IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 4, April 2015.
www.ijiset.com
ISSN 2348 – 7968

An Approach to GUI Identification for Printed Gurumukhi and English Text

Inderpreet Kaur[1] and Saurabh Mahajan[2]

[1] P.G. Scholar, Department of ECE, Chandigarh University,
Gharuan, Punjab, India.

[2] Assistant Professor, Department of ECE, Beant College of Engineering & Technology,
Gurdaspur, Punjab, India.

Abstract

Optical Character Recognition system is used to recognize printed and handwritten alphanumeric text from input image. A numerous of methods have been published based on optical character recognition. In proposed work expansion of optical character recognition to recognize multi-scripts is done which in infancy. Such type of expansion is crucial in India where each state has diverse language. The planned work is done to recognize Gurumukhi and English text only. Arial font is consider for English language and Gurbanikalmi font is consider for Gurumukhi language for the training of the proposed system and the recognition of image text evaluates with respect to following various stages; pre-processing, segmentation, feature extraction and classification are used to process the assumed image. The text of assumed image is segmented in lines, words and characters using proposed segmentation scheme after its processing through pre-processing stage. Then histogram projection profiles and number of holes features are calculated to recognize text. A GUI has been proposed to process the assume image and to calculate the recognition accuracy of system for different set of test samples. The end result of proposed system offers high accuracy and also provided information about correct and wrong recognized number of samples.

Keywords: *OCR, English, Gurumukhi, Segmentation, GUI*

1. Introduction

Optical Character Recognition is a software tool to analyze text image through an optical mechanism. The main objective of OCR is to extract important text data and information from input text image to retrieve that text information from stored database. To design methods to achieve these kinds of tasks is fast growing area for researchers. The researchers have published several methods based on optical character recognition in different areas due to its interesting nature and practical importance of its applications. This usage of optical character recognition is limited to recognize single script such as English and Arabic [1] – [11]. Because different languages has different character features which are extracted to recognize character that depends on structural properties, style and nature of writing, for instance features of Gurumukhi language are not useful to recognize English language [2]. Therefore, it is mandatory to select those features which are common to desired languages. These features would help to design an appropriate optical character recognition system to identify selected languages.

A literature survey has been presented for various published OCR methods. Weinman *et al* have discussed Scene text recognition using similarity and a lexicon with sparse belief propagation [4]. Saidane and Garcia have purposed an automatic scene text recognition using a convolution neural network [14]. Feng Pan *et al* have offered text localization in natural scene images based on conditional random field [13]. Lehal and Singh [3] have presented segmentation schemes for Gurumukhi text. Jindal *et al* developed a system to study language by touching characters in degraded Gurumukhi text [6]. Ranzato *et al* employ on sparse feature learning for deep belief networks [5]. Faaborg has described a technique using neural networks to create an adaptive character recognition system [1]. Mahasukkh *et al* presented hand-printed English character recognition based on fuzzy theory [7]. Kasaei *et al* have described new morphology based method for robust Iranian car plate detection and recognition [10]. Xin and Guoliang provide a graphical model for joint segmentation and recognition of license plate characters [12]. Here, an attempt has been made to design a system to recognize Gurumukhi and English text. The different important features are used for recognition purpose and such features have been chosen based on main characteristics of languages.

The detail description of language's related important points is given in Section II. Section III involves detail explanation of projected method architecture and functions of each stage. The results are discussed in Section IV which shows efficiency of projected method and the conclusion of work is presented in Section V.

2. Properties of Languages

The properties of desired languages play an important role in the design of an appropriate optical character recognition system to recognize the characters with more accuracy. In this section a brief introduction of relevant properties of both English and Gurumukhi languages is given.

- English language is widely learned as 2nd language along with mother language.
- The English language consist 26 characters (both uppercase and lowercase) in which 5 vowels (a, e, i, o, u) and 21 consonants.
- The writing style of this language is from left to right.
- Basically the English characters are in isolated form.
- Punjabi is primarily a Gurumukhi language was popularized by Guru Angad Dev Ji and is 14th most widely spoken language in the world.
- The Gurumukhi language includes the basic 35 different characters.
- The writing style of this script is also from left to right.
- Alike English language this language does not contain lowercase and uppercase characters concept.
- In Gurumukhi word, headline at the top of characters connect the characters with each other.
- The characters share same pixels of headline. Thus segmentation of Gurumukhi characters is difficult as compared to English characters segmentation.

Hence these above described properties of languages are used to design optical character recognition system for recognizing these languages. The detail description of proposed system is given in the next section.

3. Proposed System Description

The proposed system consists of pre-processing, segmentation, feature extraction and classification steps to recognize English as well as Gurumukhi languages. This is shown in Fig. 1.

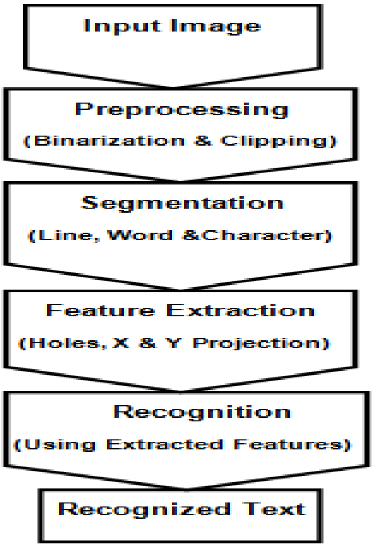


Fig. 1. Proposed framework

Initially, with the objective of modification of raw image, the image is passed through pre-processing stage. At this stage system performs binarization and clipping functions on image to make it useable in the forthcoming levels of system [8]. Binarization technique involves conversion of gray scale images to binary images with a standard threshold value. The pixel value above threshold value set as 1 and pixel value below threshold value set as 0. After binarization, clipping function is performed on image to extract text from the image and deduct spare pixels around the text. Then image is further passed to segmentation level.

Text segmentation is a three step process as shown in Fig. 2. This process divides the text image into lines, words and characters to get a single segmented character to extract unique features of characters. In segmentation algorithm the input text image is inverted (background pixels have 0 value and text pixels have 1 value) for both English and Gurumukhi languages, to achieve a fine segmentation of text.

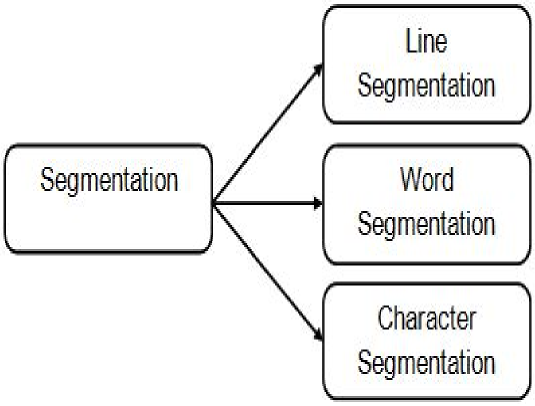


Fig. 2. Steps of segmentation

In line segmentation, horizontal projection of text image is obtained to find empty rows from all rows that contain text [11]. For this purpose image is scanned row wise starting from first row and sum of whole pixels in a row is calculated. If sum is found 0, it means all the pixels in that row are background pixels and this row is considered as empty row. Thus image from first row to this last processed row is segmented and considered as the first line of the image. The process is repeated till all rows of the image are scanned. The height of text lines is almost same in both languages so same line segmentation technique is applicable for both languages.

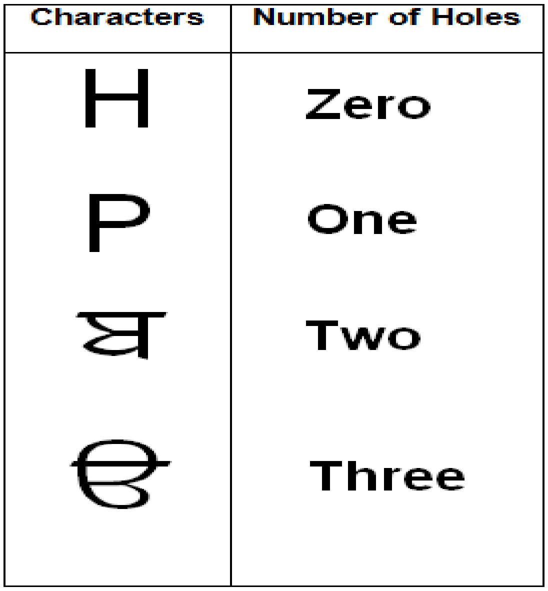
The next step is segmenting words in extracted text line. Unlike the line segmentation same word segmentation technique is not applicable for both languages. As it is known English characters are in isolated form in a word. So while segmentation, it creates confusion between character width and word width (that are of more width). Therefore a threshold value is calculated by analyzing the all blank spaces among the characters as well as words of English line. Then vertical projection profile is calculated to scan an image throughout columns for detection of particular spaces which exceeded threshold value to separate the words from a text line. In Gurumukhi language, headline at top of characters connect them in a word by sharing same pixel value. But words in any line are separated through the empty columns. So for word segmentation, vertical projection profile is calculated to find sum of whole pixels in a column. If sum is found 0, column is considered as empty space between words. Thus image from first column to this last processed column is segmented and considered as first word of processing line. The process is repeated till whole columns of the line are scanned.

Now third level of segmentation stage is character segmentation in every isolated word. Similar to word segmentation different character segmentation techniques are followed for both languages. In English words the characters are separated by empty space. Hence characters are separated by calculating vertical projection profile in the same manner as discussed above. The Gurumukhi characters in any word are connected with a connector line. Here the width of this connector line is set as the threshold value. Then vertical projection profile of word image is calculated to scan throughout the columns for detection of columns having values greater than threshold. The columns having values greater than threshold value are character pixels. The column whose value is found less than or equal to threshold value is considered as connector line. The character image from first column to this last processed column is segmented and considered as first character of processing word image. Next step is to extract features from single segmented character so it further passed to feature extraction level.

Features reveal importance of a character which describes its own identity and distinct it from other characters. Therefore features have great importance in the OCR system to recognize the characters of target languages. The selection of features depends on type of script (English or Gurumukhi or any other language), style of characters (printed or handwritten) [2]. Generally features are categorized in two categories structural and statistical. Structural feature describe geometry whereas statistical feature describe topology of a character. The selected features maximize distinction between two target languages. These features are:

- Number of holes
- Histogram projection (X &Y projection) profile

These features belong to structural and statistical approach respectively. The number of holes feature is illustrated in Fig. 3.



Characters	Number of Holes
H	Zero
P	One
ਬ	Two
ੲ	Three

Fig. 3. Example of holes concept in characters

The histogram projections of character is calculated to represents unique horizontal (X) and vertical (Y) projection profile of character which help to distinct one class character from other class. The horizontal and vertical projections of different characters are shown in Fig. 4.

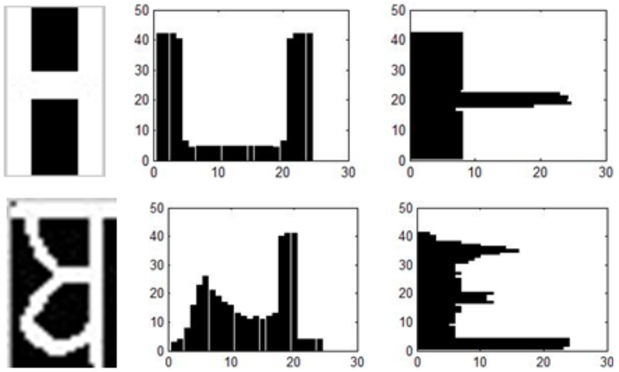


Fig. 4. Illustration for X-projection and Y-projection of characters

After feature extraction, recognition is a next most significant level of optical character recognition system. In research publication various methods of recognition has been presented by researchers such as clustering methods, feature extraction methods, artificial neural network, kernel methods and pattern matching [9]. In this work, feature extraction method has been used for recognition of characters. Here recognition of characters is achieved by using the features which were extracted in the previous level. The number of holes feature is used to divide all characters into four different groups according to number of holes present in the characters such as zero hole character, one hole characters, two holes characters, three holes characters. This feature reduces processing time of system by enabling the system to find segmented character only inside corresponding group instead of whole database. Along with number of holes the groups also include other features of character such as Letter Value, Letter horizontal projection profile, Letter vertical projection profile. Hence incoming character is recognized by comparing its features with the features stored in corresponding group of characters and result is saved in a text file.

4. Results

In this section, results are described to validate proposed work. To show the results a GUI has been made. The processing steps of GUI are as:

Step 1 First step is to select type of script to be recognized as shown in Fig. 5.

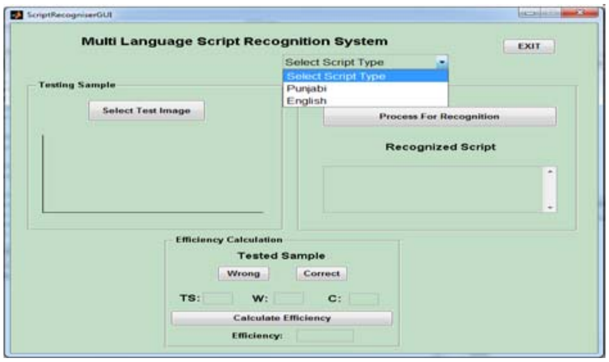
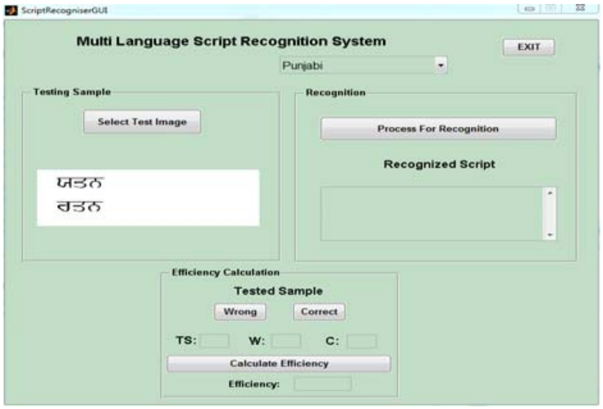


Fig. 5. Illustration for step 1

Step 2 Select the test image from stored database as shown in Fig. 6

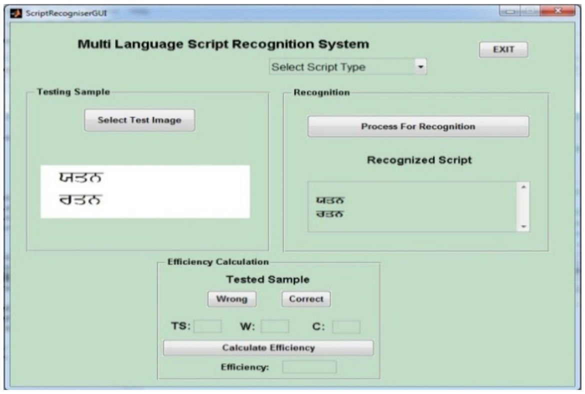


Punjabi

ਘਤਨ
ਰਤਨ

Fig. 6. Illustration for step 2

Step 3 Click on “Process for Recognition” button for recognition processing of test image. The recognized sample will be displayed in edit text box. GUI window is shown in fig. 7.



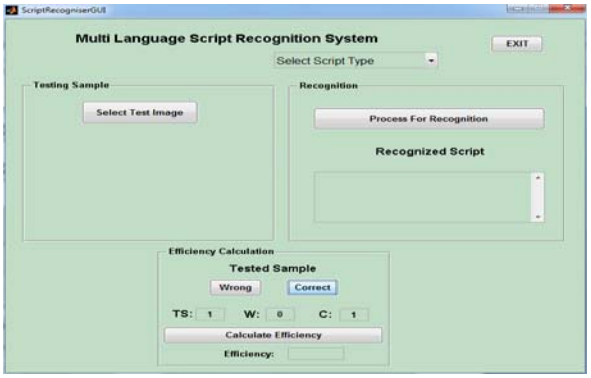
ਘਤਨ
ਰਤਨ

Recognized Script

ਘਤਨ
ਰਤਨ

Fig. 7. Illustration for step 3

Step 4 Now update the status of recognized sample as correct or wrong. Once the status has updated then system is ready for further use as demonstrate in Fig. 8.

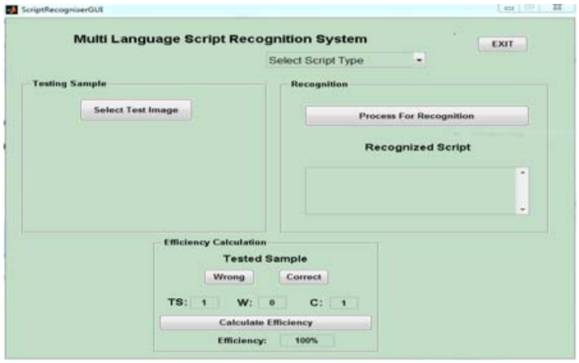


TS: 1 W: 0 C: 1

Efficiency:

Fig. 8. Illustration for step 4

Step 5 Click on “Calculate Efficiency” button to calculate accuracy rate of system as in Fig. 9.

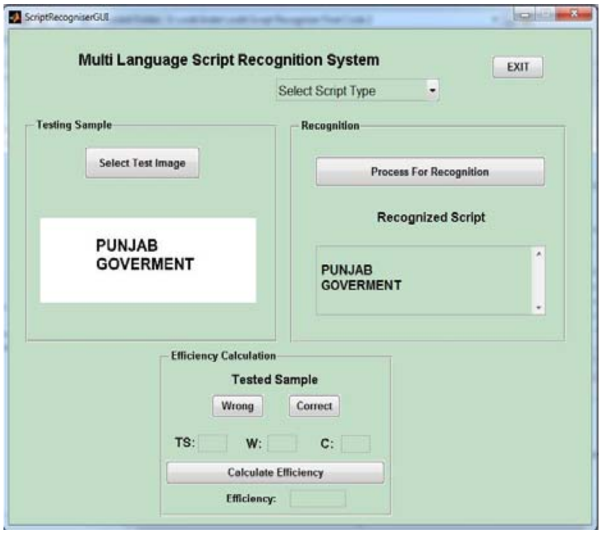


TS: 1 W: 0 C: 1

Efficiency: 100%

Fig. 9. Illustration for step 5

Similarly recognition result of two lines English text shows in Fig. 10.



PUNJAB
GOVERMENT

Recognized Script:

PUNJAB
GOVERMENT

TS: W: C:

Efficiency:

Fig. 10. Illustration for recognized English text

For experiment, testing samples of various sizes have taken for both languages. These samples have been divided into groups such as 25, 50, 75 and 100 to check variations in recognition accuracy rate. These mix samples contain single character, single word & single line and multi lines of different text size of both languages. A graph has plotted which shows a variation in recognition accuracy vs. different number of samples as shown in Fig. 11. For 25 testing samples recognition rate is 100 as illustrated in Fig. 12, for 50 testing samples recognition rate is 96, for 75 testing samples recognition rate is 93 and for 100 testing samples recognition rate is 91. As it is clear from the graph that recognition accuracy rate is decreasing with increase in number of samples. This happens because some samples of single line are not recognized correctly as shown in Fig.13.

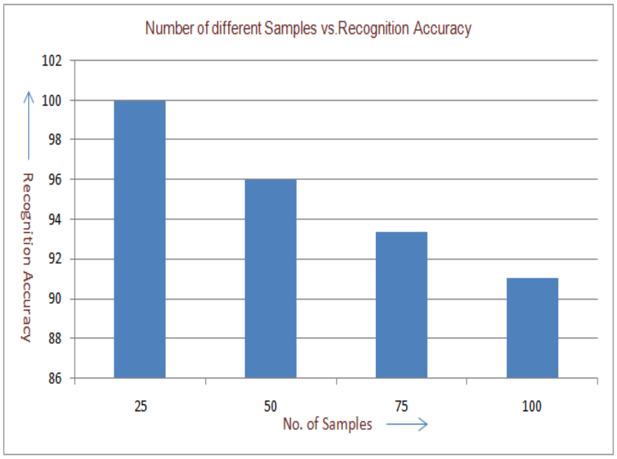
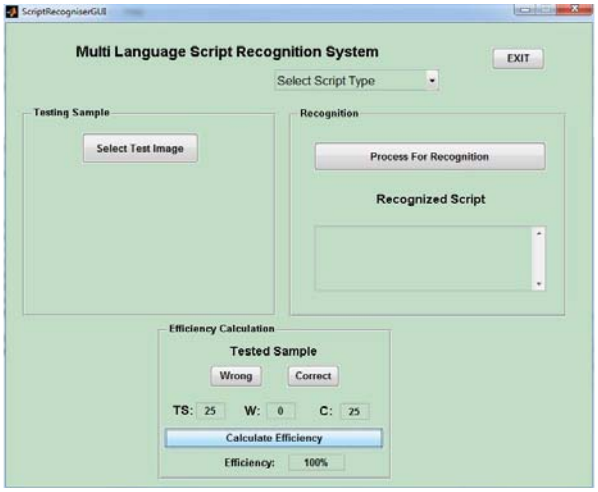


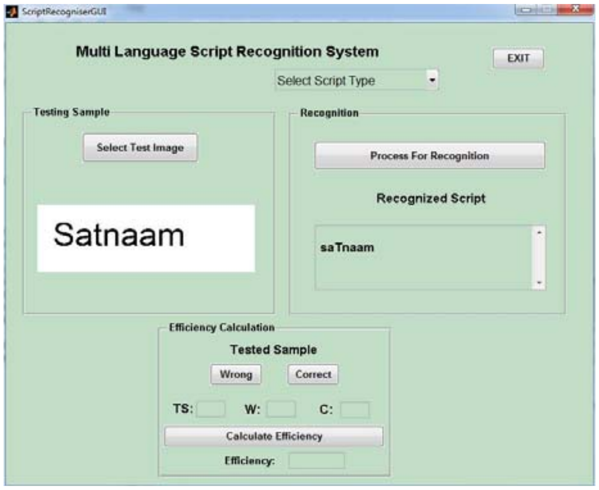
Fig. 11. Graph of recognition accuracy vs. number of samples



TS: 25 W: 0 C: 25

Efficiency: 100%

Fig. 12. Illustration for recognition accuracy of 25 samples



Satnaam

Recognized Script: saTnaam

Fig. 13. Illustration for misrecognized text

Table 1: Recognition accuracy results for english and gurumukhi

No. of Different Samples Taken For Recognition	***No. of Samples Correctly Recognized***	***No. of Samples Wrongly Recognized***	***Recognition Accuracy in (%age)***
25	25	0	100
50	48	2	96
75	70	5	93
100	91	9	91

5. Conclusion

In described work, a method to recognize multi-languages is presented. The main aim of proposed method is to identify more than one language where English and Gurumukhi are chosen. The assumed system is based on the analysis of features like numbers of holes and projection histogram which are implemented for recognition of both languages. The determined features have provided desired solution in recognition of English and Gurumukhi languages. The segmentation process also played important role in recognition process by giving fine segmented characters of text of both languages. The experimental result shows that the proposed method is appropriate to recognize English and Gurumukhi text. In future, one can hope that more features would be added to system to achieve high accuracy and trained for different fonts.

References

[1] A. J. Faaborg, “Using Neural Networks to Create an Adaptive Character Recognition System,” Cornell University, lthaca, NY, 2002.

[2] A. K. Jain, T. Taxt, “Feature Extraction Methods for Character Recognition - A Survey,” proceedings of the Pattern Recognition, vol. 29, No. 4, July 1996, pp. 641-662.

[3] G. S. Lehal and C. Singh, “A Technique for Segmentation of Gurumukhi Text,” Springer Berlin Heidelberg, Sep. 2001, pp. 191-200.

[4] J. J. Weinman, E. L. Miller and A. R. Hanson, “Scene Text Recognition using Similarity and Lexicon with Sparse Belief Propagation,” IEEE Pattern Analysis and Machine Intelligence, vol. 31, Feb. 2009, pp. 1733-1746.

[5] M. A. Ranzato, Y. L. Boureau and Y.L. Cun, “Sparse Feature Learning for Deep Belief Networks,” Advances in Neural Information Processing Systems, 2007.

[6] M. K. Jindal, G. S. Lehal and R. K. Sharma, “A Study of Touching Characters in Degraded Gurmukhi Text,” International Journal of Computer, Information Science and Engineering, vol. 1, 2007.

[7] P. Mahasukhon and H. Mousavinezhad, “Hand-Printed English Character Recognition based on Fuzzy Theory,” IEEE International Conference on Electro/Information Technology, May 2012,pp. 1-4.

[8] R. Mithe, S. Indalkar, N. Divekar, “Optical Character Recognition,” International Journal of Recent Technology and Engineering, vol. 2, Mar. 2013.

[9] S. G. Dedgaonkar, A. A. Chandavale and A. M. Sapkal, “Survey of Methods for Character Recognition,” International Journal of Engineering and Innovative Technology, vol. 1, May 2012.

[10] S. H. Kasaei, S. M. Kasaei and S. A. Kasaei, “New Morphology- Based Method for Robust Iranian Car Plate Detection and Recognition,” International Journal of Computer Theory and Engineering, vol. 2, April 2010.

[11] S. Taha, Y. Babiker and M. Abbas, “Optical Character Recognition of Arabic Printed Text,” IEEE Conference on Research and Development, Dec. 2012, pp. 235-240.

[12] Xin Fan and Guoliang Fan, “Graphical Model for Joint Segmentation and Recognition of license Plate Characters,” IEEE Signal Processing Letters,vol. 16, Jan. 2009, pp. 10-13.

[13] Y. F. Pan, X. Hou and C. L. Liu “Text Localization in Natural Scene Images on Conditional Random Field,” International Conference on Document Analysis and Recognition, 2009.

[14] Z. Saidane and S. Gracia,“Automatic Scene Text Recognition using Convolution Neural Network,” Workshop on Camera-Based document Analysis, 2007.