Word Spotting for Arabic Handwritten Historical Document Retrieval using Generalized Hough Transform

Nabil Aouadi
UTIC: research Unit of Technologies of Information and Communication
ESSTT: High School of Sciences and Techniques of Tunis, University of Tunis, Tunisia
nabil.aouadi@utic.rnu.tn

Afef Kacem
UTIC: research Unit of Technologies of Information and Communication
ESSTT: High School of Sciences and Techniques of Tunis, University of Tunis, Tunisia
afef.kacem@esstt.rnu.tn

Abstract— **Because of the high noise levels in historical documents and the great amount of variability in handwriting, handwritten historical documents are currently transcribed by hand. Easy access to such documents requires an index, which is currently created manually at great cost. The goal of the Word Spotting idea, applied to handwritten documents, is to greatly reduce the amount of annotation work that has to be performed, by grouping all words into clusters. This paper explores the use of GHT (Generalized Hough Transform) in case of word spotting for Arabic handwritten historical document retrieval. We applied GHT to identify all positions of a given word in a document. It has the advantage of being relatively unaffected by image noise. Experiments that have been conducted on the historical documents of the Tunisian national Archive show the advantage of the proposed approach.**

Keywords-Generalized Hough Transform; word spotting; pattern recognition; image processing

I. INTRODUCTION

As historical library collections across the world hold huge numbers of handwritten documents, digitizing these manuscripts should be of great interest since their content can be conserved and made available to a large community via the Internet or other electronic media. Such corpora can nowadays be shared relatively easily, but they are often large, unstructured, and only available in image formats, which makes them difficult to access. In particular, finding specific locations of interest in a handwritten image collection is generally very tedious. Easy access to such collections requires an index, which is currently created manually at great cost in terms of time and money.

Due to the large number of handwritten manuscripts, the great amount of variability in handwriting, the high noise levels in historical documents, the transition from traditional to digital libraries pose a great challenge. As automatic handwriting recognizers fail on historical manuscripts, the word spotting technique has been developed: the words in a collection are matched as images and grouped into clusters which contain all instances of the same word. By annotating "interesting" clusters, an index that links words to the locations where they occur can be built automatically. Due to the noise in historical documents, selecting the right features for matching words is crucial. This paper explores the use of GHT [18] technique to identify all positions of a given word in a document.

The HT (Hough Transform) is a technique which can be used to isolate features of a particular shape within an image [17]. Because it requires that the desired features be specified in some parametric form, the *classical* Hough Transform is most commonly used for the detection of regular curves such as lines, circles, ellipses, etc. A *generalized* Hough Transform can be employed in applications where a simple analytic description of a feature(s) is not possible.

The idea is to detect words through their parts of words outlines using GHT. This technique is also known to be tolerant of gaps in feature boundary descriptions and relatively unaffected by image noise. The outline of the paper is as follows. Section II presents some works devoted to the word spotting for Latin handwritten documents. Section III highlights on characteristics of the old Arabic handwriting and the difficulties which it poses. Section IV gives some details of the proposed system of word spotting for Arabic handwritten historical document retrieval based on GHT technique. Section V draws some conclusions and gives an outlook for future work

II. STATE OF ART

Word Spotting is an approach that clusters word images to identify and annotate content-bearing words in a collection. The goal is to greatly reduce the amount of annotation work that has to be performed, by grouping all words into clusters. Ideally, each cluster contains words with the same annotation. Once such a clustering of the data set exists, the number of words contained in a cluster can be used as a cue for determining the importance of the word as a query term. For example, highly frequent terms, such as *the*, *of*, etc. are *stop* words and can be discarded. All clusters with terms that are deemed important can then be manually annotated. This makes it possible to construct a partial index for the analyzed document collection, which can be used for retrieval.

Search works on word spotting for Latin handwriting attracted much attention [1][2][3] and have offered a starting point for the development of new systems. Several methods have been compared in [4] and showed that the

 ISBN: 978-1-61208-158-8

best accuracy was obtained with the method based on DTW (Dynamic Time Warping): an algorithm for measuring similarity between two sequences which may vary in time or speed [19].

Word spotting in the Arabic script dates back several years when we unfortunately offered little work and solutions for handwritten Arabic script. In [5], the study discusses the performance achieved by applying the concept of Word Shape applied on Arabic Handwritten Documents.

The use of CEDEARABIC system becomes very frequent. The word spotting in this system is divided into two stages: indexing and search. Each image generates a set of queries by sub-images. Note that most of the functionality of the system CEDEARABIC, including word spotting, are those of the American Fox [6][7].

You et al. [8] presented a hierarchical Chamfer matching scheme as an extension to traditional approaches of detecting edge points, and managed to detect interesting points dynamically. They created a pyramid through a dynamic thresholding scheme to find the best match for points of interest. The same hierarchical approach was used by Borgefors [9] to match edges by minimizing a generalized distance between them.

Rothfeder et al. [10], Srihari et al. [11] presented a system for spotting words in scanned document images for three scripts: Devanagari, Arabic, and Latin. Their system retrieved the candidate words from the documents and ranked them based on global word shape features.

An algorithm for robust machine recognition of keywords embedded in a poorly printed document was presented by Kuo and Agazzi [12]. For each keyword, two statistical models were generated – one represents the actual keyword and the other represents all irrelevant words. They adopted dynamic programming to enable elastic matching using the two models.

A language independent system for preprocessing and word spotting of historical document images was presented by Moghaddam et al. [13], which has no need for line and word segmentation. In this system, spotting is performed using the Euclidean distance measure enhanced by rotation and DTW.

As far as we know, up till now, there is no system that allows word spotting for handwritten documents using GHT which seems to be of great interest to detect words.

III. CHARCTERISTICS OF ARABIC HIRISTOCAL HANDWRITING

A summary of the features of Arabic writing appears below.

- Arabic text, both handwritten and printed, is cursive. The letters are joined together along a writing line (see Figure 1). This is similar to Latin 'joinedup' handwriting, which is also cursive, but in which the characters are easier to separate.
- In contrast to Latin text, Arabic is written right to left, rather than left to right.
- Arabic contains dots and other small marks that can change the meaning of a word, and need to be taken into account by any computerized recognition system.
- The shapes of the letters differ depending on whereabouts in the word they are found. The same letter at the beginning and end of a word can have a completely different appearance as shown in Figure 2. Along with the dots and other marks representing vowels, this makes the effective size of the alphabet about 160 characters.

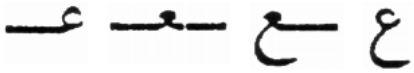


Figure 1. Different forms of the Arabic letter "ع". From left to right beginning, middle, end, isolated

- In Arabic handwriting the word is composed of one or several of parts of word. A PAW (Part of Arabic Word) is a connected component which can refer to a diacritic sign, a single letter, a sequence of letters or whole word (see Figure 2).

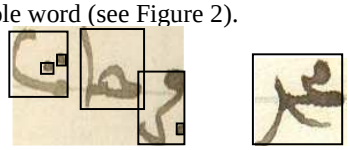


Figure 2. Example of a words composed of several number of PAWs

- Because letters can be horizontally and/or vertically ligaturated, as shown in Figure 3, the segmentation is not an easy busness.

Figure 3. Problem of segmentation due to letter ligature

For Arabic historical handwriting, other difficulties are compounded. In the historical documents that we handle, we find that for some letters, the number and/or position of their diacritic points were changed. For example, the letter ڤ is written ق: a single diacritic point above the letter body instead of two whereas the letter ڤ is written with a single diacritic point below the letter body: ڤ.

IV. PROPOSED SYSTEM

The proposed system, described here, has been built on a subset of the Tunisian National Archive collection. We are interested by documents that are from the 19th century and correspond to enumeration registers at Tunisian protectoral period. Figure 4 displays a small portion of this manuscript: a set of lines composed of list of names. These documents are written in one author's hand. This reduces the amount of handwriting variations that have to be compensated for.

The proposed system retrieves actual handwritten pages (not transcriptions) given text queries. It is based, as it will

be explained, GHT technique that we have improved by a clustering process. The system consists of training and recognition phases.

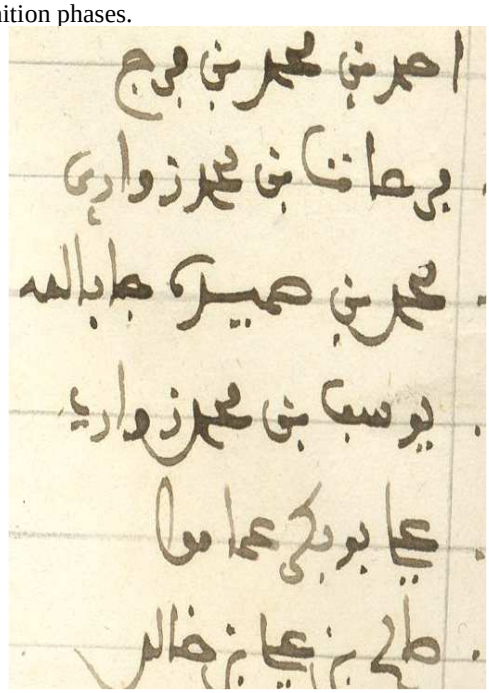


Figure 4. Example of page of the enumeration register

At the training phase, the system generates patterns of PAWs considering different samples. The PAW pattern is created from its contour points and gravitational center coordinates. PAW patterns are then registred in a table of reference as it will be explained later. Afterwards, the system builds a dictionary which includes word queries. For each word, the system saves its composing PAWs, the corresponding *d-cluster* (the minimal distance between word clusters which is empirically determined and for each PAW, the associated pattern, the Hough threshold (number of minimal votes for the PAW) and the Hough threshold for points that are close to the voting points. In phase of recognition, the system, given the word to spot, it computes GHT parameters for each of PAWs and scans the page pixel by pixel looking for the voting points and their neighborhood. In fact, the PAW patterns will largely vote for similar patterns and will form clouds of points, also called voting clusters, in the Hough space. If the distance between voting clusters is less than *d-cluster* then the system merge them into one cluster. This process has significantly improved the performance of word spotting as it will be demonstrated later. Let us recall the GHT principle then consider its application for word spotting.

A. GHT Technique

Hough Transform is a feature extraction technique to find imperfect instances of objects within a certain class of shapes by a voting procedure. This voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by the algorithm for computing the Hough Transform.

The classical HT was concerned with the identification of lines in the image, but later the Hough Transform has been extended to identifying positions of arbitrary shapes, most commonly circles or ellipses. The GHT, introduced by D.H. Ballard in 1981 [18], is the modification of the HT using the principle of template matching. This modification enables HT to be used for not only the detection of an object described with an analytic equation (e.g. line, circle, etc.). Instead, it can also be used to detect an arbitrary object described with its model.

The problem of finding the object (described with a model) in an image can be solved by finding the model's position in the image. With the GHT, the problem of finding the model's position is transformed to a problem of finding the transformation's parameter that maps the model into the image. As long as we know the value of the transformation's parameter, the position of the model in the image can be determined.

B. Word spotting using GHT Technique

The original implementation of the GHT uses edge information to define a mapping from orientation of an edge point to a reference point of the shape. In the case of a binary image where pixels can be either black or white, every black pixel of the image can be a black pixel of the desired pattern thus creating a locus of reference points in the Hough Space. Every pixel of the image votes for its corresponding reference points. The maximum points of the Hough Space indicate possible reference points of the pattern in the image.

A table of reference, called R-Table, defines the correspondence between the space image and parametric space [13]. The GHT acts on characteristic points of the image generally contour image. It makes it possible to describe the form by R-table in phase of training, while, in phase of recognition it exploits various R-tables obtained in the phase of training to generate spaces of votes which make it possible to carry out classification [14].

To build R-tables, our system extracts all the connected components of words which are the set of PAWs and generates a pattern for each one of them. The pattern is deduced from PAW contour points and a reference point, here the PAW gravitational center *G*. For each point *P*, as shown in Figure 5, the system selects the two neighboring points P_1 and P_2. Then, it determines the tangent line at *P* which is parallel to the line (P_1P_2) passing through *P*. Let β be the angle between the tangent line at *P* and the horizontal (x-axis). Let α be is the angle between the orthogonal line to the tangent at *P* that passes through *P* and the line (*GP*). All computed values are stocked in R-table in function of β as displayed in Table 1.

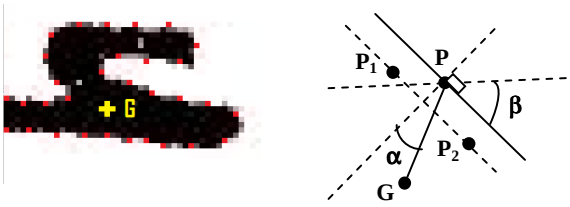


Figure 5. PAW pattern creation

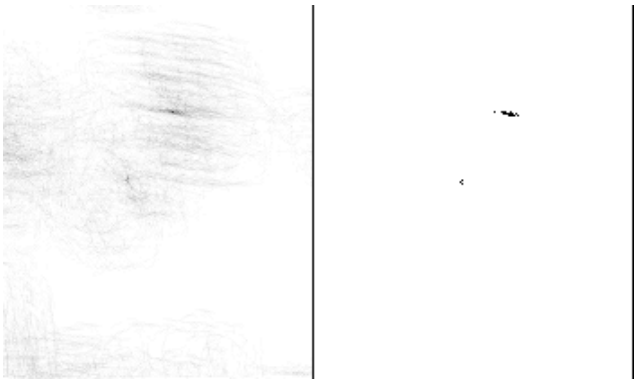
TABLE 1. R-TABLE CONSTRUCTION

β_j	(α_i, P_iG)
30°	(α_i, P_1G), (α_i, P_2G), (α_i, P_3G), …
40°	…
…	….

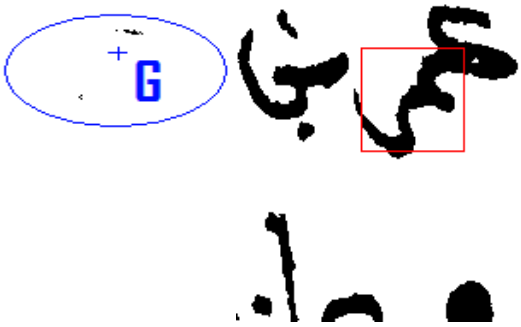
During the recognition phase, here for word spotting, the system extracts the contour of PAWs and passes a sliding window through the contour. At each passage and for a point inside the window, the system determines the tangent line at this point with the ends of the contour points that touch the border of the window. Then the system calculates β and consults R-table in search of couples α and P_iG which coincide with β. The found couples are the voting points. Finally, if the number of voting points is above a Hough threshold then the PAW is located. As our objective is to locate the entire word in the document image, the same above process is repeated for the rest parts of word. We also considered the voting clusters of PAWs and the distance, *d-cluser*, between them.

Let us give an example of word spotting using GHT for the word "عمر". In Figure 6 (a), many clusters of voting points are formed for the word parts: "ع" and "مر". The voting cluster is simply a transcription of the letter or word part image where are recorded the votes of each pixel. In Figure 6 (b), the system selects points that have the most votes. When the voting clusters are close in term of Euclidean distance: *d-cluster*, as shown in Figure 6 (c), they will be merged into one cluster which represents the votes of individual patterns that compose the word to be located.

(a) GHT application on word parts : "ع" and "مر"



(b) Selection of points that have the most votes



(c) Combinations of neighboring clusters by *d-cluster* distance

Figure 6. Word spotting of the name "عمر" using GHT

C. System evaluation

To evaluate performance of our system, we used a database which consists of 23 pages. Each page includes about 36 lines. Each line contains 4 names on average. So we handle a total of 3312 words. These pages are scanned with a resolution of 300 dpi and saved in TIFF format.

Figure 7 displays the results of a part of the word name "علي" retrieval. Several tests were carried out on other words. The observed results are quite satisfactory. Experiments that have been conducted on the historical documents of the Tunisian national Archive show the advantage of the proposed approach. This approach is relatively unaffected by image noise. But the main drawbacks of this approach, based on GHT, are its substantial computational and storage requirements that become acute when word orientation and scale have to be considered. At the current state of the system, the computing time is about one second.

 ISBN: 978-1-61208-158-8

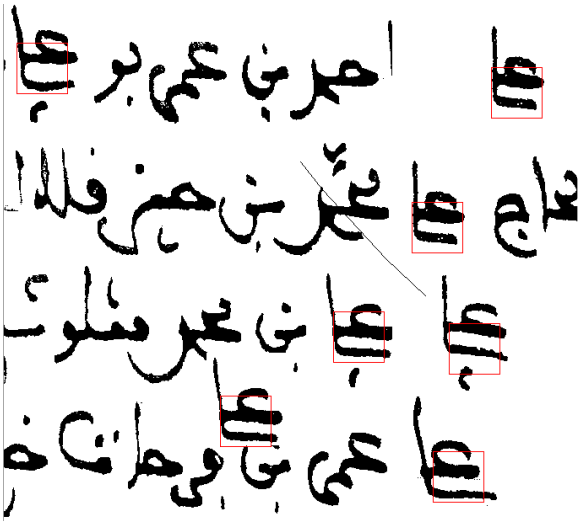


Figure 7. Word spotting of the name "علي" using GHT

V. CONCLUSION AND FUTURE WORK

This paper discusses and summarizes the work covered, its contribution to the word spotting for Tunisian historical handwritten documents. The proposed system is based on GHT to identify all positions of a given word in a document. The GHT is generally used to detect arbitrary shapes. As handritten words have no simple analytical forms, we tried here to benefit from this technique for word spotting. The system attempts to improve performance by grouping vote clusters of letters or part of words which compose the enquires words. How to choose models to optimize GHT? How to enhance system robustness against noise and poor quality of pages? There is still a lot to do but it is really worthwhile.

ACKNOWLEDGEMENT

We thank O. Elghoul for kindly providing us with the GHT source code.

REFERENCES

[1] N. Ben Amara, A. Belaïd, and N. Ellouze, "Modélisation Pseudo bidimensionnelle pour la Reconnaissance de Chaînes de Caractères Arabes Imprimés," Proc. CIFED 98, Québec, Canada, pp. 131-140, 1998.

[2] S. Kuo and O. Agazzi, "Keyword Spotting in Poorly Printed Documents using 2-d Hidden Markov Models", IEEE Trans. PAMI, 16, pp. 842–848, 1994.

[3] M. Burl and P. Perona, "Using Hierarchical Shape Models to Spot Keywords in Cursive handwriting," IEEECS Conference on Computer Vision and Pattern Recognition, pp. 535–540, June 1998.

[4] A. Kolz, J. Alspector, M. Augusteijn, R. Carlson, and G. V. Popescu, "A Line-oriented Approach to Word Spotting in Handwritten Documents," Pattern Analysis and Applications, 2(3), pp. 153–168, 2000.

[5] R. Manmatha and T. M. Rath, "Indexing of Handwritten Historical Documents-recent Progress," Proc. SDIUT 03, pp. 77–85, 2003.

[6] B. Zhang S. N. Srihari and C. Huang, "Word Image Retrieval using Binary Features," Proc. DRR 04, SPIE Vol. 5296, pp. 45–53, 2004.

[7] S. N. Srihari, S.H. Cha, H. Arora, and S. Lee, "Individuality of handwriting," Forensic Sciences, 47(4), pp. 856–872, 2002.

[8] S. N. Srihari, B. Zhang, C. Tomai, S. Lee, Z. Shi, and Y. C. Shin, "A system for handwriting matching and recognition," Proc. SDIUT 03, Greenbelt, MD, pp. 67–75, 2003.

[9] J. You, E. Pissaloux, W. Zhu, and H. Cohen, "Efficient Image Matching: A hierarchical Chamfer Matching Scheme via Distributed System", Real-Time Imaging, 1(4), pp. 245 – 259, 1995.

[10] G. Borgefors, "Hierarchical Chamfer Matching: Aparametric Edge Matching Algorithm," IEEE Trans. Pattern Anal. Mach. Intell., 10(6), pp. 849–865, 1988.

[11] J. L. Rothfeder, Shaolei Feng, and Toni M. Rath, "Using Corner Feature Correspondences to Rankword Images by Similarity," Proc. DIAR 03, Madison, WI, June 2003.

[12] S. Srihari, H. Srinivasan, C. Huang, and S. Shetty, "Spotting Words in Latin, Devanagari and Arabic scripts," Vivek: Indian Journal of Artificial Intelligence, 16(3), pp. 2–9, 2003.

[13] S. Kuo and O. Agazzi, "Keyword Spotting in Poorlyprinted Documents using Pseudo 2-d Hidden Markov Models," IEEE Trans. Pattern Anal. Mach. Intell., 16(8), pp. 842–848, 1994.

[14] R. Moghaddam, D. Rivest-Hénault, and M. Cheriet, "Restoration and Segmentation of Highly Degraded Characters using a Shape-Independent Level Set Approach and Multi-level Classifiers," Proc. ICDAR 09, Barcelona, Spain, pp. 828–832, 2009.

[15] M. Ulrich, C. steger, A. Baumgartner, and H. Ubner, "3 Real Time Object Recognition Using a Modified Generalized Hough Transform," Eckhardt Seyfert, editor, photogrammetrie -- Fernerkundung-- Géoinformation : Geodaten schaffen verbindungun, 21. Wissems chaftlich-Technische Jahrestagung der DGPF, Berlin, 2001, pp 571-578.

[16] S.Touj, N. Ben Amara, and H. Amiri, "Reconnaissance de l'écriture Arabe Imprimée par une Approche de Segmentation par Reconnaissance Basée sur la Transformée de Hough Généralisée," SETIT 03, Mars, Sousse, Tunisie 2003.

[17] R. Fisher, S. Perkins, A. Walker, and Erik Wolf, "Image Transforms - Hough Transform," Homepages.inf.ed.ac.uk. Retrieved 2009-08-17.

[18] D.H. Ballard, "Generalizing the Hough Transform to Detect Arbitrary Shapes", Pattern Recognition, Vol.13, No.2, pp. 111-122, 1981.

[19] C. S. Myers and L. R. Rabiner, "A Comparative Study of Several Dynamic Time-warping Algorithms for Connected Word Recognition," The Bell System Technical Journal, 60(7), pp. 1389-1409, September 1981.

 ISBN: 978-1-61208-158-8