Yanbu Journal of Engineering and Science

Volume 17 | Issue 1 Article 7

2019

# OCR FOR ENGLISH CHARACTERS BASED ON POLAR HISTOGRAM FEATURE EXTRACTION AND EUCLIDEAN DISTANCE

Saleh Ali Alshehri
*Jubail University College, Royal Commission at Jubail, Jubail Industrial City, Saudi Arabia*,
shehri@ucj.edu.sa

Follow this and additional works at: https://yjes.researchcommons.org/yjes
The journal in which this article appears is hosted on Digital Commons, an Elsevier platform.

Recommended Citation
Alshehri, Saleh Ali (2019), OCR FOR ENGLISH CHARACTERS BASED ON POLAR HISTOGRAM FEATURE EXTRACTION AND EUCLIDEAN DISTANCE, *Yanbu Journal of Engineering and Science*: Vol. 17: Iss. 1, 65-71.
DOI: https://doi.org/10.53370/001c.23733
Available at: https://yjes.researchcommons.org/yjes/vol17/iss1/7

This Research Article is brought to you for free and open access by Yanbu Journal of Engineering and Science. It has been accepted for inclusion in Yanbu Journal of Engineering and Science by an authorized editor of Yanbu Journal of Engineering and Science.

Yanbu Journal
of
Engineering
and Science
ISSN: 1658-5321

Vol. 17, December 2019 (1441H)
www.yjes.org.sa

# OCR FOR ENGLISH CHARACTERS BASED ON POLAR HISTOGRAM FEATURE EXTRACTION AND EUCLIDEAN DISTANCE

**Saleh Ali Alshehri**

*Department of Computer Science and Engineering, Jubail University College, Royal Commission at Jubail, Jubail Industrial City, Saudi Arabia*
*E-mail: shehri@ucj.edu.sa*

**ABSTRACT**

Optical character recognition is the process of converting characters from image format to text format. The process includes four main stages namely: pre-processing, feature extraction, character recognition, and post-processing. The success is mainly based on the feature extraction method and the character recognition algorithm. When optical character recognition systems are developed for mobile devices, two main constraints must be addressed. They are the system's size and speed. In this research, both were considered. A small feature dataset was created using polar histogram of x and y projections of the character image. In addition, a simple character recognition algorithm based on Euclidean distance was adapted. An accuracy of over 99.9% was achieved in near microsecond scale recognizing execution time on the development system.

**Keywords***: OCR, Polar Histogram, Euclidian Distance*

## 1. INTRODUCTION

Optical character recognition (OCR) is the process of character recognition where a character in image pixels format is converted to character text format such as ASCII or Unicode [1], [2], [3]. The process starts by preprocessing the text image that contains characters to prepare it in a form that can be used to extract unique features. This is the preprocessing stage. It may include noise reduction, skewness alignment, and binarization [4], [5]. The second stage is feature extraction. This is the most important stage. There are a huge number of features that can be used in OCR systems [6]. However, for each recognition algorithm there are certain features that are fit for that algorithm. Researchers keep discovering new features from time to time. Once the features are obtained, the recognition algorithm is chosen and implemented. In many situations, post-processing such as output formatting may be needed. These are the four main OCR system stages [4]. Each of which may include a few sub-stages. There are three main goals of any OCR system: speed, accuracy, and storage capacity [1]. Any OCR system focuses on one or two of these goals. Satisfying all three goals might not be possible. Because of the storage and processing speed limitation, mobile devices have some constraints on the OCR applications. The algorithm and dataset must have a lightweight footprint. The algorithm must not consume too much processing power [7], [8], [9]. The OCR system may be language dependent or language independent. There are many surveys for general OCR and for language specific OCR [10], [11], [12].



Received: 3 October 2019, Accepted: 4 December 2019

There have been considerable amount of researches on English OCR systems. English character OCR system can be considered a primitive one among other language script OCRs [13], [14]. For example, simple artificial neural network with neighbor of scanned images was used to build successful OCR with accuracy higher than 99% [15]. In that study, discrete cosine transform (DCT) was used for feature extraction with various classifiers [13]. The classifiers were fuzzy multilayer perceptron (FMP), rough fuzzy multilayer perceptron (RFMP), fuzzy support vector machine (FSVM) and fuzzy rough support vector machine (FRSVM) [13]. Artificial neural network (ANN) which produces the character Unicode were developed [15]. Roman and English characters OCR are robust in most of the widely used fonts and styles while it is not for many other fonts and styles [14]. Fast recognition based on distance profile feature with respect to left, right and diagonal direction of the character image was presented with accuracy over 80% [14]. Studying historical corpora is a field where OCR plays an important rule. There have been some studies to correct OCR error based on advance neural machine translation (NMT) [16], [17], [18]. An accuracy of 58% was achieved [18]. Among the used feature extraction is image histogram and x and y projections [11], [12], [13], [14], [15], [16], [17], [18], [19]. While both methods use large number of features which slowdown the recognition, we propose a method that reduces the feature to range between 4 and 24. Up to our knowledge, this feature reduction method was not presented in the literature so far. In this research, the focus is on the feature extraction and character recognition algorithm. Other OCR stages are not covered in this research. The techniques of these OCR stages can be adapted from many sources [1], [20], [21], [22], [23].

## 2. OCR SYSTEM STRUCTURE

The proposed recognition system consists of two stages. First, feature extraction was done leading to a small feature's dataset. Second, simple Euclidean distance was adapted. The small dataset is suitable for a mobile device. In addition, Euclidean distance calculation is simple and does not consume much processing power. English character sets with 1016 different fonts for each character was used in this research [24].

### 2.1 CHARACTER FEATURE EXTRACTION

The shape of any character is a kind of unique feature. Also, the number of pixels in a character image may differ from other characters. Building a template that can be used for character recognition produces good results. However, the template will have a large number of pixels. Thus, the recognition algorithm is a time consuming process. Another feature extraction is processed by calculating the x and y projections [11], [19]. However, the data point could be very large. For example, if the image size is 128 x 128 pixels, the sizes of x and y projections are 128 and 128 respectively. The total feature points will be 256 features. This is too large a number to be used in systems where processing power is limited, such as mobile devices. Character image histogram is widely used in character recognition as well as the image projections on the two dimensions [5]. Image projections constitute large feature size. For example, an image sized 128 x 128 produces 128 bins in x-projection and 128 in y-projection, which adds up to 256 data points. To capture adequate image histogram, a minimum of 128 data points are needed. In contrast, polar histogram can be used. The number of data points can be specified. In this research, the number of data points was varied between (4,8,12,16,20,24). This is done on the original image x projection and y projection. Figure 1 shows samples of the dataset characters. Figure 2 shows one font of

character “t” and its polar histogram with various bin sizes.



Fig. 1. Some samples of the characters dataset



Fig. 2. One font of letter ‘t’ in the dataset

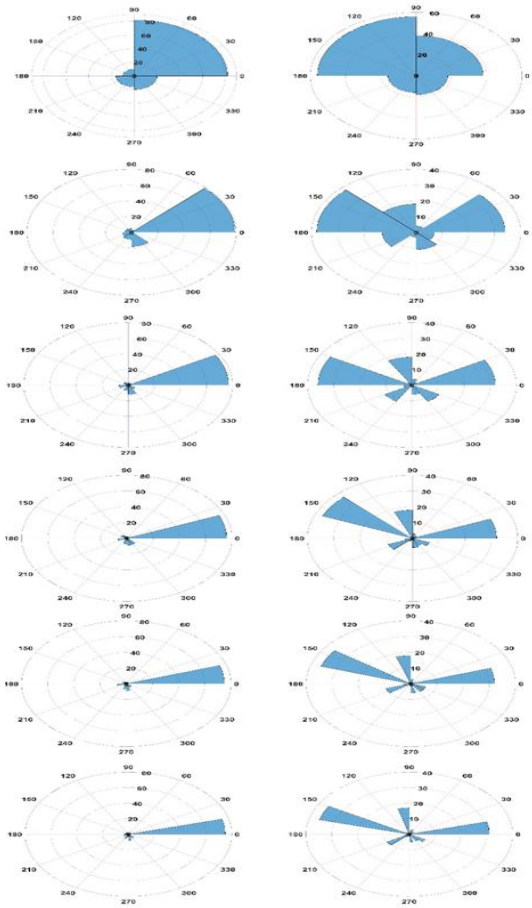

Fig. 3. Upper charts are for x projections while lower charts are for y-projections with the order from left to right (4,8,12,16,20,24) bin size

It is clear that the various histogram sizes show different feature values. This is true even for x and y projections with the same bin numbers. For example, Figure 3 has the main loop of x-projection at the angel between 0 and 30, while the main loop in y-projection is between about 135 to 180.

### 2.2 DISTANCE MEASURE FOR CHARACTER RECOGNITION

The recognition algorithm can be applied once the feature dataset is created. There are a huge number of algorithms, each of which has its advantages and disadvantages. Some are fast recognition algorithms while others have high recognition accuracy. This research focuses on storage capacity size and execution speed. The aim is to minimize both. The goal of this research is to use as minimum an amount as possible of data to be used as character features. This guarantees less computation time and less needed storage capacity. Still, it is preferable to capture the character shape in a small size format. Obtaining the image histogram in polar format produces good features in addition to small feature size. It was decided to calculate the polar histogram of x and y projection and plug the result into the Euclidean distance algorithm. Polar histogram in this case acts as a data transform tool. The size of polar histogram can be controlled by specifying the angle step size, which is a property of polar histogram. So, it can be generated for step size between 1, 180. The number of bins in the histogram specifies the step size. Polar histogram assumes the values are in degrees. In this research, the data is represented in x and y projections. Then these two vectors are inserted into a polar histogram function to produce the transferred data. The numbers of bins were chosen to be varied among (4, 8, 12, 16, 20, 24). All these bin values were investigated to obtain the best speed, accuracy, and storage capacity. Table 1 shows the percentage error and the execution time for each bin number. The calculation was done for 20 fonts, 200 fonts,

500 fonts and all fonts of each character in the dataset. For its simple arithmetic calculation, Euclidian distance is used in this research.Euclidian distance between a given vector (v) and all vectors in the dataset (vi) is obtained using Eq (1) and Eq (2):

Where n is the size of the dataset.

$$FD_i = \sqrt{\sum_{i=1}^{n}(v - v_i)^2} \quad (1)$$

$$RC = \text{argmin}(FD_i) \quad (2)$$

TABLE 1: PERCENTAGE ERROR AND THE EXECUTION TIME FOR VARIOUS FONTS USING VARIOUS NO. OF FEATURES.

| No. of Features | 20 | | 200 | | 500 | | All | |
|---|---|---|---|---|---|---|---|---|
| | Error | Execution time | Error | Execution time | Error | Execution time | Error | Execution time |
| 4xp | 12.661 | $8.207 \times 10^{-5}$ | 40.920 | $1.6511 \times 10^{-4}$ | 59.213 | $3.110 \times 10^{-4}$ | 70.347 | $1.150 \times 10^{-3}$ |
| 8xp | 0.726 | $9.044 \times 10^{-5}$ | 0.323 | $2.0927 \times 10^{-4}$ | 0.736 | $9.884 \times 10^{-4}$ | 1.030 | $2.227 \times 10^{-3}$ |
| 12xp | 0.726 | $9.354 \times 10^{-5}$ | 0.291 | $5.960 \times 10^{-4}$ | 0.652 | $14.775 \times 10^{-4}$ | 0.774 | $3.494 \times 10^{-3}$ |
| 16xp | 0.323 | $9.807 \times 10^{-5}$ | 0.258 | $7.494 \times 10^{-4}$ | 0.604 | $19.78 \times 10^{-4}$ | 0.722 | $4.618 \times 10^{-3}$ |
| 20xp | 0.323 | $10.729 \times 10^{-5}$ | 0.258 | $9.474 \times 10^{-4}$ | 0.597 | $27.907 \times 10^{-4}$ | 0.709 | $6.512 \times 10^{-3}$ |
| 24xp | 0.323 | $11.311 \times 10^{-5}$ | 0.242 | $11.381 \times 10^{-4}$ | 0.578 | $34.991 \times 10^{-4}$ | 0.677 | $9.409 \times 10^{-3}$ |
| 4yp | 3.548 | $8.816 \times 10^{-5}$ | 19.467 | $1.721 \times 10^{-4}$ | 37.904 | $3.202 \times 10^{-4}$ | 53.038 | $1.236 \times 10^{-3}$ |
| 8yp | 0.403 | $9.329 \times 10^{-5}$ | 0.872 | $2.108 \times 10^{-4}$ | 1.162 | $10.143 \times 10^{-4}$ | 1.437 | $2.411 \times 10^{-3}$ |
| 12yp | 0.565 | $9.591 \times 10^{-5}$ | 0.565 | $6.060 \times 10^{-4}$ | 1.020 | $14.922 \times 10^{-4}$ | 1.284 | $3.278 \times 10^{-3}$ |
| 16yp | 0.403 | $10.2 \times 10^{-5}$ | 0.508 | $7.568 \times 10^{-4}$ | 0.882 | $20.336 \times 10^{-4}$ | 1.073 | $4.445 \times 10^{-3}$ |
| 20yp | 0.403 | $10.574 \times 10^{-5}$ | 0.282 | $09.858 \times 10^{-4}$ | 0.817 | $29.607 \times 10^{-4}$ | 1.013 | $6.567 \times 10^{-3}$ |
| 24yp | 0.403 | $11.185 \times 10^{-5}$ | 0.282 | $11.702 \times 10^{-4}$ | 0.791 | $35.602 \times 10^{-4}$ | 0.973 | $8.698 \times 10^{-3}$ |
| 4xp,4yp | 0.565 | $9.404 \times 10^{-5}$ | 0.694 | $2.097 \times 10^{-4}$ | 0.849 | $10.137 \times 10^{-4}$ | 1.082 | $2.648 \times 10^{-3}$ |
| 8xp,8yp | 0.323 | $10.898 \times 10^{-5}$ | 0.258 | $8.159 \times 10^{-4}$ | 0.601 | $22.653 \times 10^{-4}$ | 0.650 | $4.898 \times 10^{-3}$ |
| 12xp,12yp | 0.484 | $11.568 \times 10^{-5}$ | 0.258 | $12.438 \times 10^{-4}$ | 0.584 | $36.889 \times 10^{-4}$ | 0.634 | $8.676 \times 10^{-3}$ |
| 16xp,16yp | 0.323 | $12.567 \times 10^{-5}$ | 0.242 | $17.874 \times 10^{-4}$ | 0.572 | $50.490 \times 10^{-4}$ | 0.629 | $11.348 \times 10^{-3}$ |
| 20xp,20yp | 0.323 | $13.49 \times 10^{-5}$ | 0.242 | $22.184 \times 10^{-4}$ | 0.568 | $66.887 \times 10^{-4}$ | 0.626 | $14.549 \times 10^{-3}$ |
| 24xp, 24yp | 0.323 | $16.021 \times 10^{-5}$ | 0.242 | $27.578 \times 10^{-4}$ | 0.565 | $81.393 \times 10^{-4}$ | 0.622 | $17.529 \times 10^{-3}$ |

## 3. RESULTS AND SUMMARY

Applying Euclidean distance on the collected dataset shows a fast and accurate OCR algorithm. From Table 1, one can infer that the polar histogram of y-project always gives better results than the polar histogram of x-projection. Figure 4 shows the percentage error when selecting 20 characters' fonts out of 1016 fonts. It shows that the difference in performance accuracy started to be very comparable for 8 bins and higher. So, it would be more economical to use 8 bins. However, the execution speed is slower with a higher number of bins. The execution time for x-projection is very close to the execution time for y-project as shown in Figure 5. However, it is much higher when both projections were used. Choosing 8 to 12 bins for either x-projection or y-projection will be the optimum choice in term of accuracy and speed. To evaluate the chosen number of fonts, the performance error of 20, 200, 500 and all fonts for y-projection only was calculated. Figure 6 shows the performance error comparison result while Figure 7 shows the execution speed. The execution time is almost constant for 20 fonts choice for all bins choices. Figure 7 shows that it is a little higher for 200 fonts choice with little increasing behavior for 4 to 24 bins range. It developed a smart phone application for car number plate recognition [25]. The system reached an accuracy of 92%. Another study was developed, where mobile phone OCR application with 92% accuracy was built [7]. It takes about 1.25 seconds to recognize 200 characters in a scanned business card. The proposed method shows an accuracy of above 99% with less than 0.02 seconds to recognize the same number of characters.

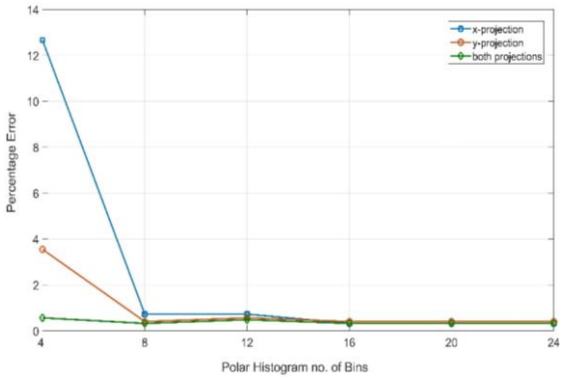


**Fig.** 4. Error when selecting 20 characters' fonts out of 1016 fonts

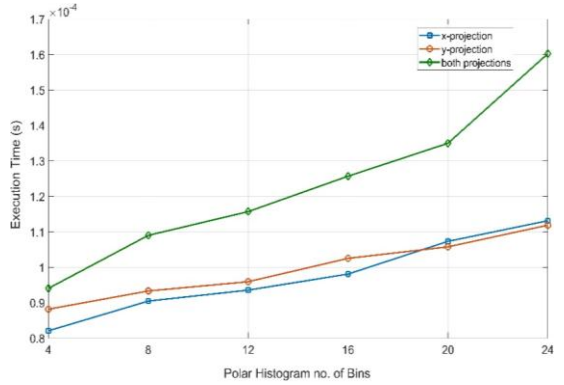


**Fig.** 5. Execution time for x and y projections

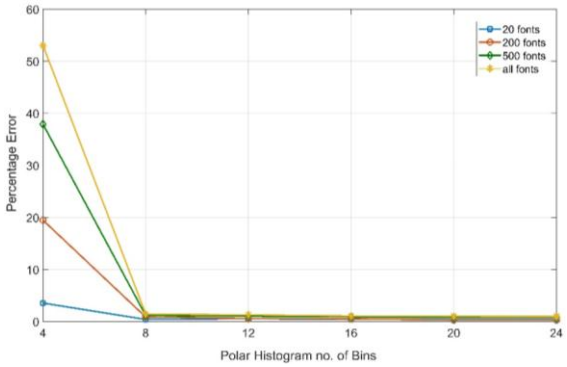


Fig. 6. Error for various font choices for all bins

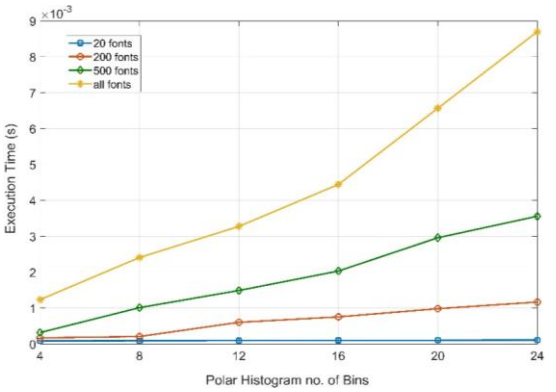


Fig. 7. Execution time for various font choices for all bins

## 4. CONCLUSION

A dataset with 1016 fonts for each character was used for feature extraction that was based on the polar histogram of the x and y projections of character image. The extracted features size of the generated dataset was small since few data points were generated from the polar histogram. Euclidean distance algorithm was used as the recognition tool. The results show high accuracy was achieved. In addition, the execution time was acceptable. The choice of the feature size is left to the user. For mobile devices, both speed and dataset size are important. It needs further investigation. This is the subject of research currently undertaken by the author to develop more compact dataset size and migrate it to a mobile phone for testing purposes.

## REFERENCES


[1] Borovikov, E., "A survey of modern optical character recognition techniques," AMS, pp. 1-4. 2004.

[2] Bhatia, E., "Optical Character Recognition Technique: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, 4, pp. 1219-1223. 2014.

[3] Dedgaonkar, S., Chandavale, A., and Sapkal, A., "Survey of Methods for Character Recognition," Internaltional Journal of Engineering and Innovative Technology, 1, pp. 180-189, 2012.

[4] Aharranw, N., El Moutaouakil, K. and Satori, K., "A comparison of supervised classification methods for a statistical set of features: Application: Amazigh OCR" in Intelligent Systems and Computer Vision (ISCV) 2015, IEEE, pp. 1-8, 2015.

[5] Singh, A., Baccuwar, K., and Bhasin, A., "A Survey of OCR Applications," International Journal of MAchine Learning and Computing, 2, pp. 314-318, 2012.

[6] Singh, A., Bacchuwar, K. and Bhasin, A., "A Survey of OCR Applications", International Journal of Machine Learning and Computing, Vol. 2, No. 3, pp. 314- 318, 2012.

[7] Mollah, A., Majumder, N., Basu, S., and Nasipuri, M., "Design of an Optical Character Recognition System for Camera-based Handheld Devices," Internationl Journal of Computer Science Issues, 8, pp. 283-289, 2011.

[8] Ch, S., Mahna, S., and Kashyap, N., "Optical Character Recognition on Handheld Devices," International Journal of Computer Applications, 115, pp. 10-13, 2015.

[9] Kwon, S., An, H., and Choi, Y., "Character Recognition System Based On Android Smart Phone," International Journal of Modern Engineering Research, 2, pp. 4091-4093, 2012.

[10] AlShoshan, A., "Arabic OCR Based on Image Invariants", Proceedings of the Geometric Modeling and Imaging-New Trends (GMAI'06), 2006.

[11] Parvez, M. and Mahmoud, S., "Offline Arabic Handwritten Text Recognition: A Survey", ACM Computing Surveys, 45(2), pp. 1-35, 2013.

[12] Abdul Sattar, S. and Shah, S., "Character recognition of Arabic script languages", Proceedings of the International Conference on Computer and Information Technology (ICCIT'12), 2012.

[13] Chaudhuri, A., Mandaviya, K., Badelia, P. and Ghosh, s., "Optical Character Recognition Systems for English Language", Springer, 2017.

[14] Bharath, V. and Rani, N., "A font style classification system for English OCR" International Conference on Intelligent Computing and Control Systems (ICICCS 2017), Madurai, India, 15-16, June 2017.

[15] Mehta, H., Singla, S. and Mahajan, A., "Optical character recognition (OCR) system for Roman script and English language using Artificial Neural Network (ANN)", 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS), Bangalore, India, May 6-7, 2016.

[16] Reul, C., Springmann, U., Wick C. and Puppe, F., "Improving OCR Accuracy on Early Printed Books by Combining Pretraining, Voting, and Active Learning", Journal for Language Technology and Computational Linguistics, v2, pp. 1-22, 2018.

[17] Reul, C., Springmann, U., Wick C. and Puppe, F., "Improving OCR Accuracy on Early Printed Books by utilizing Cross Fold Training and Voting", Journal for Language Technology and Computational Linguistics, 2017.

[18] Hämäläinen, M. and Hengchen, S., "From the Paft to the Fiiture: a Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction", Proceedings of the International Conference on Recent Advances in Natural

Language Processing (RANLP 2019), Varna, Bulgaria, September 2019.

[19] Casey, R., and Lecolinet, E., "A survey of methods and strategies in character segmentation", IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, no. 7, pp. 690-706, 1996.

[20] Dedgaonkar, S., Chandavale, A., and Sapkal, A., "Survey of Methods for Character Recognition," Internaltional Journal of Engineering and Innovative Technology, 1, pp. 180-189, 2012.

[21] Burie, J. et al., "ICDAR2015 Competition on Smartphone Document Capture and OCR (Smartdoc)", International Conference on Document Analysis and Recognition (lCDAR), 2015.

[22] Alshehri, S., "OCR for Mobile Phone App Based on Partial Projection of Letter Pixels," International Journal of Applied Engineering Research, pp. 9180-9184, 2016.

[23] Ravina, M., Indalkar, S. and Divekar, N., "Optical character recognition." International Journal of Recent Technology and Engineering (IJRTE) 2.1, 2013.

[24] de Campos, T., Babu, B. and Varma, M., "Character recognition in natural images.", In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, February 2009.

[25] Gunawan, T., Mutholib, A. and Kartiwi, M., "Performance Evaluation of Automatic Number Plate Recognition on Android Smartphone Platform", International Journal of Electrical and Computer Engineering , pp1973-1982, 2017.