

An Algorithm for Japanese Character Recognition

Soumendu Das, Sreeparna Banerjee

School of Engineering Technology West Bengal University of Technology, India
E-mail:sdphotoes@gmail.com, sreeparnab@hotmail.com

Abstract—In this paper we propose a geometry- topology based algorithm for Japanese Hiragana character recognition. This algorithm is based on center of gravity identification and is size, translation and rotation invariant. In addition, to the center of gravity, topology based landmarks like conjunction points masking the intersection of closed loops and multiple strokes, as well as end points have been used to compute centers of gravity of these points located in the individual quadrants of the circles enclosing the characters. After initial pre-processing steps like notarization, resizing, cropping, noise removal, synchronization, the total number of conjunction points as well as the total number of end points are computed and stored. The character is then encircled and divided into four quadrants. The center of gravity (cog) of the entire character as well as the cogs of each of the four quadrants are computed and the Euclidean distances of the conjunction and end points in each of the quadrants with the cogs are computed and stored. Values of these quantities both for target and template images are computed and a match is made with the character having the minimum Euclidean distance. Average accuracy obtained is 94.1 %.

Index Terms—Japanese Optical Character Recognition, geometry, topology, image processing.

I. INTRODUCTION

Optical character recognition (OCR) for both handwritten and printed text is a crucial step towards document analysis and retrieval for the purpose of storing and transmitting text in digital form via computers and networks. Furthermore, different languages have very different characteristics of their alphabets which form the basis of this written text. In Indian languages, for instance, the written script can be broadly divided into Devanagari script for the North Indian languages and the Tamil script for the South Indian languages. OCR for both printed Hindi character [1] s as well as handwritten Devanagari characters [2] constitute a complex task. South Indian languages like Malayan [3] also possess a complex written character system. Japanese and Chinese languages also possess very complex character sets as these include both syllabic character sets as well as ideograms. In this paper we focus on Japanese, which has over three thousand characters comprising of syllabic character characters called Kama and ideographic

characters called Kania. Furthermore, Japanese language is agglutinative and moratorium. It has a relatively small sound inventory and lexically significant pitch accent system and is distinguished by a complex system of honorifics. Japanese text does not have delimiters like spaces, separating different words. Also, several characters in the Japanese alphabet could be home-morphic, i.e. have similar shape definition which could add to the complexity of the recognition process. Thus, Japanese OCR is a very challenging task and many research efforts have been conducted to perform these task. A survey of some of the approaches to OCR for the Japanese language have been discussed in [4].

This paper proposes a geometric topological based algorithm for Japanese character recognition by combining the Size Translation Rotation Invariant Character Recognition and Feature vector Based (STRICR-FB) algorithm originally proposed by Barnes and Manic [5] along with some topological features of the individual characters.

The remainder of the paper is organized as follows. The next section describes the Japanese language model, followed by a description of allied work in the following section. In section IV, a review of the original STRICR-FB is presented. The proposed algorithm is described in Section V. After that, application of the proposed algorithm is discussed in section VI, and our conclusions are given in in section VII.

II. JAPANESE LANGUAGE MODEL

A. Japanese Character Sets

The Japanese language is written in a mixture of three scripts; kana and kanji. The two kana are called hiragana and katakana shown in Figure 1 and 2, respectively. For Japanese words, Hiragana is used, mostly for grammatical morphemes. Katakanas are used for transcribing foreign words, mostly western, borrowing and non-standard areas. In addition, diacritic signs like dakuten and handakuten are used (see Fig 3 and 4.) Dakuten are used for syllables with a voiced consonant phoneme. The dakuten glyph (`) resembles a quotation mark and is directly attached to a character.

Several thousand kanji are in regular use, while the two syllabaries each contain 48 basic Handakuten characters which are used for syllables with a /p/ morpheme. The glyph for a 'maru' is a little circle (°)

that is directly attached to a character. Kanji is derived from Chinese characters and represents logographic or morphological units.

Thus Hiragana is used primarily for grammatical elements - particles, inflectional endings and auxiliary verbs. Katakana is used for writing loan words, onomatopoeic words, to give emphasis, to suggest a conversational tone, or to indicate irony or a euphemism and for transcribing foreign words, mostly western.



Fig 1: Hiragana script



Fig 2: Katakana script

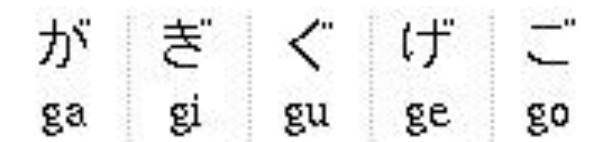


Fig 3: Dakuten alphabets

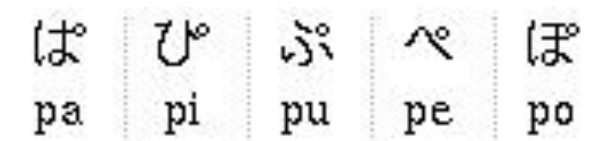


Fig 4: Handakuten alphabets

Kanji are content bearing morphemes. In Japanese text Kanji are written according to building principles like Pictograms (graphically simplified images of real artefacts), ideograms (combinations of two or more pictographically characters) and phonograms (combinations of two Kanji characters). Another important feature in Japanese is interword spacing in Japanese text.

B. Interword Spacing in Japanese

Japanese words are not separated by delimiters like spaces, thus making character recognition difficult. Although Japanese is a word based language, segmenting text into word is not as clear cut as in languages using word spacing as a rule. Spacing is incorporated as in at least two ways. The first way is by adding spaces not

only between their grammatical modifiers and post positions. The second way is to consider the modifiers and post positions as a part of the modified word. Based on the study conducted by Saint et.al [6] using 16 subjects in Japanese reading, 60 word texts from excerpts of newspapers and internet columns, it was concluded that in pure Katakana text, inter-word spacing is an effective segmentation method, in contrast to Kanji-Hiragana text, since visually silent kanji characters serve as effective segmentation uses by themselves.

C. Character Features Vectors Identification

Every character has its own features and identities. By identifying features we can recognize characters from a textual image document. By feature extraction the critical characteristics of characters gets isolated, and that reduces the complexities of the pattern. After classification it compares with known patterns and then matched with the character that has the same characteristics. The characters can be further subdivided into segments and the strokes in each of these segments exhibit certain characteristics in terms of shape, angle of inclination. In addition, presence of dakuten and handakuten also changes the character. All these aspects need to be taken into account in devising feature vectors for identification.

III. ALLIED WORK

As described in [4], most of the recent Japanese character recognition approaches, both for handwritten and printed text, either use soft computing based approaches for classification, or, image shape/morphology characteristics for classification. The most popular of the soft computing approaches is the neural network [7] followed by Genetic Algorithms [8], Hidden Markov Model (HMM) [9] and Support Vector Machine [10].

Image shape based approaches characteristically perform a segmentation of the candidate character and compare with a template of the prototype character template. The size, shape, extent and angular inclinations of the different strokes have also been taken into account in many instances. Additionally, the images of the characters (both candidate and prototype) have been divided into segments and the contours of the strokes pertaining to each of the segments have been studied. These processing prove to be helpful, especially for handwritten character texts where there are variations. Image shape based approaches characteristically perform a segmentation of the candidate character and compare with a template of the prototype character template. The size, shape, extent and angular inclinations of the different strokes have also been taken into account in many instances. Additionally, the images of the characters (both candidate and prototype) have been divided into segments and the contours of the strokes pertaining to each of the segments have been studied. These processing prove to be helpful, especially for handwritten character texts where there are variations.

Some of the major efforts are stated in the following. Convergence of the shortest path has been used as a criterion for segmentation in [11]. An algorithm that takes into account the variations of angles due to pen position trajectories has been presented in [12]. Adaptive context processing [13] is also another technique that has been used. An efficient indexing scheme for image storage and recognition has been reported in [14]. Improvement strategies for template matching have been discussed in [15]. Reference [16] proposes a multistage pre-candidate selection procedure in handwritten Chinese OCR. Recognition enhancement by linear tournament verification is suggested in [17]. Korean OCR is described in [18]. Hull [19] used multiple distortion invariant descriptors for document image matching. Snead et al. [20] generated character candidates for document image retrieval. In [21] Nina, Kagoshima and Shimmer used keyword information for post processing. Linguistic knowledge was found to improve OCR accuracy by Dal, Norton and Taylor [22]. However, an exhaustive list of all the approaches is beyond the scope of the present work.

In all these efforts, the reported accuracy is between 80 and 90 %. In both soft computing and image shape based methods the candidate character, which happens to be a handwritten sample, is matched with template characters. However, extracted features of the characters should be local, stable and discriminative [10]. The problems with the methods described above are that any change in size, translation and rotation affect the feature extraction process. Horizontal and vertical lines might be longer than a standard character. swirls and loops might be tight or open. Besides, the slanting orientation of left and right handed writers can be different. These orientations result in variations of angles of the individual strokes of the characters. Calligraphy styles might be plain or ornate. Handwritten characters are affected by all of these changes. The presence of dakuten and handwritten characters in Japanese Katakana script, i.e. characters with the presence of double tick marks or small circles at the top right hand of the characters, respectively, poses additional problems. The STRICR-FB algorithm [5], based on Kohonen's Winner Take All (WTA) type of unsupervised learning addresses all these problems. The original method has been tested on the MS-Mincho font character set. The extension of the STRICR-FB algorithm proposed in this paper has been tested on several handwritten samples and a high accuracy has been obtained by including topological features of the individual characters of the character sets. These topological features included conjunction points as well as end points. Also, by dividing the image into four quadrants and finding the COG and Euclidean distances within the four quadrants enhanced the accuracy.

VI. OVERVIEW OF STRICR-FB

The original STRICR-FB algorithm is performed in two phases: (i) the construction of Character Unique Feature Vectors (CUFV) (ii) passing of these CUFVs

through a neural network for identification. After locating the Center of Gravity of each of the characters, the four CUFV, namely, mean variance, character density and decentricity have been calculated. In the second phase, an unsupervised clustering algorithm has been used, where clusters are built with real data.

V. PROPOSED ALGORITHM

The proposed algorithm is based on STRICR-FB [5] in the first phase, where the center of gravity (COG) is first determined. Then a normalization of the character image (bitonic) is performed. Subsequently, the extraction of conjunction points are carried out. The features of the candidate images are compared by template matching with the prototype/template images using an Euclidean distance measure defined below.

The entire character image is enclosed by a circle whose center and radius are determined from the image. This circle is subsequently divided into four quadrants. The CoG of the bitonic image of the character is computed using the number of pixels (black) and their locations (x,y). Once the COG is obtained, we try to locate the COG position among the four quadrants of the encircled character.

Due to presence of multiple stroke type and orders; there are characters with intersection points and non-intersection points. A conjunction point is a point in the image where multiple strokes intersect. The characters with multiple strokes and conjunction point(s) create a pattern of relationship between COG location and the number of conjunction points for each of the individual characters of the Hiragana script. Moreover, the stroke orders and curvature types are comparatively less complex than Kanji characters. Thus, a template matching scheme based on the distances of the conjunction points from the COGs in each of the quadrants provides a robust method.

Japanese characters are homeomorphic and so the computation of COG alone might not be sufficient to recognize individual characters. Furthermore, handwritten characters vary widely from person to person. Thus distinctive landmarks are required to impart a unique identity to the character. In particular, conjunction points, which mark the intersection of closed loops and also multiple strokes play an important role in identifying the topology of the character. Thus, after normalizing a handwritten character for size, shape and orientation and subsequent fitting into a circle, the conjunction points in each quadrant are located and COGs are measured for each character w.r.t to the conjunction points of that character. Euclidean distances of test and template images are calculated for these conjunction points. The outline of the process is given in Figure 2.

After pre-processing and normalization, the matching process involves (a) COG identification, (b) location of conjunction points and end points (c) measuring the Euclidean distance (d) comparison of template and target images and (e) check matching. These steps have been described as follows:

A. COG Identifications

The COG identification is followed by the calculation of locating i_{cog} and j_{cog} where i_{cog} represents COG of i th coordinate and j_{cog} as that of j th axis

$$i_{\text{cog}} = (\sum_{i=1}^m \sum_{j=1}^n i.C_{ij}) / (\sum_{i=1}^m \sum_{j=1}^n C_{ij}) \quad (1)$$

and

$$j_{\text{cog}} = (\sum_{i=1}^m \sum_{j=1}^n j.C_{ij}) / (\sum_{i=1}^m \sum_{j=1}^n C_{ij}) \quad (2)$$

B. Locating End points and Conjunction Point(s)

The conjunction points are located initially visually inspection.

The total number of conjunction points and the distance between those and COG vary from character to characters. For the characters which don't have conjunction points, end points were calculated. And then the coordinate of the conjunction and end points are banked up.

C. Measuring Euclidean Distance(s)

The Euclidean distances between COG and each of the conjunction points are calculated next (for the characters and intersection of their multiple strokes). The characters which don't have conjunction points, the Euclidean distances were measure this COG and end points. The calculation of Euclidean distance is measured as:

$$ed = \sqrt{[(x - i_{\text{cog}})^2 + (y - j_{\text{cog}})^2]} \quad (3)$$

Where 'ed' represents Euclidean distance and (x, y) is the coordinate of a conjunction point of the sample character in the image and $(i_{\text{cog}}, j_{\text{cog}})$ is the COG obtained.

D. Comparison between Template and Target Images

Once the COG and Euclidean distance is determined for the all points, the steps A, B and C is repeated for target image of the handwritten text also. And then the comparison of the numerical values between system generated template image and handwritten target image(s) are performed. If the character has two conjunction points, then the calculation will be followed for template is:

$$ed1_{ij} = \sqrt{[(x1 - i_{\text{cog}})^2 + (y1 - j_{\text{cog}})^2]} \quad (4)$$

// with cog and point 'a' $(x1, y1)$ in template

The Euclidean distances of the hand written character in target image (pictorial text) with COG and three conjunction points are:

$$ed2_{ij} = \sqrt{[(x4 - i_{\text{cog}})^2 + (y4 - j_{\text{cog}})^2]} \quad (5)$$

// with cog and point 'p' $(x4, y4)$ in target

If the deference of the average value of the Euclidean distance of template and target image tends to zero then the target image that is chosen to be matched with

template image is considered as similar.

E. Check Matching

Applying on the formula (7), given below, we can get the percentage of matching for target image to template image, followed by the average Euclidean distance of the collected conjunction points (6). The target character can be identified.

$$ed_{\text{avg}} = \sum_{i=1}^n ed_n / n \quad (6)$$

$$(ed_{\text{avg}}(\text{target}) / ed_{\text{avg}}(\text{template})) \times 100 = \% \text{ of match} \quad (7)$$

We determined average Euclidean distance based on total number of conjunction points 'n' for template and target images and then we found the amount of match in percentage of the target character with template character.

PREPROCESSING

1. Binarization
2. Resize the image
3. Rotation
4. Cropping
5. Error/noise removal using Gabor filters

LANDMARK POINT LOCATION

1. Skeletonization
2. Calculation of the total number of conjunction points
3. Storing the coordinates of conjunction points
4. Calculation of the total number of end points
5. Storing the coordinates of end points

COG COMPUTATION

1. Encircle character
2. Divide circle into four quadrants
3. Locate the COG
4. Locate the COG in quadrant (n)
5. Check for conjunction points with multiple strokes
6. Calculate Euclidean distance between COG and conjunction point
7. Get average Euclidean distance for all conjunction points
8. Get % match between template and target images
9. End

Both the system generated text (template image) and handwritten samples (target image) are pre-processed, normalized and the numerical statistics of their features are tallied by visually inspected.

In the result, it is observed that (A) the location of the COG of the encircled character object is not always at the center, but somewhere close to the center (according to the stroke density of the pixels present in the image) and (B) the location of the COG (in the quadrant of the encircled character) is multiple of the total number of the conjunction points obtained for that particular character.

The conjunction points are only obtained when two strokes or curvature of the strokes intersects between themselves. And hence the methodology is restricted to

the characters with multiple strokes and presence of intersection of the strokes.

The matching similarity of the characters having single stroke or multiple strokes with zero intersection point could be obtained from equation (7). According to the



Fig 5. Figure for Euclidean distance measurement and COG location.

Flow chart, the normalization (after preprocessing and segmentation) starts with identifying center of gravity of the character in the image[12-17]. This is initiated to check the density strength of the pixels of the character. Japanese characters are very complex especially when they are kanji scripts because of multiple strokes.

VI. RESULTS

Here a comparative study has been performed where the numeric values are compared with visual inspection. An analysis based on complete visual inspection is made. We applied the approaches on 45 Hiragana template images and six different handwritten samples of 45 Hiragana characters, individually. And then the samples were checked accordingly. An example of the calculations performed for the Hiragana character “あ” is given in Table II.

Table 1. Comparison Between Template And Target Image

Images	Template	Target
Size	(268, 258)	(242, 247)
COG	(133.5, 130.3)	(118.7, 126.8)
Ed1	(122, 188.3) 59.2	(112.7, 188.5) 62
Ed2	(104.9, 125.6) 28.8	(104.7, 136.8) 17.3
Ed3	(108.6, 74.9) 60	(109.7, 67.5) 59.9
Ed4	(167, 113) 37.1	(143.8, 146.3) 31.5
Average	46.275	42.675
%	(42.675 / 46.275).100 = 92.22	

It was observed that a match of 92.22% was obtained between the template and target images by applying the approach. The numeric values of differences were manually checked. Matching covers the groups of techniques based on similarity measures where the distance between the feature vectors, describing the extracted character and the description of each class is calculated. Different measures may be used, but the most

common is the Euclidean distance. Table II shows the measurement of six target samples of the handwritten Japanese Hiragana character “あ”.

Handwritten characters differ from person to person due to various ways to of writing, by various people. Collection of multiple handwritten samples proves COG location could be one of the factors for identifying the handwritten pictorial character sample. The location of the COG among the four quadrants, in an enclosed circle.

Table 2. Six Different Samples Of Hiragana “あ”

Char : あ	Ed1	Ed2	Ed3	Ed4	%
	66.7	24.0	65.4	26	98.4
	67.4	25.1	55.7	37.3	100
	52.1	25.7	59.2	26.6	88.4
	62.3	36.6	49	27.3	94.7
	60.8	21.4	50	32.2	88.8
	64.8	21.8	58.3	29.3	94.1

Distinguishes handwritten characters from sample to sample.

Table 3. Relationship Between Cog In Quadrant And Total Number Of Conjunction Points

Characters with multiple intersecting strokes				
Char.	Quadrant	Conj. Pt.	End pt.	Strokes
あ	4	4	5	3
き	2	2	6	3
む	2, 3	2	4	2
ふ	3	1	4	2
Characters with multiple non-Intersecting strokes				
Char.	Quadrant	Conj. Pt.	End pt.	Strokes
か	2	1, 2	6	3
こ	3	3	4	2
に	2	2, 3	6	3
た	2	2	8	4
Characters with single non-intersecting strokes				
Char.	Quadrant	Conj. Pt.	End pt.	Strokes
し	3	0	2	1
て	4	0	3	1
へ	3	0	3	1
ろ	3	0	4	1

Examples of characters with multiple intersecting and non-intersecting strokes as well as single non-intersecting strokes are given in Table III. For these characters the number of conjunction points, end points and strokes are given.

Our average recognition rate was 94.1% with the best three results corresponding to 100 %, 98.4 % and 95%. This is better than the result obtained in [5] using STRICR-FB for the MS Mincho set with a recognition rate of 90%.

VII. CONCLUSION

In this paper, we have proposed a semi-analytical approach based on physical and linguistic features of the Japanese Hiragana script and worked on six different handwritten samples for each of 45 hiragana characters for identification by template matching. Due to the nature of homeomorphism of Japanese characters, we have combined the original STRICR-FB algorithm for COG identification along with topological feature identification, namely, conjunction points and end points for handwritten characters which is an improvement over the result obtained using STRICR-FB for printed character sets. This is due to the fact that the inclusion of topological features in the identification process has added robustness to the technique.

In the future, the technique will be tested on Katakana and Kanji characters, both printed and handwritten.

REFERENCES

- [1] Indira B., Shalini M., Ramana Murthy M.V., Mahaboob Sharief Shaik, Classification and Recognition of printed Hindi characters using Artificial Neural Networks, *Int. J. Image, Graphics, Signal Processing*, 4(6), 15-21 (2012).
- [2] Dewangan Shailendra Kumar, Real Time Recognition of Handwritten Devanagiri Signatures without segmentation using Artificial Neural Networks, *Int. J. Image, Graphics, Signal Processing*, 5(4), 30-37 (2013).
- [3] John Jomy, Balakrishnan Kannan, Pramode K.V., A system for Offline Recognition of Handwritten characters in Malayam Script, *nt. J. Image, Graphics, Signal Processing*, 5(4), 53-59 (2013).
- [4] Das S and Banerjee S (2014), Survey of Pattern Recognition Approaches in Japanese Character Recognition, *International Journal of Computer Science and Information Technology*, Vol. 5(1) 93-99.
- [5] D. Barnes and M. Manic, STRICR-FB a Novel SIze-Translation Rotation Invariant Character Recognition Method (2010), *Proceed. 6th Human System Interaction Conference*, Rzeszow, Poland, 163-168
- [6] Miia Sainio, Kazuo Bingushi, Raymond Bertram; *Vision Research* (2007); "The role of interword spacing in reading Japanese: An eye movement study"; Volume: 47, Issue: 20, Pages: 2577-2586
- [7] Neural Networks which consist of Simple Recurrent Networks for Character Recognition by Template Matching, (2008) Y. Shimozawa, *Journal of the Information Processing Society* 49(10), 3703-3714.
- [8] The post processing of Character Recognition by Genetic Algorithms, Y Shimozawa, S. Okoma, (1999), *Journal of the Information Processing Society*, 40(3) 1106-1116.
- [9] Off-line Hand Printed Character Recognition System using directional HMM based on features of connected pixels, H. Nishimura, M. Tsutsumi, M. Maruyama, H. Miyao and Y. Nakano (2002), *Journal of the Information Processing Society* 43(12) 4051-4058.
- [10] Maruyama, K.-I.; Maruyama, M.; Miyao, H.; Nakano, Y.; "Hand printed Hiragana recognition using support vector machines"; *Frontiers in Handwriting Recognition*, 2002. *Proceedings. Eighth International Workshop on Digital Object Identifier: 10.1109/IWFHR.2002.1030884*; 2002, Pages: 55 – 60.
- [11] Proposal of Character Segmentation using Convergence of the Shortest Path, E. Tan aka (2011) *Meeting on Image Recognition and Understanding Proceedings* (2011) MIRU 2011, 331-336.
- [12] An Online Character Recognition Algorithm RAV (Parametrized Angle Variations) M. Lubumbashi, S. Masai, O. Minamoto, Y. Nagasaki, Y. Kommunizma and T. Maturation, (2000), *Journal of the Information Processing Society*, 41(9), 2536-2544.
- [13] A Handwriting Character Recognition System Using Adaptive Context Processing, N. Okayama, (1999) *IPSIJ Technical Report on Information and Media (IM)* 85-90.
- [14] An Efficient Indexing Scheme for Image Storage and Recognition, *IEEE Transactions on Industrial Electronics*, M. Al Mohamed, (1999) 46(2).
- [15] S. H. Kim, "Performance Improvement Strategies on Template Matching for Large Set Character Recognition", *Proc. 17th International Conference on Computer Processing of Oriental Languages*, pp 250-253, Hongkong, April 1997.
- [16] C. H. Ting, H. J. Lee, Y. J. Thai, "Multi stage per-candidate selection handwritten Chinese character recognition systems", *Pattern Recognition*, vol. 27, no.8, pp.1093-1102, 1994."
- [17] H. Tallahassee and T.D. Griffin, "Recognition enhancement by linear tournament verification", *Prof. 2nd ICDAR*, Tsunami, Japan, 1993, pp.585-588.
- [18] D.H. Kim, Y.S. Twang, S.T. Park, E.J. Kim, S.H. Park and S.Y. Bang, "Handwritten Korean character image database PE92", *Prof. 2nd ICDAR*, Tsunami, Japan, 1993, pp.470-473.
- [19] J. J. Hull, Document image matching and retrieval with multiple distortion--invariant descriptors, *Prof. IAPR Workshop on Document Analysis Systems*, pp. 383-399 (1994).
- [20] Snead, M. Minos and K. locked, Document image retrieval system using character candidates generated by character recognition process, *Proc. Second Int. Cong. Document Analysis and Recognition*, pp. 541- 546 (1993).
- [21] N. Nina, K. Kagoshima and Y. Shimmer, Post processing for character recognition using keyword information, *IAPR Workshop Machine Vision Applications*, pp. 519-522 (1992).
- [22] D.A. Dahl, L. M. Norton and S. L. Taylor, Improving OCR accuracy with linguistic knowledge, *Proc. Second Ann. Symp. Document Analysis and Information Retrieval*, pp. 169-177 (1993).

Authors' Profiles



Soumendu Das, male, was a student of West Bengal University of Technology, Kolkata, India, and he is currently employed at Infosys, India. His research interests are image processing and OCR in Japanese



Sreeparna Banerjee, female, associate professor of West Bengal University of Technology, India. Her research interests include multimedia, digital image processing.

How to cite this paper: Soumendu Das, Sreeparna Banerjee, "An Algorithm for Japanese Character Recognition", IJIGSP, vol.7, no.1, pp.9-15, 2015. DOI: 10.5815/ijigsp.2015.01.02