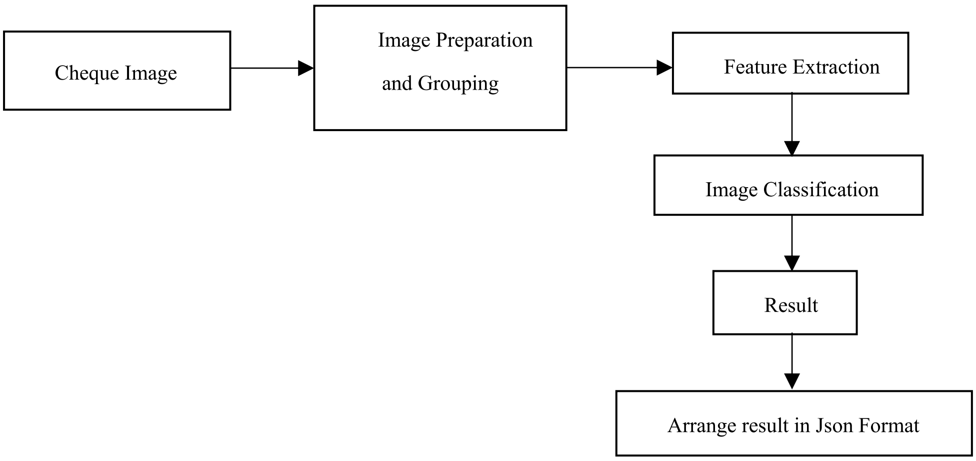

Fig. 5. Diagram of OCR processes on Web API

3 Experimental Result

The evaluation of the system is separated into two parts. The first part is the comparison of accuracy and runtime among 3 classifiers for train data and blind test data. The comparison results of 3 classifiers is show in Table 2 and Table 3.

Table 2. The comparison results of 3 classifiers with train data.

Classifier	Average of Runtime (m)	Average of Accuracy (%)
KNN	0.152	99.002
SVM	0.211	99.003
GBM	20.243	99.007

Table 3. The comparison results of 3 classifiers with blind test data.

Classifier	Average of Runtime (s)	Average of Accuracy (%)
KNN	0.024	98.883
SVM	0.015	99.760
GBM	0.013	100.000

The second part of the evaluation is based on a survey from 10 experts who are either the developers or the person in charges of claiming process in the insurance industry. There are 7 topics in the survey related to accuracy, speediness, usability and performance. Score of survey is based on Likert scale. The result of survey is shown in Fig. 6.

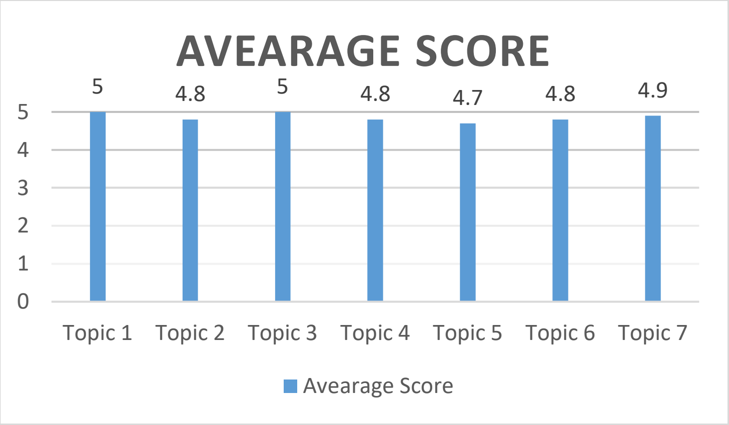

Fig. 6. The result of survey from 10 experts after demonstration for the system

4 Discussion

The training results of Table 2 show that KNN took the least runtime but GBM took the most. The training time of GBM was sacrificed to finding the best learning rate parameter. However, this training time can trade-off for the highest accuracy, as presented in Table 3. Anyhow, for the prediction, GBE demands the least time usage for prediction, followed by SVM and KNN. For the survey results presented in Fig.6, the results related to the accuracy and speediness of the system developed could obtain 5 out of 5 scale based on Likert score. The other noteworthy result is topic 7, which is about well-appointed of the required information. The author sees that the system can provide the expected accuracy and speed of prediction for users. But some information still need that it can imply that users still need more information that the author could not provide through Web API.

5 Conclusion

From the evaluation of among 3 classifiers, GBM was found to be the best by means of accuracy and prediction speed. The proposed system also has its usability evaluated by getting from survey by 10 experts, including 3 of developers, 3 of quality assurance specialists, 2 of business analysts, and 2 of customer support specialists. The survey result indicates the good average scores in every topic. The both of the accuracy and speediness of the system developed topic could obtain 5 out of 5 scale based on Likert score. Therefore, it can be concluded that the purposed system can increase capability of data input process of the insurance claim payment process. Lessons learned from the study includes the cause of invalid prediction of classifier. The author found that noise located nearly digit number on a business cheque can be an effect for accuracy. This makes the author decides to add the noise reduction process to image preprocessing step for reducing noise on business cheque image. The future work includes grouping of number can be done by using surrounding of symbol. Furthermore, the work principle of OCR process can apply to other areas of characters in business cheques.

References

1. Digital Transformation in Insurance, https://www.adacta-fintech.com/blog/digital-transformation-in-insurance-apis-platforms-and-ecosystems, last accessed 2020/05/21
2. Patel, I., Jagtap, V., & Kale, O. A Survey on Feature Extraction Methods for Handwritten Digits Recognition. International Journal of Computer Applications, pp. 14-15 (2014)
3. IDRBT Cheque Image Dataset, https://www.idrbt.ac.in//icid.html, last accessed 2019/11/5
4. Purohit, Ayush & Chauhan, Shardul. A Literature Survey on Handwritten Character Recognition. International Journal of Computer Science and Information Technology. 7. pp.1-5. (2016)
5. Noise Reduction using OpenCV, https://docs.opencv.org/3.4/d5/d69/tutorial_py_non_local_means.html, last accessed 2020/01/09.
6. Morphology Techniques using OpenCV, https://docs.opencv.org/trunk/d9/d61/tutorial_py_morphological_ops.html, last accessed 2020/08/2.
7. Image Thresholding using OpenCV, https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_thresholding/py_thresholding.html, last accessed 2020/05/12.
8. Structural Analysis and Shape Descriptors, https://docs.opencv.org/3.4/d3/dc0/group__imgproc__shape.html#ga103fcbda2f540f3ef1c042d6a9b35ac7, last accessed 2020/05/12.
9. Find the best parameter using Hyperopt, https://github.com/hyperopt/hyperopt/wiki/FMin, last accessed 2020/06/15.
10. k-fold-Cross-Validation, https://machinelearningmastery.com/k-fold-cross-validation/,last accessed 2020/01/08
11. Phannachitta, P., & Matsumoto, K. Model-based software effort estimation–a robust comparison of 14 algorithms widely used in the data science community. International journal of innovative computing information and control, 15(2), pp. 569-589 (2019)
12. Locate the Bank Routing Numbers on a Cheque, https://www.nation wide.com/lc/resources/personal-finance/articles/routing-and-account-numbers, last accessed 2019/11/09

13. Hussein Moghnieh, Scanned Numbers Recognition using k-Nearest Neighbor (k-NN), https://towardsdatascience.com/scanned-digits-recognition-using-k-nearest-neighbor-k-nn-d1a1528f0dea, last accessed 2019/11/09
14. MICR Font, https://www.micrgauge.com/about-micr.htm, last accessed 2020/01/09
15. Random Noise Using Scikit-image, https://theailearner.com/tag/skimage-util-random_noise/, last accessed 2020/01/09