World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:2, 2008

Ottoman Script Recognition Using Hidden Markov Model

Ayşe Onat, Ferruh Yildiz, and Mesut Gündüz

***Abstract*—**In this study, an OCR system for segmentation, feature extraction and recognition of Ottoman Scripts has been developed using handwritten characters. Detection of handwritten characters written by humans is a difficult process. Segmentation and feature extraction stages are based on geometrical feature analysis, followed by the chain code transformation of the main strokes of each character. The output of segmentation is well-defined segments that can be fed into any classification approach. The classes of main strokes are identified through left-right Hidden Markov Model (HMM).

***Keywords*—**Chain Code, HMM, Ottoman Script Recognition, OCR

I. Introduction

THE Ottoman Empire lasted until the twentieth century. Ottoman Empire comprised an area of about 5.6 million km². After the decline of Ottoman Empire nearly 30 countries appeared in this area. Therefore the written state archives of these countries were Ottoman Script. Also in Turkey the state archives are in Ottoman Script. For these reason it is important to recognize Ottoman Script.

Ottoman Script is a variant of the Turkish language which was used as the administrative and literary language of the Ottoman Empire, containing extensive borrowings from Persian, which in turn had been permeated with Arabic borrowings. Spoken Turkish lived and developed alongside Ottoman Turkish and was greatly influenced by its extensive borrowings from Arabic and Persian.

Optical Character Recognition (OCR), involves a system designed to translate images of typewritten or handwritten text into machine-editable text. By the development of artificial intelligence techniques many OCR applications developed and satisfactory results obtained. Most of the works done in this topic is about Latin character recognition. It is difficult to work Arabic characters or Ottoman characters because of the script characteristic. There are some works done in Arabic character recognition, but for Ottoman scripts the studies is not sufficient. In this study Hidden Markov

Models are used for recognition of Ottoman Scripts. The detailed study and the satisfactory results explained.

Character recognition problem is transferring a page to the computer that contains symbols and matching these symbols with previously known or recognized symbols. After extraction the features of these symbols via appropriate with preprocessing methods.

Manuscript received June 30, 2006.

A. Onat is with Selcuk University, Bozkır Vocational School of Higher Education, Computer Department, Turkey (phone: +90-332-426 1444, e-mail: aonat@selcuk.edu.tr).

F. Yildiz is with Selcuk University, Engineering and Architecture Faculty, Geodesy & Photogrammetry Engineering Department, Turkey (e-mail: fyildiz@selcuk.edu.tr).

M. Gündüz is with Selcuk University, Engineering and Architecture Faculty, Computer Engineering Department, Turkey (e-mail: mgunduz@selcuk.edu.tr).

II. Ottoman Scripts

The Ottoman alphabet contains 28 letters. Each character has between two and four shapes. This shape depends on the position of the letter within its word or subword. The shapes have the four conditions: beginning of a (sub)word, middle of a (sub)word, end of a (sub)word, and in isolation. Table I shows each shape for each letter. For example, the letter which hasn't initial or medial shape, can not be connected to the following letter as in the Table I [1].

III. Digitization and Preprocessing

A 300 dpi scanner was used to digitize the image for this investigation. After the colored image was taken, it is converted to gray. Then this gray image is converted to binary image to use for investigation. Finally, skeletonization algorithm was used for thinning.

IV. Skeletonization

The skeleton of a binary object is a collection of lines and curves that encapsulate the size and shape of the object. There are in fact many different methods of defining a skeleton. In this study Zhang-Suen's Skeletonization Algorithm is used for thinning. The algorithm steps are shown below;

N Flag a foreground pixel p=1 to be deletable if

1. $2 =< B(p) =< 6$
2. $X(p) = 1$,
3. If N is odd, then
   $p2 * p4 * p6=0$
   $p4 * p6 * p8=0$

If N is even, then
$p2 * p4 * p8=0$
$p2 * p6 * p8=0$

Item 1 in the algorithm ensures that the pixels that have only one neighbor or have seen or more are not deleted. If a

pixel has only one neighbor, it would be at the end of skeletal line and should not be deleted. If a pixel has seven neighbors, then deleting it would starts unacceptable erosion into the object's shape. This item thus ensures that the basic that the basic shape of the object is kept by the skeleton [2].

TABLE I
SHAPES OF OTTOMAN SCRIPTS

Isolated	Initial	Medial
ا	-	
ب	بـ	ـبـ
ت	تـ	ـتـ
ث	ثـ	ـثـ
ج	جـ	ـجـ
ح	حـ	ـحـ
خ	خـ	ـخـ
د	-	
ذ	-	
ر	-	
ز	-	
س	سـ	ـسـ
ش	شـ	ـشـ
ص	صـ	ـصـ
ض	ضـ	ـضـ
ط	طـ	ـطـ
ظ	ظـ	ـظـ
ع	عـ	ـعـ
غ	غـ	ـغـ
ف	فـ	ـفـ
ق	قـ	ـقـ
ك	كـ	ـكـ
ل	لـ	ـلـ
م	مـ	ـمـ
ن	نـ	ـنـ
ه	هـ	ـهـ
و	-	
ي	يـ	ـيـ

In Fig. 1 an example of Ottoman scripts is given. By using this example script thinning, segmentation and other step of the study will be shown.

In Fig. 2 given example script is thinned by using Zhang-Suen algorithm. After thinning of the example image, image is ready for segmentation step and the next step is applied.

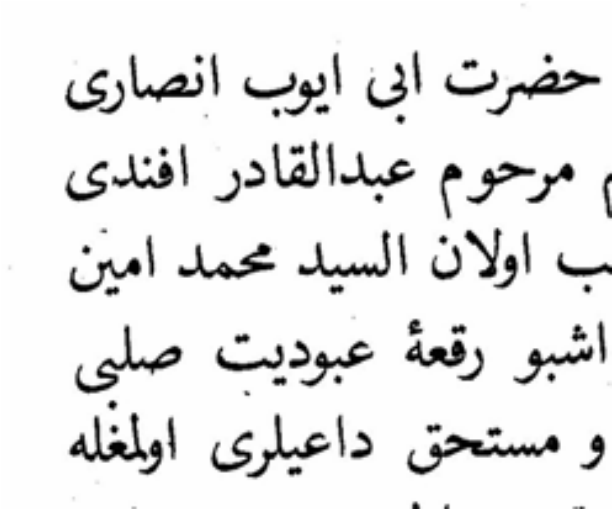


Fig. 1 An example of Ottoman Script

In Fig. 2 Zhang-Suen's Skeletonization Algorithm applied to the image.

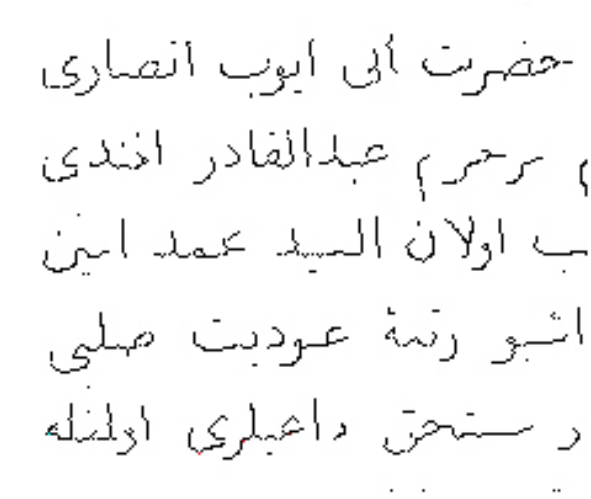


Fig. 2 Thinned image of example script

V. SEGMENTATION

The segmentation is the one of important phase in recognizing handwritten Ottoman text. If there is any error in segmenting, the basic shape of Ottoman characters will produce a different representation of the character component. [4]

The segmentation stages are based on geometrical feature analysis as in the example. First the scripts are separated to the lines, then words. After that, the words separated to the characters. After choosing a word or a character, for simplification the place of it is painted with white color. Therefore selection of characters or words becomes easy. This is shown in Fig. 3.

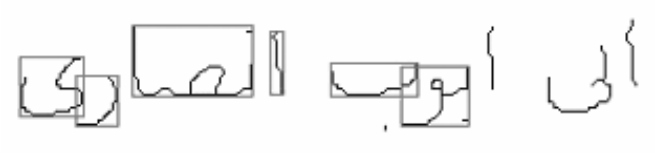
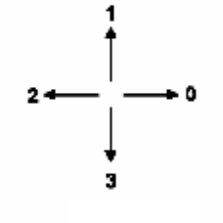


Fig. 3 Segmented image

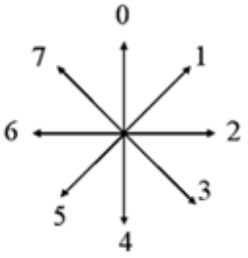
While separating the characters number of neighbors is checked. If the number of the neighbors is bigger than 3 words are separated to the characters. After the separating of words to characters, chain codes obtained for each of the character.

VI. Chain Codes

Chain codes are used to represent a boundary by a connected sequence of line segments of specified length and direction. The chain codes are taken for the features of characters. Representation is based on 4 or 8 connectivity segments. In this paper 8 connectivity is used, because when working on OCR data loss should be minimum. Therefore 8 connectivity referred.



Direction for 4- connectivity



Direction for 8- connectivity
Fig. 4 4 connectivity and 8 connectivity

In Fig. 5 an example for chain code representation is shown.

In the example in Fig. 5, after applying the chain code, chain code string "44444646606000200222" is obtained.

VII. Hidden Markov Models

Hidden Markov Models (HMMs) are finite state machines and powerful statistical models for modeling sequential or time-series data, and have been successfully used in many tasks such as speech recognition, protein/DNA sequence analysis, robot control, and information extraction from text data. [5] The HMM is called "Hidden" Markov Model because the process affects the observed sequence of events.

Each HMM consists of a number of states. When the model is in state, the process may be measured and one of the symbols may be produced, according to an observation probability distribution. At each time step, the model will undergo a transition to a new state, according to a transition probability distribution. [6] These transitions occur between any two states which contains self-transitions. The transitions will continue until the transition probability distribution is non-zero. In this study, discrete left-to-right HMM is used as shown in Fig. 6.

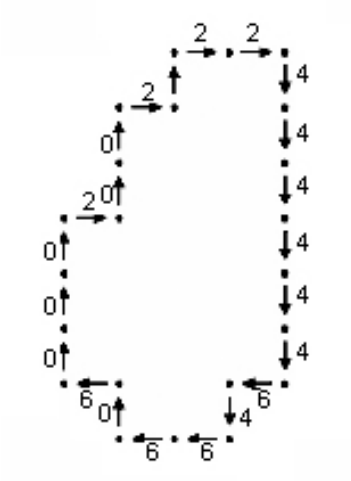


Fig. 5 Chain code representation

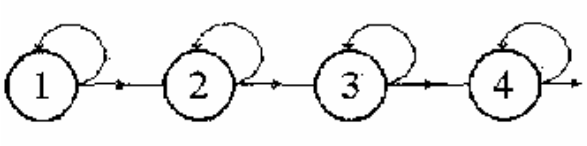


Fig. 6 Left to right Hidden Markov Model

VIII. Conclusion

In this paper, the proposed system was built with C++ language. Firstly, taken pictures were preprocessed for preparing of some function. Then, the scripts are separated to the words and characters. HMM techniques were used for testing this system using the simplified and traditional Ottoman fonts. We had a database which contained the beginning, middle and ending of the characters. The system provided accuracy approximately 65%.

References

[1] Lorigo Liana M., *Offline Arabic Handwriting Recognition*, IEEE Transactions On Pattern Analysis and Machine Intelligience,Vol 28, No 5, May 2006.

[2] Mcandrew Alas Dair, *Digital Image Processing with Matlab,* Thomson Course Technology.

[3] Atıcı Alper, *Segmentation, Feature Extraction and Recognition of Ottoman Script*, September 1994.

[4] Motawa Deya, Amin Adnan and Sabourin Robert, *Segmentation of Arabic Cursive Script.*

[5] ChengXiang Zhai, *A Brief Note on the Hidden Markov Models (HMMs),* March 16, 2003.

[6] Alaa M.Gouda, M.A.Rashwan, *Segmentation of Connected Arabic Characters Using Hidden Markov Models* , CIMSA 2004 IEEE lnternetional Conference *on* Computational Intelligence for Measurement Systems and Applications Baston, YD, USA, 14-16 July 2004.