Embedded Character Recognition System using Random Forest Algorithm for IC Inspection System

Chong Wei Jian, M.Z. Ibrahim, Thum Wei Seong, Ting Ei Wei and S. Khatun
Faculty of Electrical and Electronics Engineering, Universiti Malaysia Pahang, 26300 Pahang, Malaysia
zamri@ump.edu.my

Abstract—**Character recognition system based on human inspection is unpractical due to lack of accuracy and high cost. Therefore, investigating on automated character inspection system by computer is needed to improve the accuracy, reduce the cost and inspection time. In this project, a Beagle Bone Black (BBB) was used as a processing device and Logitech webcam was used for as an image acquisition device. Total of 1080 training samples will undergo the image pre-processing, character segmentation, feature extraction and training using random forest classifier. The optimal parameter values of random forest classifier are determined by computing cross-validation misclassification rate. The maximum number of splits, number of trees, and learning rate that yields the zero-misclassification rate is 1, 39 and 0.10 respectively. The process of testing random forest classifier was done using SN74LS27N chip under five different illuminations: no LED, one LED, two LED, three LED and four LED. From the experiments, it shows that the proposed system able to achieve 90.00% of accuracy within 1second to recognize characters on the SN74LS27N chip compared to 65.56% accuracy of human inspection.**

Index Terms—**Beagle Bone Black; Character Segmentation; Character Classification; Embedded System; Random Forest Algorithm.**

I. Introduction

In electronics manufacturing industries, character recognition system based on human inspection is unpractical due to lack of accuracy and costly. Besides that, different perspectives to identify a defeat of different people had increased the inspection time. In addition, the consequence of improper inspection will cause wrong installation of integrated circuit (IC) chip on a board at the same time damaged a device. The human inspection was getting replaced by automated marking inspection system.

In [1], Hua Yang et al. developed a real-time marking defect inspection system for TSSOP20 packaging chip. The system is developed by a combination of the artificial neural network with optical character recognition (OCR). The OCR techniques are including image pre-processing, marking the location, character segmentation, feature extraction and defect character classification. Bhatia et al. [2] had reviewed various techniques of the OCR system. In 2006, Ivan et al. [3] have proposed an OCR depend on a number of factors. The number of factors is consisting of feature extraction and classification algorithms. The authors in [3] look at the results of the application of a set of classifiers to datasets received through multiple fundamental of feature extraction techniques.

An OCR system suggested by Thomas et al. [4] was used for handwritten characters recognition and converting into digital text. Pal et al. [5] used Support Vector Machines (SVM) method for Bangle and Devnagari text recognition and achieved accuracy 99.18% and 98.86% respectively. In [6], Zahedi et al. proposed random forest (RF) classifier for Persian handwritten single character recognition, since RF is a robust and fast classifier which able to manage a large number of feature variables.

The research on character recognition was continued with the used of the microcontroller. Richard et al. [7] have developed digit recognition by using Beagle Board xM with MATLAB Simulink environment while Shelke et al. [8] has developed real-time character reading system for Marathi script by using Raspberry Pi with Python and OpenCV library. From the research, the Beagle Board xM has high recognition rate which is 92.38% compared to the Raspberry Pi with 92%. Furthermore, Bhatia et al. [9] had reviewed various techniques of the OCR system. In 2006, Ivan et al. [10] have proposed an OCR depend on a number of factors. The number of factors is consisting of feature extraction and classification algorithms. The authors in [10] look at the results of the application of a set of classifiers to datasets received through multiple fundamental of feature extraction techniques.

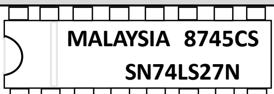
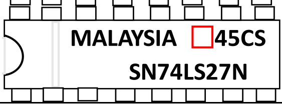
Random forest (RF) algorithm [11] was applied on character recognition [12], handwriting digit recognition [13] and even land-cover classification [14], this mostly because of RF is easily built and is one of the best accurate learning algorithm available, but surprisingly RF is not yet applied on IC marking inspection from previous research. In this paper, RF algorithm was applied to inspect IC marking and compare performance with SVM.

II. Methodology

In this work, an embedded character recognition system is constructed using Logitech webcam C525 which are built on a Beagle Bone board. The webcam is being assigned as the image acquisition tool and the Beagle Bone board acts as an interface between webcam and computer for image processing of SN74LS27N IC chip.

There are three types of IC chips which with correct character, misprint character and missing character as shown in Table 1. The correct character is represented the chip is marking correctly. The misprint character is represented the chip is marking with 623 numbers instead of 8745 numbers. The missing character is represented the chip is lack of 87 number on it.

Table 1
Type of Character on SN74LS27N IC Chip

Type of character on SN74LS27N IC chip	Top view of SN74LS27N IC chip
Correct character	
Misprint character	
Missing character	

The flow chart of the system is illustrated in Figure 1. A list of capital letter and number were created for training process using Microsoft Word 2016 in Arial as font type and 36 as font size. There are total 52 characters for both uppercase and lowercase alphabet, while there are total 10 characters for numbers as shown in Figure 2.

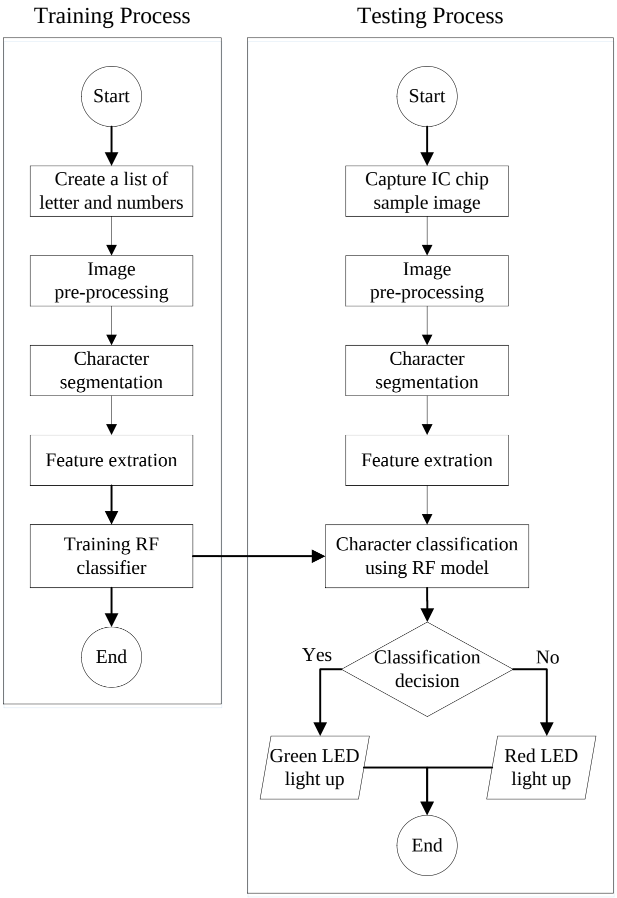


Figure 1: Flowchart of embedded character recognition system

For the character segmentation process, each capital letter and number in the binary image are bounded and segmented into individual character. The framework of character segmentation contains three basic steps which are label connected components, measure properties of the image region and character extraction.

First, the connected components in the 2-D binary image will be labeled by using 'bwlabel' function in MATLAB. Then, the labeled region will be bounded by a bounding box using 'regionprops' function as shown in Figure 3.

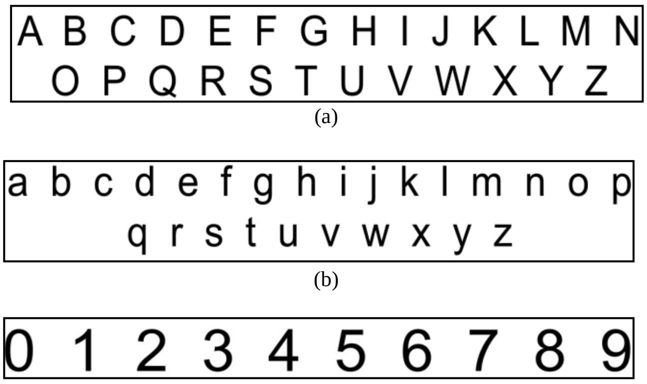


Figure 2: Characters of alphabet and numbers before image pre-processing (a) uppercase letter from A to Z, (b) lowercase letter from a to z, (c) number from 0 to 9

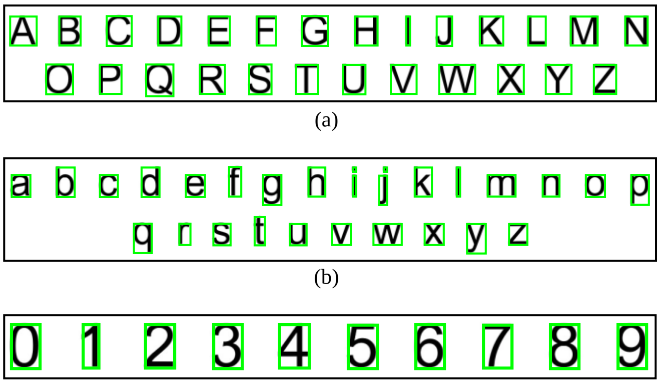


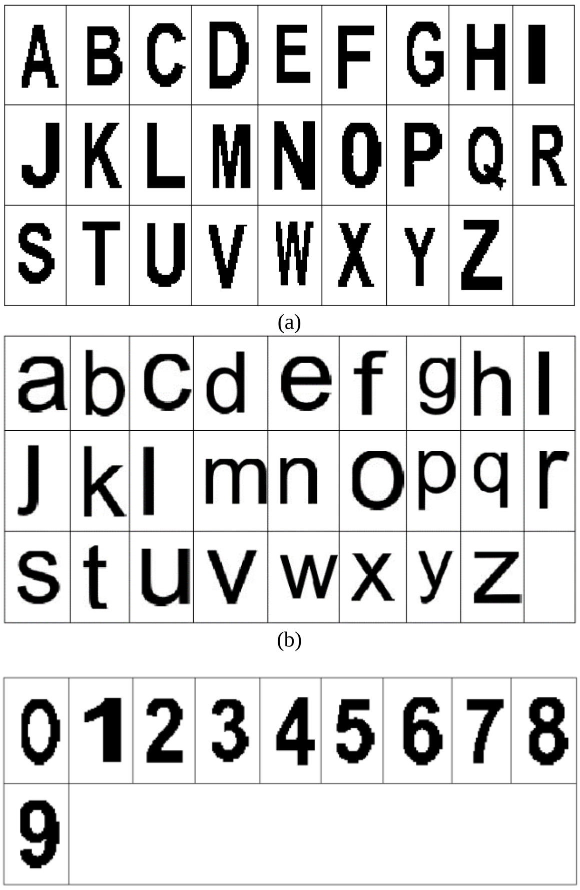
Figure 3: Bounding box for image (a) uppercase letter from A to Z, (b) lowercase letter from a to z, (c) number from 0 to 9

The non-zero element in the bounding box will be extracted out as a binary image with 11 row and 13 columns. Then the binary image is undergoing complement process to reversed black (zeros) and white (white) to obtain final training images.

Due to the difference output resolution and pixel values of each character, all individual characters were complemented and resized to 24 x 42 pixels and save as picture with the format of Joint Photographic Experts Group (JPEG). For the character recognition, a histogram of oriented gradient (HOG) [15] and local binary pattern (LBP) [16] techniques were used to extract features from the binary segmented image while Random Forest [11] and support vector machine (SVM) [17] were used as a classifier.

The IC chip image acquired from the Logitech webcam C525 will be undergone image pre-processing. During image pre-processing process, the region of the chip's image that does not consist of any character is cropped out. The cropping process is used to get a region of interest (ROI) which consists of characters only. After that, the ROI region image is undergoing histogram equalization for image enhancement.

Then, the ROI image was converted to the grayscale image from the color image, then further converted into a binary image for character segmentation. For the character segmentation process, each capital letter and number in the binary image are bounded and segmented into individual character.



(a)

(b)

(c)

Figure 4: Image complement and resize for (a) uppercase letter from A to Z, (b) lowercase from a to z, (c) number from 0 to 9

Next, characters in the ROI image are classified by the random forest classifier to determine the class membership for the characters. During the classification process, the green and red light emitting diodes (LED) will turn on and off to indicate the classification result. Green LED indicates that character is being recognized completely, then red LED indicates classification result is not similar to the characters on the SN74LS27N IC chip. The embedded character recognition system is constructed and setup is shown in Figure 5.

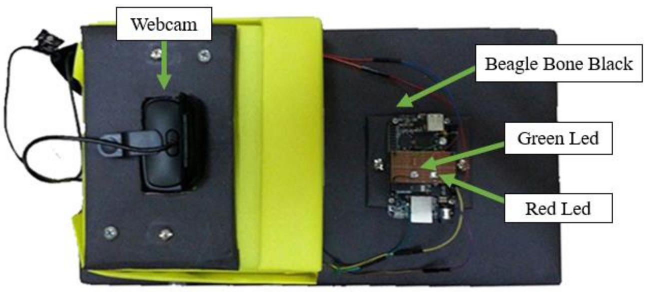


Figure 5: Top view of embedded character recognition system

III. Result and Discussion

An experiment was carried out for analyzing the accuracy of human eyes character recognition. There are total 30 volunteers involved in character inspection on SN74LS27N IC chips and accuracy achieved 66% only.

Table 2
Type of Feature Extraction and Classifier with Percent of Correct Classification

Type of feature extraction	Type of classifier	Percent of correct classification (%)
LBP	SVM	74.44
LBP	Random Forest	81.11
HOG	SVM	85.56
HOG	Random Forest	90.00

Table 2 compared the performance for character recognition system using LBP and HOG as feature extraction and, Random Forest and SVM as a classifier. It can be shown that the HOG with Random Forest is suitable for this character recognition because it obtains the highest percent of correct character classification which is 90% compared to other types of feature extraction and classifier. Remaining of the section will use HOG and Random Forest as feature extraction and classifier.

The optimal distance between the Logitech Webcam with SN74LS27N IC chip was investigated. By obtaining the optimal distance, webcam able to capture the IC chip clearly without any blurring. According to the result shown in the Figure 6, the optimal distance between the webcam and IC chip is 1.0cm.

The relationship between the light intensity with the efficiency of character recognition classification was also investigated. Brightness is significant for the webcam to capture a clear image, the low light intensity will cause the images captured are dark and then the characters are hard to recognized by the system. The calculation percent of correct character classification is shown in Figure 7.

From the Figure 7, it can be shown that three LEDs are the best light source for this embedded character recognition system which achieved 90% of correct classification. Three LEDs provide sufficient light intensity for the webcam to capture a clear image, but four LEDs provide too much light intensity and the image was too bright and not suitable for character recognition.

To validate the consistency and reliability of RF in character classification, a few experiments were conducted to pick the best parameter values for Random Forest classifier. Comparison between a different number of tree, the maximum number of tree split and learning rate were implemented.

Based on the Table 3, it can be deduced that the best parameter values obtained during the 39 number of trees, 1 maximum number of tree splits and 0.10 learning rate. The best parameters values are determined by observing the number of trees, a maximum number of tree and learning rate which able to achieve zero cross-validation misclassification rate. A higher number of trees should be selected to ensure all observations will be predicted once or even more. In order to prevent overfitting happen, low learning rate should be picked. Hence, the 39 number of trees is achieved zero cross-validation misclassification rate by 1 maximum number of tree splits and 0.1 learning rate.

Meanwhile, the 19 number of trees is achieved zero cross-validation misclassification rate by 3 maximum number of tree splits and 0.25 learning rate. By comparing the 39 number of trees with the 19 number of tree, the 19 number of trees has two highest values for the number of trees and a maximum number of tree splits. Therefore, the 19 number of trees is not the best or optimal parameter values for the random forest classifier.

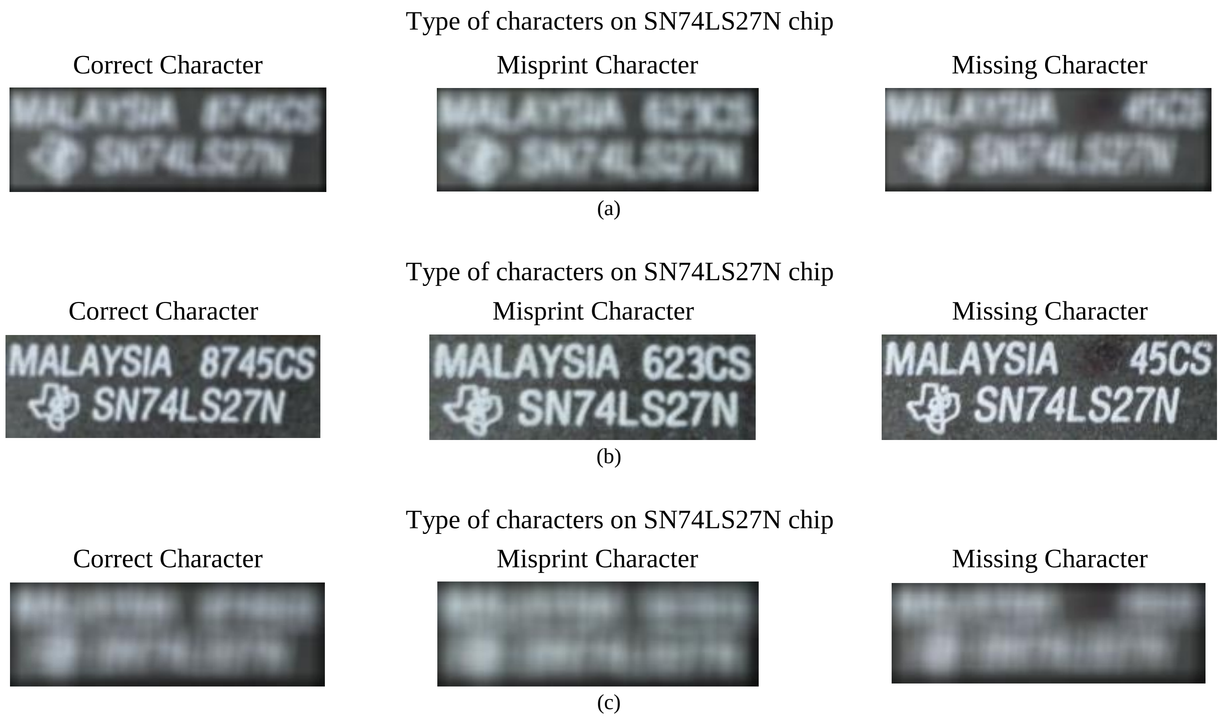


Figure 6: Distance between the Logitech webcam and SN74LS27N IC chip (a) 0.5cm, (b) 1.0cm, (c) 1.5cm

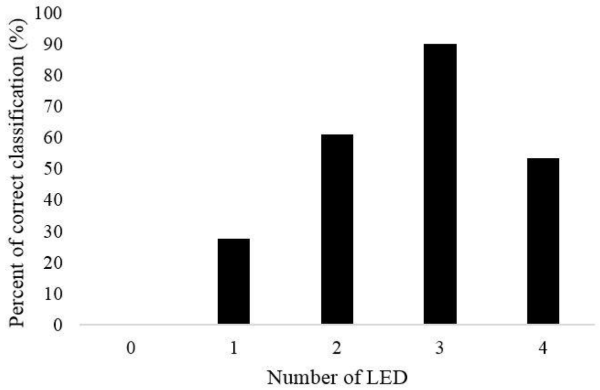


Figure 7: Performance of correct classification with different number of LEDs

Table 3
Comparison of Cross-Validation Misclassification Rate with Number of Tree, Learning Rate And Maximum Number of Tree Split

Number of trees	Maximum number of tree splits	Learning rate	Cross-validation misclassification rate
39	1	0.10	0
19	3	0.25	0
1	27	1.00	0

On the other hand, the 1 number of the tree is achieved zero cross-validation misclassification rate by 27 maximum number of tree splits and 1.00 learning rate. By comparing the 39 number of the tree with the 1 number of tree, the 1 number of the tree has two highest values for the number of trees and a maximum number of tree splits. Therefore, the 1 number of the tree is not the best or optimal parameter values for the random forest classifier.

As a conclusion, the 39 number of trees with the 1 maximum number of tree split and 0.10 learning rate has the lowest parameter values for achieving zero cross-validation misclassification rate.

IV. Conclusion

A random forest algorithm has been implemented successfully in an automated character recognition system. The system has been evaluated in terms of accuracy of classification based on the type of feature extraction technique and classifier. In addition, the optimal webcam distance from IC and light intensity were also investigated in this work. In a nutshell, propose automatic recognition system able to detect correct characters on the SN74LS27N chip is achieved at 90% compared to a human at 66%.

Acknowledgment

This research is supported by Universiti Malaysia Pahang Internal Grant of RDU150363. The authors would also like to thank the Faculty of Electrical & Electronics Engineering Universiti Malaysia Pahang for providing the facilities to conduct this research and for financial support throughout the process.

References

[1] Yang, Hua, Buyang Zhang, and Yang Hu, "A real-time marking defect inspection method for IC chips." In *Seventh International Conference on Graphic and Image Processing. International Society for Optics and Photonics*. 2015.

[2] Bhatia, Er Neetu, "Optical character recognition techniques: a review." *International Journal of Advanced Research in Computer Science and Software Engineering* 4, no. 5, 2014.

[3] Dervisevic, Ivan, "Machine Learning Methods for Optical Character Recognition." *University of Novi Sad*, 2006.

[4] Deselaers, Thomas, Tobias Gass, Georg Heigold, and Hermann Ney, "Latent log-linear models for handwritten digit classification." *IEEE transactions on pattern analysis and machine intelligence* 34, no. 6, pp. 1105-1117, 2012.

[5] Pal, Umapada, Partha Pratim Roy, Nilamadhaba Tripathy, and Josep Lladós, "Multi-oriented Bangla and Devnagari text recognition." *Pattern Recognition* 43, no. 12, pp. 4124-4136, 2010.

[6] Zahedi, M., and S. Eslami, "Improvement of random forest classifier through localization of persian handwritten OCR." *ACEEE Int. J. Inf. Technol.* 1, no. 2, pp. 31-36, 2012.

[7] Borbély, Richard, Michal Blaho, Leo Mrafko, and Tatiana Mudráková, "Digits Recognition With Beagleboard Platform."

[8] S. D. Shelke and S. D. Apte, "Real Time Character Reading System for Marathi Script using Raspberry Pi," *3rd Int. Conf. Electr. Electron. Eng. Trends, Commun. Optim. Sci*, pp. 868–872, 2016.

[9] Bhatia, Er Neetu, "Optical character recognition techniques: a review." *International Journal of Advanced Research in Computer Science and Software Engineering* 4, no. 5, pp. 1219–1223, 2014.

[10] I. Dervisevic, "Machine Learning Methods for Optical Character Recognition," 2006.

[11] Garzón, Marta Benito, Rut Sánchez de Dios, and Helios Sainz Ollero, "Effects of climate change on the distribution of Iberian tree species." *Applied Vegetation Science* 11, no. 2, pp. 169-178, 2008.

[12] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets : A Learned Multi-Scale Representation for Scene Text Recognition," *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit*, pp. 4042–4049, 2014.

[13] Bernard, Simon, Sébastien Adam, and Laurent Heutte, "Using random forests for handwritten digit recognition." In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, pp. 1043-1047. IEEE, 2007.

[14] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 93–104, 2012.

[15] Dalal, Navneet, and Bill Triggs, "Histograms of oriented gradients for human detection." In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886-893. IEEE, 2005.

[16] Ojala, Timo, Matti Pietikäinen, and David Harwood, "A comparative study of texture measures with classification based on featured distributions." *Pattern recognition* 29, no. 1, pp. 51-59, 1996.

[17] Suykens, Johan AK, and Joos Vandewalle, "Least squares support vector machine classifiers." *Neural processing letters* 9, no. 3, pp. 293-300, 1999.