A New Pattern Matching Approach to the Recognition of Printed Arabic

Ali M. OBAID
Department of Measurement and Information Systems
Technical University of Budapest
H-1521 Budapest, Műegyetem rkp. 9.
e-mail: obaid@mmt.bme.hu

Abstract

The paper presents a new segmentation-free approach to the Arabic optical character recognition. Extended with a suitable pre- and post-processing the method offers a simple and fast framework to develop a full OCR system. The method was developed primarily for the Naskhi font, it is however robust and flexible and can be easily extended.

Introduction

The most difficult problem in Arabic optical character recognition (AOCR) is to decide how to handle the cursiveness of the text. Thus while the segmentation is relatively simple in printed Roman texts, it is still an open question in Arabic. In most of the reported AOCR research the segmentation is considered the main source of recognition errors, see e.g. Al-Badr (1995). In addition, the presence of ligatures, especially those composed from dotted characters, adds to the problem so much, that until recently they were almost entirely omitted from the research. For a review of some of the problems of AOCR see Fig. 1.

AOCR followed the main approaches tried in Roman OCR research, consequently it focused for a long time on the issue of segmentation. Although various segmentation algorithms had been devised, see e.g. Amin (1989), cursiveness introduced serious problems, difficult to compensate even by additional processing. The application of advanced techniques, like neural networks, fuzzy techniques and hidden Markov models did not bring the expected breakthrough, due to the inherent segmentation problems, see Walker (1993). Recently, Al-Badr (1995) attempted to avoid segmentation at all. Using morphological operators he tried to recognize at least a part of a word and then the entire word by searching a large data-base of references. The scheme was handicapped however by the extensive Arabic vocabulary.

1 The Outline of the Approach

The proposed approach can dispense with the traditional segmentation, and in advanced version even with the segmentation into sub-words and text lines. The basic idea is that not every component of an character is essential to the OCR process. Consequently, computing features from non-informative segments wouldn't contribute to the recognition. Contrary to the traditional segmentation where words are scanned looking for segment boundaries, in presented approach special points are identified in the interior of the characters. These points serve as references for configurations of sensors (referred to as focal points and N-markers respectively) designed to identify the essential character strokes. By distributing enough markers over a character, a letter or a group of letters (ligature) can be positively detected. The approach is related to some early ideas of the OCR of the isolated Roman characters (N-tuples and Character Loci), see Ullman (1969).

For all of the practical purposes the presented method must be extended with obligatory and

optional processing steps. The scanned text should be treated for minor noise removal, skew corrected and normalized. In the basic algorithm text lines and words should be separated, then thinned and smoothed. The method should be completed with symbolic rules resolving the residual ambiguities of the kernel method.

The present research was aimed at the recognition of printed Arabic, widely used in books and renown periodicals. Such texts are printed almost entirely in so-called Naskhi font, sometimes even identified with the printed Arabic. Various typesetting sources introduce however variations to the basic Naskhi shapes, which means further problems in the recognition. The tackled AOCR problem is hence, on the one hand, restricted to a single albeit extensively used font. On the other hand minor font variations and a wide spectrum of ligatures are treated. We aimed also at a method robust enough to handle degraded text and adaptable to other fonts if required.

2 N-Markers

As an example let us consider the isolated characters 'Waw' and 'Qaf' (Fig. 2.). Both shapes posses loops and tails, however the topology of tails differ. Let us assume that a suitable focal point (the junction below the loop) is already detected. Then the presence or the absence of a tail can be measured by placing marker $\mathbf{m}_1$ below the junction on the expected path of the tail. That way we attempted to restrict the class of shapes to both 'Waw' and 'Qaf'. Another marker $\mathbf{m}_2$, placed well to the left, will distinguish now between them. Interpreting the presence of the shape fragment under the markers in terms of logical functions we have accordingly:

'Qaf' = ($\mathbf{m}_1$ = YES) AND ($\mathbf{m}_2$ = YES), while
'Waw' = ($\mathbf{m}_1$ = YES) AND ($\mathbf{m}_2$ = NOT).

For a meaningful detection focal points should be properly selected. Such points (line ends, junctions and a number of special patterns) should be 'stable', i.e. easy to detect, relatively immune to distortions and of pronounced appearance in all of the investigated font variations. Definition of the markers requires further an uniform normalization of the text lines (chosen as 100 pixels high, which together with the assumed minimal size of 12 point still yields an acceptable quantization noise).

Contrary to the schemes found in the literature (e.g. Ching (1982)), classifying the shapes is not based primarily on the shape similarity, but rather on focal points and marker (stroke) configurations instead. E.g. 'initial Lam' is similar to 'medial Lam', yet their focal points are different (a line end vs. a junction). Consequently, they belong to different classes. The rationale of this approach is that it allows recognition of multiple shapes by the same marker configuration, making thus the treatment of ligatures more straightforward.

Consider for example the shape class in Fig. 3. It contains six shapes: four characters (initial, medial, terminal and isolated 'Hha') and two ligatures ('Lam+Hha' and 'Meem+Hha'). A well developed 3-way junction is used as focal point. Markers $\mathbf{m}_3$, $\mathbf{m}_4$, $\mathbf{m}_7$, and $\mathbf{m}_8$ detect strokes common to the class members. Remaining markers are used to differentiate between particular shapes. For every shape Boolean test functions are defined, e.g.:

medial 'Hha' =
$\neg(\mathbf{m}_1 \vee \mathbf{m}_2 \vee \mathbf{m}_6 \vee \mathbf{m}_7 \vee \mathbf{m}_8) \wedge (\mathbf{m}_3 \wedge \mathbf{m}_4 \wedge \mathbf{m}_5 \wedge \mathbf{m}_9)$

where the logical value of $\mathbf{m}_i$ depends whether the required stroke is present or not.

For a moment N-markers configurations were designed manually by collecting sufficient number of thinned samples. After choosing a suitable focal point the shapes were superimposed and aligned to show how much they vary. Marker configurations were defined around the designated focal point, by assigning markers to every critical line segment. They were then iteratively tested and modified.

3 Pre- and Post-Processing

The kernel of any OCR system are feature extraction and classification. In practice these operations must be preceded by suitable

procedures collectively called pre-processing. In the proposed system pre-processing includes minor noise removal, correction for skewness, line separation, normalization of the text lines, word separation, dot extraction, thinning of the isolated words, and smoothing of the word skeletons.

Although the proposed method (by defining focal point patterns in larger windows and introducing approximate instead of exact matching) could be applied to regular (non-thinned) words, focal points are much easier to find along thinned skeletons. Several thinning algorithms are available in the literature, see Lam (1992). The single dots are, however, a critical issue, and should be extracted before thinning. In time of development of the method no satisfactory thinning algorithm could be found, consequently further processing (smoothing) of the skeleton was necessary.

To complete the method a suitable post-processing is also required to correct recognition errors and side-effects introduced by pre-processing and classification. In the proposed system post-processing is performed by symbolic rules. *Redundancy removal rules* are needed due to the necessary trade-offs in designing N-markers configurations. Using simple rules is a more straightforward strategy than increasing the number of markers. *Dot and 'Hamza' association rules* complete the recognition of shapes (especially ligatures), differentiated solely by the presence of dots or 'Hamza'. *Ambiguity resolution rules* handle cases when (in poor quality text) thinned images of 'Hamza' and three dots coincide. Finally *Combining shapes rules* connect subcharacters into characters if necessary.

4 Verification of the Method

For the testing ten densely printed pages (including ligatures) were scanned, using HP Scan Jet IIcx Scanner at 300 dpi resolution, from a good quality book type-set in Naskhi font, Haekl (1983), together with two pages of degraded text taken from a magazine printed on a highly reflective and smooth paper, Al-Arabi (1996). Due to the very low incidence of some of the ligatures, a collection of ligatures (2 pages of unrelated words each containing at least one ligature) was prepared (printed and scanned) for testing purposes. In addition, suitable files from the test repository of Al-Badr, see WWW in References, were also borrowed for testing.

Frequent shapes showed recognition rates of 95%-98%. Testing for rare shapes (on artificial samples) yielded similar results. Recognition rate for degraded image (filled loops) dropped, as expected. Last test involved two degraded pages from the magazine. To deal with degradation, markers were enlarged from 1-dim line segments to 2-dim windows, covering larger portions of the strokes, without disturbing however their configuration relative to focal points and being that way more robust, yet still selective enough (see Fig. 4.). The method yielded results comparable to those obtained for good quality pages. The only problem observed in the testing were loops.

Time needed to process a full page (pre-processing, detection of focal points, worst-case application of the markers, and post-processing) was estimated as app. 135 sec. which was equivalent to the recognition rate of 340 characters/minute. This result is promising considering that it represented the lower limit of performance.

5 Extensions to the Method

The basic implementation does not exploit fully other advantageous aspects of the N-markers. Treatment of elongation is easy, considering that no focal points appear along an elongated line and nothing disturbs the detection. Most shapes include several possible focal points. In applying markers, in the worst case, all candidate focal points have to be considered. An intelligent selection of the focal points with respect to their usefulness would considerably speed-up the process.

Another possibility is to apply markers also over unthinned strokes. To this purpose the processing of focal points should be extended,

as mentioned, to an approximate matching. A further natural extension is shape extraction from a full non-segmented page. Knowledge of where the text line and word strokes are, is not essential. Windows detecting focal points can be slid along the whole page in any direction, with a possible parallel implementation.

A question to solve is at least a partial automation of the manual designing of the markers. Efficient heuristic algorithm could possibly be developed, theoretically the problem is intractable due to the equivalence to the NP-complete N-tuple configuration problem, shown by Jung (1996). Related question is the extension of the marker configurations to the new fonts. Although the target font was the widely spread Naskhi, the concept of N-markers is not confined to this font alone. One way of extending the method is by constructing markers for new fonts in a manner describe above. Another approach could be to identify the relation between the fonts as a local nonlinear image transform. Then this transform could be used to deform the marker configurations to fit the new shapes.

Conclusion

Experiments with N-markers show promising results. The main source of errors in AOCR is avoided. The method is intuitive and works with unified features. Handling large and diverse set of shapes including ligatures is relatively easy. 'Shape similarity' is based on focal points, rather, than on the apparent visual similarity, which can lead to mistakes. The accommodation of the possible variations of the font is straightforward and is insensitive to the character of the shape differences. The method is simple to implement and does not require lengthy numerical computations. The very idea is open to extensions and is relatively immune to degradation of the text. The primary disadvantage of the basic (thinned words) technique is its dependency on the size and orientation of the text, redundancy of the focal points, sensitivity of the focal points to degradation, dependence on the thinned image. These problems can be largely solved by switching to the unthinned text processing, which is under investigation. A question is the heuristic automation of the 'manual tuning' of the classes. Finally some of the essential processing steps of the method are illustrated in Fig. 5.

References

Al-Arabi (1996) Ministry of Culture, Kuwait, April.

Al-Badr B. and Haralick R. (1995) *Segmentation-free word recognition with application to Arabic*, in Proc. of the IEEE 3rd Int. Conf. on Document Analysis and Recognition - ICDAR'95, p. 355, August.

Amin A. and Mari J. (1989) *Machine Recognition and Correction of Printed Arabic Text*, IEEE Trans. on Systems, Man and Cybernetics, vol. 19, no.5, pp. 1300, Sept/Oct.

Ching Y. Suen (1982) *Distinctive Features in Automatic Recognition of Handprinted Characters*, Signal Processing, vol. 4, pp. 193-207.

Haekl M. (1983) *At the crossroads*, Printing and Publishing Comp., Beirut, Lebanon.

Jung D., Krishnamoorthy M., Nagy G. and Shapira A. (1996) *N-Tuple Features for OCR revisited*, IEEE Trans. on Pattern Recognition and Machine Intelligence, vol. 18, no. 7, pp. 734-745, July.

Lam L., Lee S., and Suen C. (1992) *Thinning Methodologies - a Comprehensive Survey*, IEEE Trans. on PAMI, Vol.14, No.9, pp. 869-884, Sept.

Ullman J. (1969) *Experiments with the N-tuple method of Patter Recognition*, IEEE Trans. on Computers, vol. 18, no.12, pp. 1135-1137, Dec.

Walker R. F., Bennamoun M. and Boashash B. (1993) *Comparative Results for Arabic Character Recognition Using Artificial Neural Networks*, in Proc. of WoSPA'93, SPRC Workshop on Signal Processing and its Applications, Dec., Brisbane, Australia

WWW
http://george.ee.washington.edu/~badr/ARABIC,
http://george.ee.washington.edu/~badr/SAIC

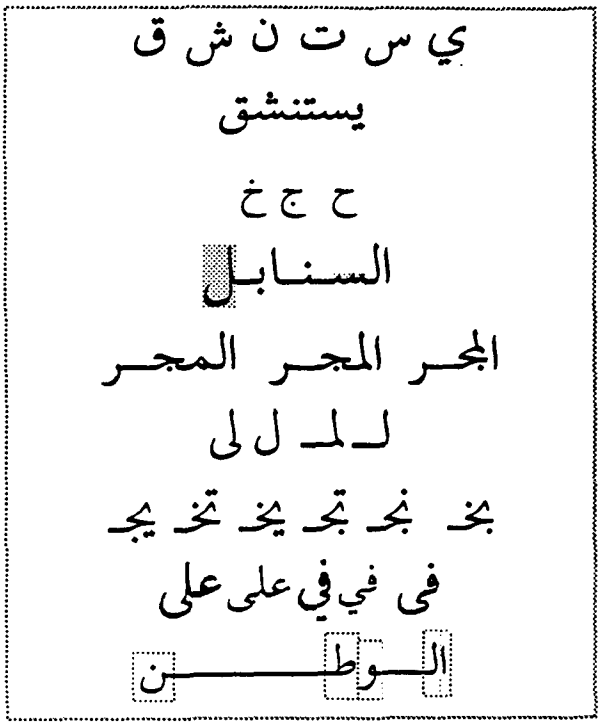

Fig. 1. Problems encountered in the Arabic OCR: Difference between isolated and connected characters; Character shapes differentiated only by dots; Variation in size: medial 'Seen' (right) and terminal 'Lam' (left); Changing shape due to ligation; Similarity of ligatures to the normal characters ('Lam'+'Ya' vs. isolated 'Lam', and 'Lam'+medial 'Meem' vs. initial 'Lam'); Ligatures with relatively small body and complicated dot constellations (from left to right: 'Ya'+'Jeem', 'Ta'+'Kha', 'Ya'+'Kha', 'Ta'+'Jeem', 'Noon'+'Jeem', 'Ba'+'Kha'); Minute variation in the Naskhi font; Elongation of words.

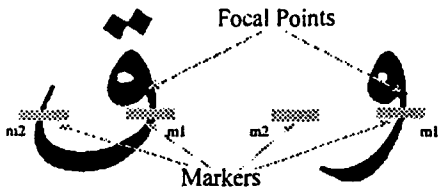

Fig. 2. Selective detection with focal points and markers (Qaf - left, and Qaw - right)..

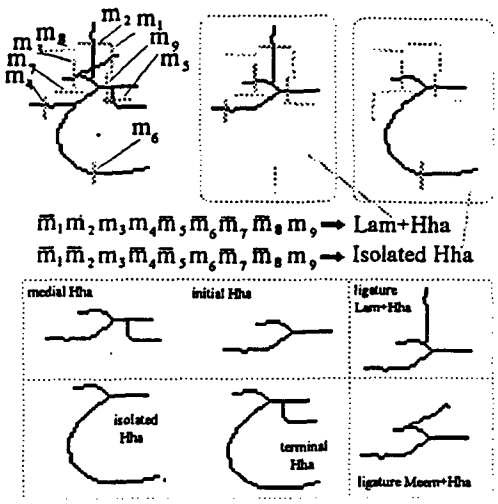

Fig. 3. One of the shape classes used in detection (composite shape and N-markers configuration, basic shapes, and test functions for two particular shapes).

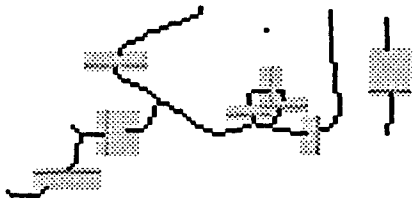

Fig. 4. Extended markers to compensate distortion problems.

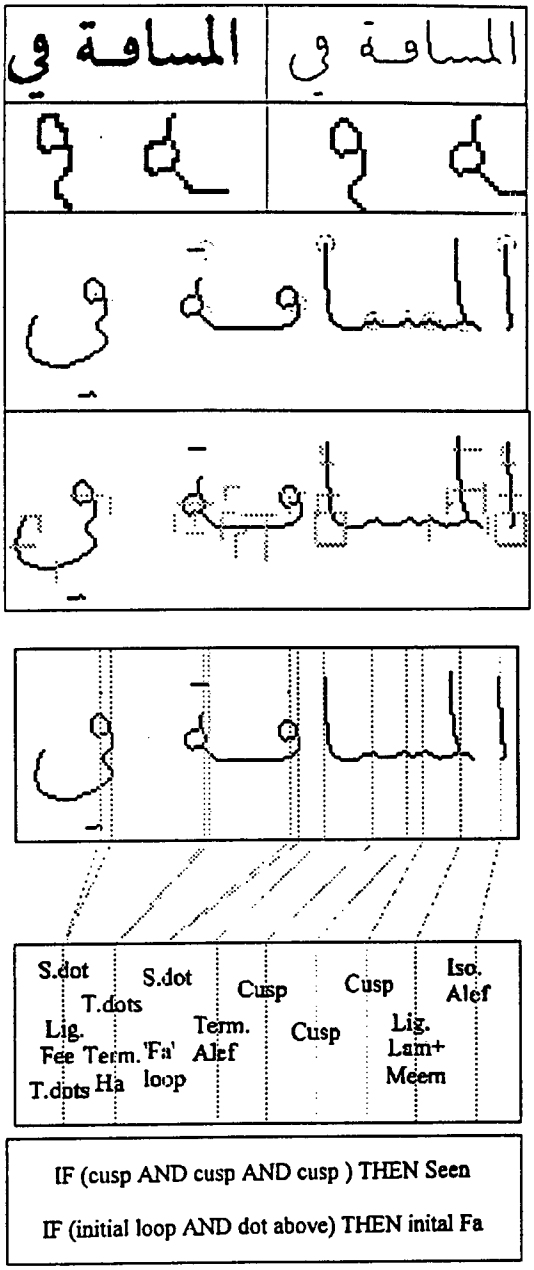

Fig. 5. Essential phases of the proposed character detection: thinning, skeleton smoothing, detecting focal points, applying markers, preliminary classification, applying symbolic rules.