University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 2, Issue 2, April 2018
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Optical Character Recognition System for Sindhi Text: A Survey

Waseem Javaid Soomro, Dil Nawaz Hakro, Imdad Ali Ismaili, Ghulam Mustafa Shoro

Institute of Information and Communication Technology, University of Sindh, Jamshoro
waseem.soomro@usindh.edu.pk, dilnawaz@usindh.edu.pk, iai_a@yahoo.com, gm.shoro@usindh.edu.pk,

Abstract: Optical character recognition is popular field for researchers during last decade of research, which is able to successfully recognize the scanned English image into editable text form. However, optical character systems for other regional languages such as Urdu, Arabic, and Sindhi, still presents a huge challenge and implementation problems. Thus, in this paper various techniques of optical character recognition system for such low level regional languages have been discussed and analyzed. This survey paper consolidates all such techniques and presents an overview to aid researcher understand the methodology of performing and implementing OCR system for Sindhi language.

Keywords: Optical Character Recognition, Sindhi OCR, Sindhi Computing, Sindhi Language, Character Recognition, Segmentation

I. INTRODUCTION

In this particular section most of us make clear the actual material publicized in a variety of newsletters, seminars symposiums. Almost all of the strategies reviewed within this section are manufactured with relevance to the applications. Common approaches are nevertheless to be considered at pertaining to number of applications. The latest progress associated with Information Technology shifted a new roman set of scripts practically to help their foundation along with workable and very adoptable with every one of the regions of research.

A substantial work has been done for the Persian script yet an extremely little effort has been done about the 'languages' implementing Persian script. An extraordinary work is available within computerizing Sindhi vocabulary, however the work towards Sindhi Optical Character Recognition is in the foundation place and almost non-existent. Some of the efforts have been found on this spot other than few basic work done in [1][2][3][4].

The published studies are not on the Sindhi words, additionally the difficulties related to that script have become challenging as well as a lot of causes can be known for having less attention from government about this part of Sindhi research area, but the critical trouble may be considered as financial resources [1].

The Sindhi along with Urdu has a prolonged extension of the actual Arabic script; Urdu features 39 characters in comparison with the actual 28 of authentic Arabic script, Sindhi script is written with all the modified type of the actual Arabic script, as compared with 28 in the Arabic [1] that includes the actual fifty-two unique letters. Just about every character features multiple designs using the spot along with easy use in the word. The character will change its shape in the sentence that used with contain on as well as making it character. The character may possess the "initial (start), medial (middle), final (last), remote (standalone)" shape according. Apart of both first, medium along with last characters have the instances, while isolated character is referred to as the remote as well as standalone character. Whenever Character combinations are available in its design (form), based on its spot, it is truly is known as word sensitivity [5].

However, many of the character are similar on paper, an extremely modest alter (Modification) inside an identity may generate one more identity. All the character is usually gathered in multiple categories or groups based on their own shapes (called base groups). Every category contains only two to eight characters. All the categories and character are usually notable or labeled using the number of dots and location involving dots (Accent Mark). Throughout Sindhi dialect these marks are put on over or beneath the bottom appearance. More than half of the character (out involving 52) is part of this kind of groups and one, a couple of, 3 to 4 dots are widely-used to tell apart among several characters. Arabic and Urdu dialect are producing the application of a single, a couple of, or three dots, but in Sindhi, we are able to make use of 4 dots, this kind of depth creates further complex characters and presents more

problems throughout distinguishing and identification of the character.

II. *ARTIFICIAL* INTELLIGENCE

Artificial Intelligence is a field of study which makes computer able to think and decide according to experience. It is an art to create machine to decide on the basis of Man-made thinking ability. Extensive variety of meanings is available, whereas the proper explanation concerning Artificial Intelligence has been provided by [06]. From a computer perspective it is considered as one of the important branch which provides man made thinking and abilities ranging from image Acquisition, to all processing involving images and videos [06].

A. Machine Vision

Computer vision is the scientific disciplines regarding endowing personal computers or different equipment together with imaginative and prescient vision, or the opportunity to see. Nearly all personal computer vision professionals might consent that experiencing is a lot more than the task regarding creating mild in a very kind that can be performed rear, such as creating of the camcorder. However precisely what, exactly, is needed is the detection or creation regarding mild to be able to capture a product, whether natural or produced, is experiencing. Computer vision is the scientific disciplines involving endowing desktops or various other machines using imaginative and prescient vision, or the chance to notice.

Computer vision is the research associated with endowing personal computers as well as additional devices with imaginative and prescient vision. What exactly will it indicate to discover is the majority of computer vision users would certainly recognize in which viewing is usually greater than the task associated with taking mild in a type which might be played out returning, such as the taking of a camcorder. Although what exactly, just, is needed in addition to the discovery as well as taking associated with mild to be able to claim that your device, whether it be natural as well as manufactured, is usually viewing?

Computer vision would be the technology of endowing computers or perhaps different products having imaginative and prescient vision, or perhaps to be able to discover. Nearly all computer vision scientists might concur that will be capturing can be more than the task of capturing within a style which can be converted back again, just like the saving of a video camera.

B. Document Image Understanding

From the knowledge-based image understanding, it is very important to recognize design structures in individual files specifically for working with convenient record designs in the files. Inside knowledge-based document graphic comprehending, it is necessary to recognize the document of particular image. A document may behave to work with convenient document type. The classification technique on the basis of the arrangement and confirmation model splits several categories of documents into suitable document types stepwise, as shown in the Figure 1.

Within the knowledge-based image understanding, it is important to separate the design constructions regarding character files specifically having a look at to be able to work with adaptable file format [7].

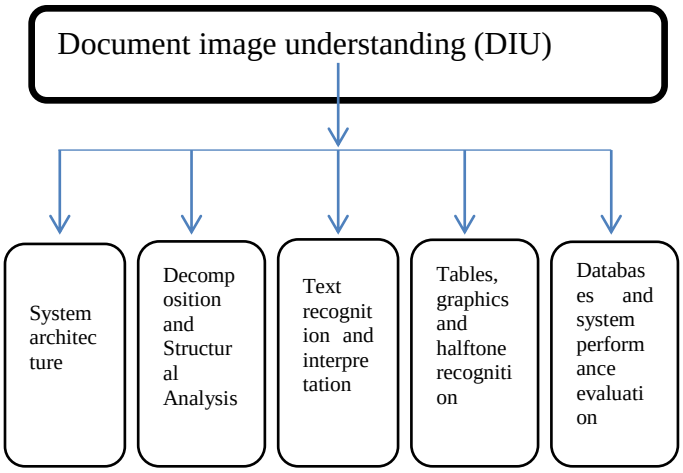

Figure 1: Types of Document Image Understanding [1]

C. Text Recognition and Interpretation

Extraction of information from the text image is difficult task which could be achieved by pattern recognition. In pattern recognition, the image is recognized with the skeletons and recognition of particular character code is generated. The most difficult task in recognition is to detect the exact information from the raw image and extraction of the information [8].

D. Optical Character Recognition (OCR)

Character Recognition is widely used system by which data available in form of the printed text and written format available in books, novels and

newspapers. Through the OCR, printed text can be converted into the editable soft copy. An OCR generally converts the geometrically based source into the binary form to make understand the computer to read it in form of ASCII or in form of the digital format.

E. Origin of OCR

The reason for development of the OCR is back in 1950's "GIZMO" which was using the job of interpretation and PC control. Jessi cryptanalyst and Harvey mark built the new developments in the area of OCR. It was observed at that time that the achievement was useful, as different companies adopt this system like Jesse Shepard including standard acrylic firm for telephone organizations through their readers which were responsible for sending and receiving of mail messages [9]. Automatic mailing appeared as the biggest achievement in this research area, at primarily level it was inside development, later on it advanced with a fast pace. Through 1965, John Rabinow was the first to get the copyright authentication of OCR and on behalf of this employed by U.O hydrates Postal Program [9]. Since the OCR is attractive arena, imposing a lot of challenges along with complication and also open-handed number of replacements inside the portions of Unusual Intelligence. These features make OCR quite interesting and exciting to have work with lots of challenges and difficulties.

F. Hidden Markov Models

Decerbo, et al. 2004 put in place the screenplay independent strategy by using Undetectable Markov Designs for Pashto Byblos technique. Mcdougal claimed the device has been tried in diverse languages and also the technique is trainable. This line segmentation is not required as well as segmentation of words. Alteration of Byblos technique can be used for new terminology with a few stage courses change. The final results usually are generated by converting British, China along with Arabic languages [10].

Cavalin (2006) utilized the particular segmentation technique for segmenting and recognition associated with guitar strings or heroes associated with individual creation. Strings or numerals are actually segmented by using two conditionals HMM dependent approaches. Foreground and background functions have been put together to adjust the loss for the words associated with recognition. The technique seemed to be applied upon remote identity guitar strings and also numerals solely [11] [12].

G. Holistic Approach

Clocksin and Khorsheed (2000) employed a fresh method of discerning cursive Persian terms. There're words are not segmented and the attributes have been extorted from unsegmented word impression. The word is usually handled since without segmentation and in this way, words are handled with overall shapes, that's the reason it's possible to state it is a holistic approach pertaining to word recognition. Each and every word may be displayed by the independent theme. Euclidean distance from web templates is used for the recognition of the word [13] [14] [15].

H. Ligature identification

Erlandson et at., (1996) proposed a technique through which ligature segmentation along with recognition is actually realized for Arabic word recognition without character segmentation. They have used one of a function vector for characterizing of Arabic phrases. They have created a databases of function vectors, while indicated function vector can be found it truly is coordinated while using the databases function vectors checked out by making use of regarded phrases easily obtainable in book. There are many function vectors for each and every term in a very book made up of forty-eight, 190 records. The saying is actually generated seeing that theory when its function vectors are extremely identical. [16].

I. Descending Window

Tolba et al., (1990) segmented Arabic characters by making use of dropping window approach. The window may be moved above each script coming from left dropping from each instant so that the segmentation parameter may be determined, followed by evaluation, the identified spot would be named specific location and in case of segmentation parameter is gloomier in comparison with threshold. The specific location will be found immediately after the start of the type identified to the finish of the characteristic. It will likely be another characteristic outset if the segmentation parameter is increased. [17].

J. Segmentation cost –free strategy

Another technique of segmentation-free strategy where segmentation is bypassed and direct recognition is executed [18]. The actual system has been directly divided into four main phases. The initial is the particular element extraction and that is the point base connected with training; the next level is actually selection matrix development by inverse place vector. Next level character is searched and

identified based on selection matrix and the last stage deals with managing the particular available interruptions [18] [19] [20].

K. Utilization of Morphological Operator in Persian OCR

Jelodar, et al., (2005) extracted local features showing some sort of design and local textual content. They've used morphological operator method; this method is called Hit/Miss operator for sub-word procedure. Elevated number of errors are claimed throughout segmentation step. This Acceptance process has been shown throughout some steps leaving one side result period. Feedback stage uses the bitmaps from first stage along with input regarding preprocessing. They divided collections along with sub-words. Counting the word and dots along with thinning algorithms would be the materials in this stage. Hit/miss operator has been employed in the radical lookup process to learn the functions inside characteristic extraction phase. The final phase is specialized in conclusion along with classification for final results. The system has been examined upon merely lotus fonts by utilizing a couple of measurements along with the accuracy claimed is 99. 9% [21].

L. Fuzzy Logic

Ouslama and Kishibe (1999) suggested an approach which in turn combines the actual structural as well as record strategy on fuzzy reasoning pertaining to characteristic extraction and classification. The authors of the studies have done segmentation firstly into principal then complement stage. Functions had been extracted via number of complement heroes as well as directory as well as side to side projection profiles from the principal character that had been employed in classification. Class ended up being using the list of fuzzy regulations. High accuracy rates are actually reached since they have got tested acceptance criteria on diverse fonts [22] [23] [25].

III. OCR FOR OTHER LANGUAGES

In this section we have surveyed the work done so for on the Arabic script adapted languages, like Urdu and Persian languages. Designing of OCR for the different languages requires special approaches to design comprehensively.

A. Urdu OCR

Pal and Sarkar (2003) presented their work based on water reservoir for Urdu isolated characters. Obtained series of segmentations by applying Hough transform. The actual element labels had been for the character segmentation. Drinking water tank approach may be for the topological, shape-based features. Sampling classifier had been created to discover the insight on the process, decides the existence and also absence of the element. 97. 8% reliability had been believed by the authors. The actual segmentation problems have been reported as 0.7% and also sampling category problems reported as 1.1% respectively [24]. A HMM and fuzzy logic was combined to produce a hybrid algorithm for recognition of Urdu based scripts [25]. Similarly, a nastaleeq based specific algorithm was also successfully implemented in [26] [27] [28].

B. Arabic OCR

Hamid & Ramziet (2001) utilized artificial Neural network to the segmentation connected with Arabic word. The Artificial Neural Network (ANN) proof is utilized just as one more time [29]. Three phases were used before the use of ANN. These types of methods are generally scanning, binarization and feature extraction. A regular similar formula can be employed by the actual segmentation of the linked prevents connected with character along with for producing pre-segmentation factors for these which prevents connected with character. Bringing in the actual topographic functions from the word along with calculating segmentation factors include the basic heuristic of the function. Soon after segmenting factors, the actual artificial Neural network is in charge of the actual proof of the accuracy and reliability of such segmentation factors [29]. Similarly, many other studies have been done to propose an efficient Arabic OCR system [30] and recognition based segmentation was done to achieve higher accuracy [31]. Similarly, Neuro-Heuristic technique was implemented to segment hand written Arabic text [32].

C. Telugu OCR

Patvardhan et al., (2004) employed particular character connected with attached models connected with Telugu textual content as the fundamental icons are regarded as the cornerstone connected with identification. Their particular program can be producing phonetic English as the OCR end result and then it really is converted to be able to editable Telugu textual content. The system manages numerous fonts and varying dimensions. The system reaches 98% accuracy [33].

D. Oriya Script

Chaudhuri et al., (2002) utilized the actual merging for typical strategies in addition to their proposed to

recognize the actual produced Oriya recognition. Skew correction, series and expression segmentation have been performed with the merging for typical strategies, such as projection and many others. Functions are actually produced by making use of strokes and run-number strategies. Water reservoir process has been employed by the authors. The recognition process contains each sub process through which one particular point offers the recognition on the modified character plus the different spotting left over characters. The character is viewed modified and characters are having diverse dimensions and positions point out within top zone or lower zone. The authors claim accuracy based on the stages of the OCR program. word segmentation with 97.5%, expression segmentation with 97. 7% and character segmentation 97%. The average accuracy on the system has been stated with 96. 3% with using the error cost for 3. 7% [34].

E. Khanada Script

Sagar et al., (2008) started work on Khannada script employ the approach associated with database. The authors with this system have reported the accuracy of 100%. The reason for such high accuracy was the non-cursive nature of Khannada language when the character (Aksharas) usually are written separately [35].

F. Chinese ICR

Li et al., (2003) proposed method associated with Adaptive duration based to segment and evaluate the identification associated with Oriental ICR. Mcdougal described the actual greater performance from the adaptive algorithm in comparison with the actual repaired duration procedure. The sliding window strategy continues to be used plus the time period is set through the shape along with to evaluate from the character. The standard stroke calculation along with query associated with selection restrictions intended for segmentation are classified as the major actions with this program. Judging by evaluate effects, the result chart is plotted. The minimum route associated with arithmetical essential mean dissimilarity is actually selected a final decision associated with remaining segmentation along with classification. Mc Dougal features extracted the machine about 500 Oriental address wrinkles associated with email made up of cracked, handled, usually designed character images in binary format. The top 5 characteristically correct identification is actually described 93% as the correct identification intended for address text is actually over 90% [36].

IV. CONCLUSION

In this survey we presented a brief overview of the various OCR systems used for multiple regional languages of the world. Initially various techniques used in OCR systems are discussed from Artificial Intelligence, holistic, ligature based to fuzzy logic approach. Then the applications of these techniques with respect to various regional languages are analyzed and elaborated for languages such as Arabic, Urdu, Telegu and Chinese. This survey gave a detailed overview of various OCR techniques and their uses in each language, to develop an understanding and conceptual overview to develop a suitable OCR system for Sindhi language. This survey paper consolidates all such techniques and presents an overview to aid researcher understand the methodology of performing and implementing OCR system for Sindhi language.

V. REFERENCES

[1] Hakro, D.N., Ismaili, I.A., Talib, A.Z., Bhatti, Z., & Mojai, G.N., (2014) Issues and Challenges in Sindhi OCR. Sindh University Research Journal (Science Series). Vol.46 (2). Pp. 143-152. Sindh University Press. June 2014.

[2] Hakro, D.N., Talib, A.Z., Bhatti, Z., & Mojai, G.N., (2014) A Study of Sindhi Related and Arabic Script Adapted languages Recognition. Sindh University Research Journal (Science Series). Vol.46 (3). Pp. 323-334. Sindh University Press. October

[3] Bhatti, Z., Waqas, A., Ismaili, I. A., Hakro, D. N., & Soomro, W. J. (2014). Phonetic based SoundEx & ShapeEx algorithm for Sindhi Spell Checker System. arXiv preprint arXiv:1405.3033.

[4] Bhatti, Z., Ismaili, I. A., & Soomro, W. J. (2015). Phonetic-Based Sindhi Spellchecker System Using a Hybrid Model. Digital Scholarship in the Humanities, fqv005

[5] C. Vasantha Lakshmi1 and C. Patvardhan "An optical character recognition system for printed Telugu text , Pattern Analysis & Applications", Category, Theoretical Advances, Volume 7, Number 2 / July, 2004 Pages 190-204

[6] C. Vasantha Lakshmi, C. Patvardhan, "A high accuracy OCR System for Printed Telugu Text".TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region Volume 2, Issue , 15-17 Oct. 2003 Page(s): 725 - 729 Vol.2

Digital Object Identifier 10.1109/ TENCON. 2003.1273274

[7] Hiroyuki Masai 1'2 and Toyohide Watanabe I, Document Categorization for Document Image Understanding, Department of Information Engineering, Graduate School of Engineering, Nagoya University Furo-cho, Chikusa-ku, Nagoya 464-01, Japan

[8] Document Image Processing of Indian Scripts,. Special Issue of Sadhana 2002.

[9]http://en.wikipedia.org/wiki/Optical_character_recognition, this page was last modified on 1 November 2010 at 12:31, Mindmatrix (14,095 bytes).

[10] Michael Decerbo, Ehry MacRostie, Premkumar Natarajan, (2004) " The BBN Byblos Pashto OCR System", HDP'04, November 12, 2004, Washington, DC, USA. Copyright 2004 ACM 1-58113-976-4/04/0011

[11] Cavalin, P. R. (2006, April). An implicit segmentation-based method for recognition of handwritten strings of characters. In *Proceedings of the 2006 ACM symposium on Applied computing* (pp. 836-840). ACM.

[12] Cavalin, P. R., Sabourin, R., & Suen, C. Y. (2010, March). Dynamic Selection of Ensembles of Classifiers Using Contextual Information. In *MCS* (Vol. 10, pp. 145-154).

[13] Khorsheed MS, Clocksin WF, Spectral features for Arabic word recognition. The IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'2000, Istanbul, Turkey, June 5-9 , 2000, pp,3574-3577.

[14] Baecher, P., Büscher, N., Fischlin, M., & Milde, B. (2011). Breaking reCAPTCHA: a holistic approach via shape recognition. *Future challenges in security and privacy for academia and industry*, 56-67.

[15] Bhattacharya, S., Sukthankar, R., & Shah, M. (2011). A holistic approach to aesthetic enhancement of photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *7*(1), 21.

[16] Erlandson E, Trenkle J, Vogt R, Word level recognition of multifont Arabic text using a feature vector matching approach, Proceedings of International Society for Optical Engineers, SPIE, 1996; 2660: 63-70.

[17] Tolba M, Shaddad E. On the automatic reading of printed Arabic characters. Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Los Angeles, CA, 1990; 496-498.

[18] Casey, R. G., & Lecolinet, E. (1996). A survey of methods and strategies in character segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *18*(7), 690-706.

[19] Su, T. H., Zhang, T. W., Huang, H. J., & Zhou, Y. (2007, September). HMM-based recognizer with segmentation-free strategy for unconstrained Chinese handwritten text. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on* (Vol. 1, pp. 133-137). IEEE.

[20] Gatos, B., Ntzios, K., Pratikakis, I., Petridis, S., Konidaris, T., & Perantonis, S. J. (2006). An efficient segmentation-free approach to assist old Greek handwritten manuscript OCR. *Pattern analysis and applications*, *8*(4), 305-320.

[21] M.Salmani Jelodar, M.J.Fadaeieslam , N.Mozayani,M.Fazeli, (2005) "A Persian OCR System using Morphological Operators", Transactions on Engineering, Computing and Technology v4 February 2005 ISSN 1305-5313.

[22] Bouslama F, Kishibe H. Fuzzy logic in the recognition of printed Arabic text. IEEE Transactions on 1999: 1150-1154

[23] Singh, R., Yadav, C. S., Verma, P., & Yadav, V. (2010). Optical character recognition (OCR) for printed devnagari script using artificial neural network. *International Journal of Computer Science & Communication*, *1*(1), 91-95.

[24] Pal, U., & Sarkar, A. (2003, August). Recognition of printed Urdu script. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on* (pp. 1183-1187). IEEE.

[25] Razzak, M. I., Anwar, F., Husain, S. A., Belaid, A., & Sher, M. (2010). HMM and fuzzy logic: a hybrid approach for online Urdu script-based languages' character recognition. *Knowledge-Based Systems*, *23*(8), 914-923.

[26] Ahmad, Z., Orakzai, J. K., Shamsher, I., & Adnan, A. (2007, December). Urdu nastaleeq optical character recognition. In *Proceedings of world academy of science, engineering and technology* (Vol. 26, pp. 249-252).

[27] Ul-Hasan, A., Ahmed, S. B., Rashid, F., Shafait, F., & Breuel, T. M. (2013, August). Offline printed Urdu Nastaleeq script recognition with bidirectional LSTM networks. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on* (pp. 1061-1065). IEEE.

[28] Shah, Z. A. (2002, December). Ligature based optical character recognition of Urdu-Nastaleeq font. In *Multi Topic Conference, 2002. Abstracts. INMIC 2002. International* (pp. 25-25). IEEE.

[29] Hamid, A.and Haraty, R., "A Neuro-Heuristice Approach for Segmenting Hand written Arabic Tex", ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, 25-06-2001 – 29-06-2001, pp: 110-113.

[30] Bazzi, I., Schwartz, R., & Makhoul, J. (1999). An omnifont open-vocabulary OCR system for English and Arabic. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(6), 495-504.

[31] Cheung, A., Bennamoun, M., & Bergmann, N. W. (2001). An Arabic optical character recognition system using recognition-based segmentation. *Pattern recognition*, *34*(2), 215-233.

[32] Hamid, A.and Haraty, R., "A Neuro-Heuristice Approach for Segmenting Hand written Arabic Tex", ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, 25-06-2001 – 29-06-2001, pp: 110-113.

[33] C. Vasantha Lakshmi1 and C. Patvardhan " An optical character recognition system for printed Telugu text , Pattern Analysis & Applications", *Category, Theoretical Advances, Volume 7, Number 2 / July, 2004 Pages 190-204*

[34] B B CHAUDHURI, U PAL and M MITRA,"Automatic recognition of printed Oriya script".Sadhana Vol.27, Part 1,February 2002,pp.23-34.(c) Printed in India

[35] B.M. Sagar, Dr. Shobha G, Dr. Ramakanth Kumar P."OCR for printed Kannada text to Machine editable format using Database approach ", WSEAS TRANSACTIONS on COMPUTERS, Issue 6, Volume 7, June 2008, ISSN: 1109-2750.

[36] LI Guo-hong (李国宏)†, SHI Peng-fei (施鹏飞) "An approach to offline handwritten Chinese character recognition based on segment evaluation of adaptive duration".Journal oa Zhejiang University SCIENCE, ISSN 1009-3095, Li et al/ J Zhejiang Univ SCI 2004 5(11):1392-1397.