International Journal of Ocean System Engineering 2(1) (2012) 16-23
http://dx.doi.org/10.5574/IJOSE.2012.2.1.016

Combining Different Distance Measurements Methods with Dempster-Shafer-Theory for Recognition of Urdu Character Script

Yunus Khan [1*], Chetan Nagar [2] and Devendra S Kaushal [3]

[1]*M.E Researcher JIT Borawan Khargone M.P. India*

[2]*Professor & Head Department of CSE JIT Borawan M.P. India*

[3]*M.Tech. Researcher SATI Vidisha M.P. India*

(Manuscript Received December 12, 2011; Revised January 9, 2012; Accepted February 24, 2012)

Abstract

In this paper we discussed a new methodology for Urdu Character Recognition system using Dempster-Shafer theory which can powerfully estimate the similarity ratings between a recognized character and sampling characters in the character database. Recognition of character is done by five probability calculation methods such as (similarity, hamming, linear correlation, cross-correlation, nearest neighbor) with Dempster-Shafer theory of belief functions. The main objective of this paper is to Recognition of Urdu letters and numerals through five similarity and dissimilarity algorithms to find the similarity between the given image and the standard template in the character recognition system. In this paper we develop a method to combine the results of the different distance measurement methods using the Dempster-Shafer theory .This idea enables us to obtain a single precision result. It was observed that the combination of these results ultimately enhanced the successrate.

Keywords: Dempster-Shafer Theory, Patten Recognition, Feature Extraction, Linear Correlation, Cross-Correlation, Nearest Neighbor, Hamming.

1. Introduction

We know that the character recognition is the process of identifying the optically scanned files to computer editable format. Urdu Character recognition techniques give a specific symbolic identity to an offline printed or written image of a Urdu character. Here, we tend to replicate human functions by machine involving the recognition process. Character recognition is better known as optical character recognition because it deals with the recognition of optically processed characters rather than magnetically processed ones. The main objective of character recognition is to interpret input as a sequence of characters from an already existing set of characters. Character recognition is one of the most fundamental topics in the context of pattern recognition and is included as a key issue in the recognition of hand written characters and digits [2]. Hand-written character recognition can be divided into two categories, namely online handwritten character recognition and offline handwritten character recognition [1].Online character recognition involves the identification of characters while they are being written and offline character recognition involves the recognition of already written character patterns in a scanned digital image [3]. Character Recognition usually is a mechanical or electronic transition of images of hand written or printed text into machine editable text. Character Recognition is an important area of research in pattern recognition, artificial intelligence and machine vision. The advantages of the character recognition process are that it can save both time and

*Corresponding author. Tel.: +91-7285-277862, Fax.: +91-7285-277710.
E-mail address: callyunuskhan@gmail.com

effort when developing a digital replica of the document. It provides a fast and reliable alternative to typing manually. The Character Recognition program software can convert a document in several electronic formats, like Microsoft Word, Text (and Rich text), Excel, and PDF formats. All documents created through program software are editable and allow you to modify the content as you see fit. Most character recognition procedures can be visualized as consisting of three steps which use: the pre-processor, feature extractor and recognizer [6, 7]. The following are some of the applications of character recognition

1) Signature Verification
2) Writer Identification
3) In Examination assessment as a Mark Sheet Reader, etc.

An Optical Character Recognition System is software engineered to convert hand-written or typewritten text (usually scanned) documents into machine editable text formats. The proposed method avoids feature extraction as it directly compares the test Urdu character with the template.

The paper is organized in seven sections. Section II outlines the features of Urdu script. Section III discusses about the related work, problem statement is presented in Section IV and Section V discusses proposed system. Section VI discusses about the technique used, Section VII presents the Experimental analysis and the paper concludes with future scope in Section VIII.

2. Urdu and Its Script

The Urdu alphabet is the right-to-left alphabet used for the Urdu language. It is a modification of the Persian alphabet, which is itself a derivative of the Arabic alphabet. With 38 letters, the Urdu alphabet is typically written in the calligraphic Nasta'liq script, whereas Arabic is more commonly in the Naskh style. Usually, bare transliterations of Urdū into Roman letters omit many phonemic elements that have no equivalent in English or other languages commonly written in the Roman alphabet.[citation needed] The National Language Authority of Pakistan has developed a number of systems with specific notations to signify non-English sounds, but these can only be properly read by someone already familiar with Urdū, Persian, or Arabic for letters such as ژ ص ط غ خ or ق and Hindi for letters such as ڑ.

A list of the letters of the Urdū alphabet and their pronunciation is given below. Urdū contains many historical spellings from Arabic and Persian, and therefore has many irregularities. The Arabic letters yaa and haa both have two variants in Urdū: one of the yaa variants is used at the ends of words for the sound, and one of the haa variants is used to indicate the aspirated consonants. The retroflex consonants needed to be added as well; this was accomplished by placing a superscript ط (to'e) above the corresponding dental consonants. Several letters which represent distinct consonants in Arabic are conflated in Persian, and this has carried over to Urdū. Some of the original Arabic letters are not used in Urdu. This is the list of the Urdu letters, giving the consonant pronunciation. Many of these letters also represent vowel sounds.

Fig. 1. Vowels
*

** Vowels in Urdu are represented by letters that are also considered consonants. Many vowel sounds can be represented by one letter. Confusion can arise, but context is usually enough to figure out the correct sound. This is a list of Urdu vowels found in the initial, medial, and final positions.*

Romanization	Pronunciation	Final	Medial	Initial
a	/ə/	◌ٰ	◌ٰ	◌ٰ
ā	/ɑ:/	◌ٰ	◌ٰ	◌ٰ
i	/ɪ/	◌ٖ	◌ٖ	◌ٖ
ī	/i:/	◌ٖ	◌ٖ	◌ٖ
u	/ʊ/	◌ٗ	◌ٗ	◌ٗ
ū	/u:/	◌ٗ	◌ٗ	◌ٗ
e	/e:/	◌ے	◌ٖ	◌ٖ
ai	/ə/	◌ے	◌ٖ	◌ٖ
o	/o:/	◌ٗ	◌ٗ	◌ٗ
au	/ɔ/	◌ٗ	◌ٗ	◌ٗ

Fig. 2. Short vowels

Short vowels ("a", "i", "u") are represented by marks above and below a consonant.

Vowel	Name	Transcription	IPA
◌َ	zabar	ba	/ə/
◌ِ	zer	bi	/ɪ/
◌ُ	pesh	bu	/ʊ/

Alif

Alif (□) is the first letter of the Urdu alphabet, and it is used exclusively as a vowel. At the beginning of a word, alif can be used to represent any of the short vowels, e.g. □□ ab, □□□ ism, □□□□ urdū. Also at the beginning, an alif (□) followed by either wā'o (□) or ye (□) represents a long vowel sound. However, wā'o (□) or ye (□) alone at the beginning represents a consonant.Alif also has a variant, call alif madd (□). It is used to represent a long "ā" at the beginning of a word, e.g. □□ āp, □□□□ ādmi. At the middle or end of a word, long ā is represented simply by alif (□), e.g. □□□ bāt, □□□□ ārām.

Wao

Wā'o is used to render the vowels "ū", "o", and "au". It also renders the consonant "w", but many people get confused between (W and V)sounds. (Wā'o) sound is not closer to the consonant (V) sound because V has vibrating sound. Many Urdu linguists believe that there is no (V) sound in Urdu.

Ye

Ye is divided into two variants: choṭī ye and baṛi ye. Choṭī ye (□) is written in all forms exactly as in Persian. It is used for the long vowel "ī" and the consonant "y". Baṛī ye (□) is used to render the vowels "e" and "ai" (/e:/ and /æ :/ respectively). Baṛī ye is distinguished in writing from choṭī ye only when it comes at the end of a word.

Use of specific letters

Retroflex letters: Retroflex consonants were not present in the Persian alphabet, and therefore had to be created specifically for Urdū. This was accomplished by placing a superscript □ (to'e) above the corresponding dental consonants.
Short vowels ("a", "i", "u") are represented by marks above and below a consonant.

Letter	Name	IPA
ٹ	ṭe	[ʈ]
ڈ	ḍāl	[ɖ]
ڑ	aṛ	[ɽ]

Do Chashmi he

The letter do chashmī he (□) is used in native Hindustānī words, for aspiration of certain consonants. The aspirated consonants are sometimes classified as separate letters, although it takes two characters to represent them.

بھا	bhā [bʱɑ:]
پھا	phā [pʱɑ:]
تھا	thā [tʱɑ:]
ٹھا	ṭhā [ʈʱɑ:]
جھا	jhā [d͡ʒʱɑ:]
چھا	chā [t͡ʃʱɑ:]
دھا	dhā [dʱɑ:]
ڈھا	ḍhā [ɖʱɑ:]
ڑھا	ṛhā [ɽʱɑ:]
کھا	khā [kʱɑ:]
گھا	ghā [gʱɑ:]

Fig. 3. Transcription IPA

3. Related Works

Research in Urdu script recognition has started nearly three decades ago, but it is only recently that it has gained popularity. Based online recognition system for Urdu symbols is very challenging to current researchers. Researchers have many different approaches for both segmentation and recognition tasks of word recognition such as ligature and world probabilities, artificial neural network, HMM and optical detection. The Latest concepts is the Damster-Shafer theory of belief functions and sketch a brief history of its conceptual development [15].It is fully dependent on theory of functions and conceptual development of optical character recognition system An overview of the classic works has been examined to establish a body of knowledge on belief functions, transforming the theory into a computational tool for evidential reasoning in artificial intelligence, opened up new avenues for applications, and became authoritative resources for anyone who is interested in gaining further insight into and understanding of belief functions [11, 14].Recogination of urdu character using Support vector machine are also provide a knowledge about how to detect scanned files into computer editable data. These days extract oriented features of a

handwritten character are extracted and then these features are applied to Dempster-Shafer theory which can powerfully estimate the similarity ratings between a recognized character and sampling characters in the character database.

4. Problem Statement

The functionality behind the Urdu Character Recogination system is to compute effort using Damster-Shafer theory after calculating the distances from different distance measurements and to develop a character recognizer with the help of this theory.

5. Proposed System

The Moto of this paper is to recognize offline handwritten Urdu characters and numerals using five different methods. The results from the methods are combined using the Dempster-Shafer method to arrive at a degree of belief, in other words we aim to arrive at a single precision result.

The steps involved in this process are as follows:
1. Initially a database of prototypes of handwritten Urdu characters is created
2. The probability of identifying the given input as a particular Urdu character is obtained with the use distance measurement methods
3. The results obtained are combined using the Dempster-Shafer theory
4. From the resulting single precision result the input character is identified by the use of the calculations based on the least possible error.

The design primarily involved in our offline handwritten Urdu character recognition system is outlined in the following block diagram

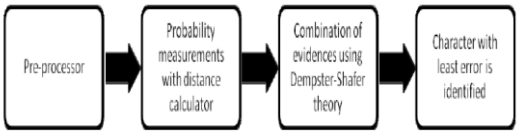

Fig. 4. Block Diagram of offline handwritten Urdu character recognition system.

- The pre-processor is in general used to prepare the raw material for recognizing the input Urdu character image. In this particular method we used the method of size normalization.

- The probabilities involved in the procedure are calculated by the use of five different methods.
  1. Similarity Based Methods
  2. Hamming Method
  3. Linear-Correlation Method
  4. Cross-Correlation Method
  5. Nearest Neighbor Method

- In the subsequent step the evidences obtained are combined using the Dempster-shafer theory to obtain a final single precision value.

- In the final step the Urdu character is identified based on the least possible error.

6. Methodology

The current methodology is divided into three phases first Pre-Processing is applied on a raw set of data, subsequently process for the similarity or the dissimilarity between the input Urdu character (Test case) and the prototypes from the existing database, and finally similarity measures are combined with Damster-Shafer theory which can powerfully estimate the similarity ratings between a recognized character and sampling characters in the character database.

6.1 Pre-Processing

The pre-processor prepares a raw input image of recognizing the given Urdu character. Here, we have normalized the image to a size of 10x10. The purpose of size normalization is to make the size of the input character image equal to the size of the existing prototype image.

6.2 Distance-Measurements

The proposed model measures similarity or the dissimilarity between the unlabelled input character (the target) and the labeled prototypes from the existing database. The unlabelled target is recognized, identified and present with labeled based on the error calculations. The methods employed here are given in detail subsequently.

6.2.1 1 Similarity Function

The similarity function S(Y, X) was used to identify the cells occupied by both the models Y and X (the target and the prototype respectively) using the following formula.

S(Y, X) = Σmi=1Σnj=1Yij.AND.Xij

Where
Yij.AND.Xij = 1, if the ijth cell is occupied by both models Y and X
= 0, otherwise

The smaller the value of S(Y, X) obtained, lesser is the area shared commonly by both the models and hence lesser the similarity between the two. The largest value among those calculated for the entire population is chosen.

6.2.2 2 Hamming Distance

The Hamming distance function H(Y, X) was used to measure the number of different cells occupied by the two models Y and X. in other words it measures the number of positions or cells the two models differ in. it is give by the following formula.

H(Y, X) = Σm i=1Σn j=1Yij.XOR.Xij

Where,
Yij.XOR.Xij = 1, if the jth cell is occupied by one model and not by the others
= 0, if otherwise

The larger the value of H(Y, X) the greater is the difference between the target and the prototype. Thus unlike the Similarity method, here the smallest value in the entire population is chosen.

6.2.3 3 Linear Correlation

The linear correlation function was obtained by the modification of the similarity function S(Y,X) keeping in mind the various degrees of misalignment and stroke width variation between the target and the prototype.

LC(Y, X) = 2*[S(Y, X) / (Ny+Nx)]

Smaller the value of LC(Y, X) smaller is the common area shared by the two models. Thus, the largest value over
the entire population is preferred

6.2.4 4 Cross-Correlation

Similar to auto-correlation, similarity function S(Y, X) was modified to obtain cross- correlation. It is given by the Following formula.

CC(Y, X) = [S(Y, X)]2 / (Ny*Nx)

Where Ny and Nx are cells occupied by models Y and X respectively Smaller the value of CC(Y, X) smaller is the normalized common area shared by the target and the prototype. Thus the largest value from among the entire population is chosen.

6.2.4 5 Nearest Neighbor

The "nearest neighbor cell distance d(Yij ,X) was used to measure the distance between ijth cell of model Y to the nearest cell of model X. For any pair of models Y and X two measurements were used to indicate the difference between the pair. They are given by the following two equations.

Nearest neighbor 1

ND1(Y, X)=1/Ny Σmi=1Σnj
=1[d(Yij ,X)]1/2+1/NxΣmi=1Σnj
=1[d(Xij ,Y)]1/2

Nearest neighbor 2

ND2(Y,X)
=[(Σmi=1Σnj=1d(Yij,X)/Ny)
+(Σmi=1Σnj=1d(Xij ,Y)/Nx)]1/2

Where, Ny and Nx are numbers of cells occupied by the two models Y and X respectively.
Larger the values by both nearest neighbor 1 and 2 greater is the cell difference between the two models X and Y. Thus, the smallest value from among the entire population is chosen.

6.3 Normalisation

Normalization is a process to put distances or similarity scores as a performance index into a range of 0 and 1. If the similarity or the dissimilarity index is in the range of [dmin, dmax] and is not in the range of 0 and 1, then

$$\delta = (d - dmin) / (dmax - dmin)$$

Where d is the similarity or the dissimilarity and δ is the normalized similarity / dissimilarity.

6.4 Dempster-Shafer Theory

The Dempster-Shafer theory, also known as the theory of belief functions, is a generalization of the Bayesian theory of subjective probability [13, 15]. Whereas the Bayesian theory requires probabilities for each question of interest, belief functions allow us to base degrees of belief for one question on probabilities for a related question. These degrees of belief may or may not have the mathematical properties of probabilities; how much they differ from probabilities will depend on how closely the two questions are related. Dempster-Shafer degrees of belief resemble the certainty factors and this resemblance suggested that they might combine the rigor of probability theory with the flexibility of rule-based systems [9]. Subsequent work has made clear that the management of uncertainty inherently requires more structure than is available in simple rule based systems, but the Dempster-Shafer theory remains attractive because of its relative flexibility [10].

The Dempster-Shafer theory is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence. Implementing the Dempster-Shafer theory in a specific problem generally involves solving two related problems. First, we must sort the uncertainties in the problem into a priori independent items of evidence [12, 14]. Second, we must carry out Dempster's rule computationally. These two problems and their solutions are closely related. Sorting the uncertainties into independent items leads to a structure involving items of evidence that bear on different but related questions, and this structure can be used to make computations feasible.

Dempster-Shafer theory is a mathematical theory for combining the evidences obtained from different sources and evaluating the conflict between them. The purpose of aggregating such information is to meaningfully summarize and simplify a corpus of data. The Dempster-Shafer theory is primarily based on the assumption that each of those multiple sources from which results have been obtained is independent of the others.

If m1(A) an m2(A) are the results evidences from two independent measurements then the combined result(evidence) is given by

$$\{m1(A) * m2(A)\} / (1-k)$$

Where, k is the normalization factor which varies from 0 to 1.

7. Experimental Analysis

Templates are created for numerals and subset of Telugu characters. Distances are measured for two test cases and one for numerals and one for Telugu characters. The results of distance measurements after applying normalization and Dempster-Shafer theory are shown in tables below. The error calculations are represented in the graph. It is found that test case has minimum errors.

Table 1. Table Showing values after Normalization and Applying D-S Theory with test case as

Digits	S	H	LC	CC	NN1	NN2	Probability	Error
٠	0.131991	0.107623	0.122224	0.123028	0.093400	0.070935	0.003644	0.9927
١	0.091723	0.105381	0.097017	0.078816	0.083062	0.073383	0.001160	0.9976
٢	0.031320	0.000000	0.030333	0.018094	0.074066	0.088515	0.000000	1
٣	0.087248	0.096413	0.091921	0.073141	0.082199	0.067222	0.000805	0.9983
٤	0.210291	0.385650	0.250947	0.314382	0.226377	0.262529	0.979190	0.00043
٥	0.073826	0.127803	0.089294	0.063463	0.136069	0.155401	0.002911	0.9941
٦	0.149888	0.085202	0.126454	0.139706	0.121735	0.120812	0.008545	0.9829
٧	0.000000	0.002242	0.000000	0.000000	0.000000	0.000000	0.000000	1
٨	0.145414	0.080717	0.123172	0.133750	0.097155	0.076361	0.003694	0.9926
٩	0.078300	0.008969	0.068638	0.055621	0.085938	0.084843	0.000050	0.9999

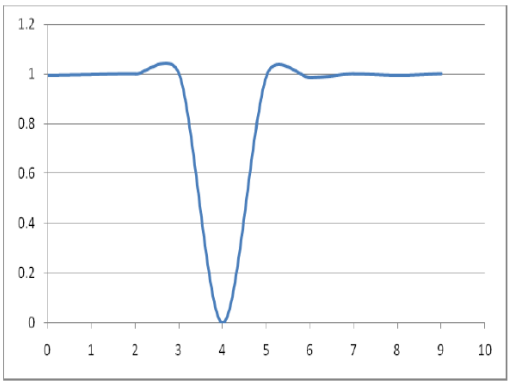

Fig. 5. Graph for Numeral Vs Error using D-S Theory with test case as shown above

Table 2. Table Showing values after Normalization and Applying D-S Theory with test case as

Distances for Number Comparisons	S	H	LCC	LC
ب	0.307692	0.120690	0.228671	0.268487
چ	0.153846	0.339080	0.287723	0.215018
ش	0.000000	0.103448	0.000000	0.000000
ل	0.076923	0.000000	0.005278	0.032858
ق	0.384615	0.189655	0.329699	0.381546
م	0.076923	0.247126	0.148629	0.102091
Distances for Number Comparisons	**NN1**	**NN2**	**Probability**	**Error**
ب	0.276669	0.222415	0.118308	0.777381
چ	0.000000	0.000000	0.000000	1
ش	0.098897	0.163789	0.000000	1
ل	0.032778	0.147137	0.000000	1
ق	0.400328	0.281850	0.873091	0.016106
م	0.191329	0.184810	0.008601	0.982872

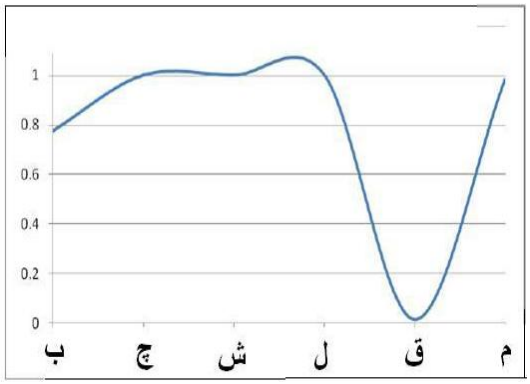

Fig. 6. Graph for Character Vs Error using D-S Theory with test case as shown above

8. Conclusion and Future Scope

Recognition of Urdu letters and numerals was done and various experiments were performed. Five similarity and dissimilarity algorithms to find the similarity between the given image and the standard template in the recognition system are implemented. Once the test case is identified it is displayed. This method becomes extensive as the no. of templates increases it is required to calculate the distances from all the templates. Another disadvantage is with characters of type 3 and 8 which may get probability values which are very near. In such cases structural verifier may be used which improves the recognition rate. This work has a future scope, it can be further implemented to feed the segmented characters to the recognition system and then recognize the characters. The segmentation algorithm can further be improved for high rate of efficiency. In future this strategy is also implemented for online Urdu character recognition System.

References

[1] V.Jagadeeshbabu.L.Prasanth, R.Raghunath, Sharma, Prabhakar Rao and G.V.Bharath, HMM –based online recognition system for telugu symbols, proceedings of ninth international conference on document analysis and recognition , (2007).

[2] M.S.Rao, Gowrishankar andV.S Chakravarthy, Online recognition of hand written Telugu characters, proceedings of the international conference on Universal knowledge, (2002).

[3] Hariharan Swethalakshmi,Anitha Jayaraman, V.Srinivasa Chakravarthy and C.Chandra Sekhar, On line character recognition of Devanagari.

[4] C.VLakshmi. Patvardahan and C,Mohit Prasad, A novel approach for improving recognition accuracies in OCR of printed telugu text, proceedings of international conference on signal processing and communications, (2004).

[5] B.BChaudhuri,O.AKumar. and K.VRamana, Automatic generation and recognition of Telugu script characters, Journal of IETE, (37) (1991) 499-511.

[6] S.Impedevo,L.Ottaviano and S.Occhingro, Optical Character Recognition — A Survey, International Journal of Pattern Recognition and Artificial Intelligence, (1991).

[7] J.Mantas,An overview of character recognition methodologies, Pattern Recognition, 6 (19) (1986) 425-430.

[8] S.N.S.Rajasekharan and B.L.Deekshatulu. Generation and Recognition of printed Telugu characters, Computer graphics and image processing, (6) (1977) 335-360.

[9] A.P.Dempster,A generalization of Bayesian inference, Journal of the Royal Statistical Society, B (30) 205-247.

[10] Shafer, Glenn, A Mathematical Theory of Evidence, Princeton University Press, (1976).

[11] Shafer, Glenn, Perspectives on the theory and practice of belief functions. International Journal of Approximate Reasoning, 1-40.

[12] Shafer, Glenn, and Judea Pearl, Readings in Uncertain Reasoning. Morgan Kaufmann.

[13] Liping Liu and R. Ronald Yager, Classic Works of the Dempster-Shafer Theory of Belief Functions: An Introduction, (219) (2008).

[14] Yager, Mario Fedrizzi, R.Ronald and Janusz, Kacprzyk. Advances in the Dempster-Shafer Theory of Evidence, Wiley, (1994).

[15] Special Issue on Dempster-Shafer Theory, Methodology, and Applications. International Journal of Approximate Reasoning, (31) (2002) 1-2.