ISSPA07 Proceedings
Al-Sharja, UAE, 15-17 Feb. 2007

A FUZZY EXPERT SYSTEM FOR RECOGNITION OF HANDWRITTEN ARABIC SUB-WORDS

ABSTRACT

A fuzzy expert system for selected Arabic sub-words recognition is presented in this paper. For each sub-word pattern, membership values are determined for a number of fuzzy sets defined on the features extracted from the pattern. These sub-words consist of two characters and are written cursively, so, the first step is to segment the sub-words into two objects, main and secondary objects, keeping the two characters connected and making use of the general and special features in the recognition process. Presence or absence of dots in a sub-word and the number of such dots are fuzzy features, since the dot(s) may not appear exactly above or below the related character and they may appear mixed together. Closed loop is a fuzzy feature also. The proposed expert system consists of two main parts. First, the preprocessing part includes the feature extraction step that provides sufficient information to the inference engine. Second, the inference engine which applies the suitable set of fuzzy rules and aggregates them towards the final decision

Keywords: Arabic ocr, fuzzy logic, fuzzy expert system, handwritten ocr.

1. INTRODUCTION

In the real world, we are primarily concerned with necessity, a measure of the extent to which the data support a conclusion. The reasoning process of establishing necessary conclusions is not the same as the process of establishing possible conclusions.

Expert systems which deals with this necessity are programs, designed to make available some of the skills of an expert to non-experts. One of the earliest methods employs rule-based systems, which use "If…Then …"rules to represent the expert's reasoning process. Other approaches include semantic or associative nets, frames, and neural nets, currently very popular in a wide variety of fields.

While fuzzy set theory has been used to model relevant features; fuzzy logic has been used to perform the decision making. Thus fuzzy set theory permits to compress the knowledge base and fuzzy logic permits to reduce complexity of the system by using partial and uncertain information.

Arabic language is always written cursive. Due to the possible connectivity of some characters from either sides and connectivity for others from one side, the Arabic word consists of one or more sub-words. Each sub-word consists of one character or more. There are four forms for the characters, namely: stand-alone, initial, final and middle forms. The sub-word consisting of one character is in its stand-alone form. The first character of a sub-word consisting of two characters should be in its initial form and the second in its final form. When the sub-word consists of more than two characters, the first is in its initial form, the last in its final form and the rest in their middle forms. A novel approach for the Arabic character recognition based on statistical analysis of a typical Arabic text has been presented to indicate the importance of sub-word recognition rather than word in Arabic OCR systems.[2] An estimate of the number of characters in the sub-words showed that about 45% of Arabic sub-words consist of one character and about 28% of the Arabic sub-words consist of two characters. This means that recognizing stand-alone characters carries the most important target and recognizing the two-character sub-words is of a second importance. This paper concentrates the effort on the recognition of the sub-words consisting of two characters, since the recognition of stand-alone characters had received attention in the first place.[3]

The data used in this paper were obtained from 48 different writers and transformed manually into a matrix of characters or sub-words so as to be processed for character recognition by various methods.[4] These data have been put in an information vector to serve as an input to the system. The final output is the final decision

Since people have different ways of writing each character, there may be variations in size, thickness and style among samples of the same character hence normalization becomes a necessity.

2. FUZZY EXPERT SYSTEM

A fuzzy expert system is an expert system that uses a collection of fuzzy membership functions and rules, instead of Boolean logic, to reason about data[1].

The rules in a fuzzy expert system are usually of a form similar to the following:

if x is low and y is high then z = medium

where x and y are input variables , z is an output variable, low, high, and medium are fuzzy variables defined by membership functions (fuzzy subsets) defined on x, y, z respectively. The antecedent (the rule's premise) describes to what degree the rule applies, while the conclusion (the rule's consequent) assigns a membership function to each of one or more output variables. The set of rules in a fuzzy expert system is known as the rule base or knowledge base.

3. INFERENCE IN A FUZZY EXPERT SYSTEM

Inference in an expert system involves the modification of data, either its value or its truth value or both, by rules.[1] Whether modification of the value of a datum is permitted or not depends on its existing truth value, and the type of inference being used. The general inference process proceeds in four steps:

3.1 Under FUZZIFICATION, the membership functions defined on the input variables are applied to their actual values, to determine the degree of truth for each rule premise.

3.2 Under INFERENCE, the truth value for the premise of each rule is computed, and applied to the conclusion part of each rule. Those results in one fuzzy subset to be assigned to each output variable for each rule. Usually only MIN's or PRODUCT's are used as inference rules. In MIN inference (which is used in the present research), the output membership function is clipped off at a height corresponding to the rule premise's computed degree of truth (fuzzy logic AND).

3.3 Under COMPOSITION, all of the fuzzy subsets assigned to each output variable are combined together to form a single fuzzy subset for each output variable. Again, usually MAX or SUM are used. In MAX composition (which is used in the present research), the combined output fuzzy subset is constructed by taking the point wise maximum over all of the fuzzy subsets assigned to variable by the inference rule (fuzzy logic OR).

3.4 Finally is the (optional) DEFUZZIFICATION, which is used when it is useful to convert the fuzzy output set to a crisp number. Two of the more common techniques of defuzzification are the MAXIMUM and CENTROID methods. In the MAXIMUM method (which is used in the present research), one of the variable values at which the fuzzy subset has its maximum truth value is chosen as the crisp value for the output variable.

4. SUB-WORD RECOGNITION SYSTEM

The implemented sub-word recognition system, involve 4 image-processing stages: Preprocessing, Segmentation (of objects), Feature Extraction and Recognition.

4.1 Preprocessing

This phase aims to produce a cleaned up version of the original image, so that it can be used directly and efficiently. This is accomplished by splitting, binarization, smoothing, and thinning, in order to enable a reliable feature extraction.

Figure-1: Sample of the input image

Figure 1 shows a sample of the input image of 48 two-character sub-words written by 48 different persons [4]. Figure 2 shows an enlarged single sub-word image. Thinning is the process of minimizing the width of a line, in an image, from many pixels wide to just one pixel as shown in Figure 3.

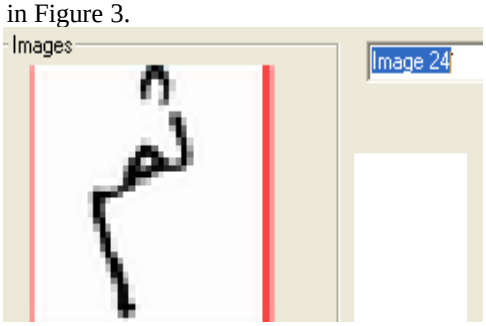


Figure 2: Single Sub-word Image

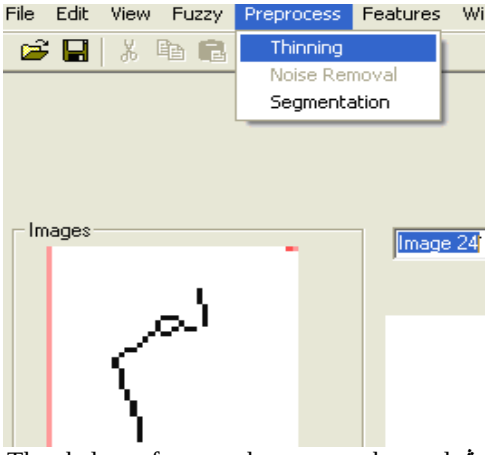


Figure 3 : The skeleton for two characters sub-word ثم after noise removal and thinning

The present research uses two types of thinning algorithm (Sequential and Parallel) in order to get a fast thinning. The result of thinning procedure is shown in Figure-3 for the sub-word ثم without the dots.

4.2 Objects Segmentation

Objects segmentation is the process of dividing an image into objects. An object may be a character, a word, a sub-word or a dot(s) or a special character. Any error made in segmentation will affect the over all recognition result. So, segmentation is a crucial step in the optical character recognition process.

After the preprocessing stage, the majority of character recognition systems perform breaking up operation, on the character to be recognized, and end up with individual objects. Such operation is also called segmentation.

The segmentation algorithm used in the proposed system depends on coloring. This algorithm used in Filtering, Dots Count, and Dots Position allocation. It depends on changing the color of an object into two or three colors, one color for the main part of sub-word and other color(s) for the secondary objects which are mainly dots above or (and) below the main object. By counting the number of pixels of each color, the position and count of dots may be found using fuzzy logic.[5]

4.3 Feature Extraction

Once the spot containing the secondary character is separated, feature extraction stage can be applied to every spot (secondary or primary) according to their belonging to one of the secondary characters or the main body of the sub-word.

4.3.1 Features of secondary character

The number of pixels comprising the dots is so little that they are easy to confuse with each other. It is difficult to find good algorithms to distinguish between them in a clear cut way. This is because there is a limited number of features to be used for secondary characters recognition, namely width, height, height to width ratio or area. Some other global features shall be given below.

4.3.2 Features of main object of sub-word

In order to distinguish between various sub-words - without segmenting the stroke into its constituent characters - features can be subdivided into ***general features*** which are used to distinguish between main strokes of the sub-words. It consists of the following features:

{ Width, Height, area }

See general features dimensions in Table-1.

Other features are called ***special features***, that are used to distinguish between the two characters comprising the sub-word. Those features are:

TABLE 1 GENERAL FEATURES DIMENSIONS

	Two Caracters Sub-Word					
	لر	ما	مع	مل	نا	ثم
Max.Height	28	26	28	34	27	38
Min. Height	15	12	9	16	13	23
Average Height	20	18.2	16.3	25.7	19.1	30
Max. Width	38	37	36	34	35	36
Min. Width	27	27	27	16	24	26
Average Width	32.5	30.7	31.4	257	29.2	30.4
Max. Area	1008	962	980	1224	837	1296
Min. Area	464	360	288	512	338	648
Average Area	652.8	504.4	514.5	906.5	559.7	911.6

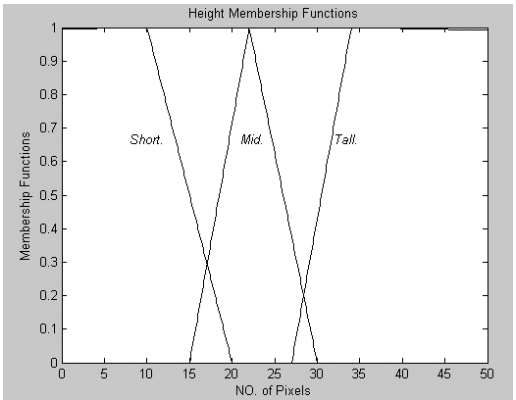
{Dot-position, Dot-count, Upper/Lower density, Connectivity, End points, Vertical and Horizontal crosses, Surrounding of the letter center (Black sides), Concavity, Closed loop}

Some other features are also used to increase accuracy.

5 ARABIC SUB-WORD AND FUZZINESS

Review of use of fuzzy Logic in OCR research has been made in [6]. The fuzzy sets are formed by the features of characters and are evolved from the existence of variation or uncertainty in the feature values of different samples. To illustrate this concept, different samples of the same sub-word have been considered. This research is dealing with 48 samples of a two character sub-word and 12 possible features for each sample, then a particular feature collected over the samples, forms a fuzzy set.

The reference fuzzy sets are obtained from these 48 sub-



word samples.[4] In order to recognize the characters, a triangular fuzzy membership function was selected. Each feature in a fuzzy set is fuzzified using this triangular membership function. The parameters for all fuzzy sets form the knowledge base (KB) for a reference character.

Figure 4: Membership Function for Height

The meaning of linguistic terms is defined by their membership functions. The graph of the sub-word height triangular membership function is shown in Figure 4.

6 CLASSIFICATION AND RECOGNITION

The classification algorithm is based initially on the number of diacritic dots. According to that, sub-words are divided into eight subgroups (i.e. one dot up e.g. نو , one dot down e.g.بر , one and two dots up e.g. نق , one dot up and two dots down e.g. ين, etc.) according to the number of the secondary characters in both letters. This number is a crisp value resulted from a prior fuzzy processing of secondary characters.

The variation in size of sub-words, which is evaluated by fuzzy model, was used for other subdivision of subgroups (Small area, Medium area, or Big area).

Then the presence and direction of unclosed path (curve) in one or both of the characters are classified into other subgroups. According to that, curves determined are those with the highest number of pixels.

Finally the classification of sub-word is then performed according to the presence of closed curve (loop) and its location (in the first and/or second character).

Combining fuzzy terms by logical operators (AND,OR), a set of fuzzy rules has been constructed for recognition. Examples of these rules are:

If ((RightMostDotPos is Up && (RightMostLoop || MidLoop) && Not (LeftMostLoop) && (RightMostDot Count is 1 || MidDotCount is 1) && LeftMostDotCount is 0 && ((RightMostConcave is Up && LeftMostConcave is Up) || SecondaryConcave is Up)) Then OutPut =' خا ' .

The features RightMostDotPos, RightMostLoop, RightMostConcave..etc. are described by fuzzy sets defined on their respective universe of discourse.

Mamdani reasoning method was used to develop the linguistic model. The fuzzy output is then defuzzified into a crisp number by using max-membership method of defuzzification.

7 RESULTS AND CONCLUSION

The proposed system combined a structural and statistical method for feature extraction and a modeling and classification technique based on fuzzy logic. Features were extracted from the main and secondary parts of a character. In total 12 chosen features were used for classification of all characters. The features were modeled by fuzzy linguistic values, which provided a more expressive natural system for the characters. A set of fuzzy rules was used for classification.[6]

A number of samples for handwritten two character sub-words has been used, and it achieved a recognition rate as 93% using the structural approach.

From study of the available sub-word images used in the present research, it has been noticed that the presence or absence of dots in two character sub-words is a fuzzy feature since the dot(s) may or may not appear exactly above or below the related main object of the character. Closed loop is a fuzzy feature also. Those properties have been managed by the fuzzy rules.

The overall character recognition rate, for some of the groups, was not satisfactory. The reasons for that was due to the following reasons [6,7]: Imperfectly or incompletely written characters, such as '**هـ**' with an open loop or even without a loop after smoothing. Radically embellished characters, such as an ending ' **ع** ' with a loop flourish at the end, misclassified characters because of similarity to other characters. The most common example is recognizing a ' **د** ' to be a '**ر** ', and vise versa.
The various writing styles of characters introduce large variations. The variability is addressed here by the use of fuzzy logic approach stems from its robustness.

REFERENCES

[1] Cox Earl, "The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems" 2nd Edition , Boston, AP Professional, 1998.

[2] Khedher M. Z. and Abandah G. A.,"Arabic Character Recognition Using Approximate Stroke Sequence", Workshop on Arabic Language Resources and Evaluation: Status and Prospects, LREC2002, Las Palmas de Gran Canaria, 1st June 2002

[3] Khedher M. Z., Abandah G. A., and Al-Khawaldeh A.M., "Optimizing Feature Selection for Recognizing Handwritten Arabic Characters", International Conference on Signal Processing - ICSP 2005 to be held on Feb. 25-26, 2005, Istanbul, Turkey

[4] Abandeh and Khedher M.Z., " Recognition of Printed and Handwritten Arabic Text: A Preliminary Study" A Report of Research Conducted with the support of Higher Council for Science and Technology, Amman, Jordan, 2003

[5] Khedher M. Z. and Al-Talib G., "*Recognition of Secondary Characters in Handwritten Arabic Using Fuzzy Logic*", International Conference on Machine Intelligence (ICMI'05), Tozeur, Tunisia, 2005.

[6] Ghaydaa A. Al-Talib, “Fuzzy Logic Based Arabic Optical Character Recognition”, Mosul University, Ph.D. Thesis, 2006.
[7] Kharma Nawwaf and Ward Rabab, “*A Novel Invariant Mapping Applied to Hand-written Arabic Character Recognition*”, Jou.: Pattern Recognition, V.34, 11, pp. 2115-2120, 2001.