Retrieval of Ottoman Documents

Esra Ataer
Bilkent University
Department of Computer Engineering
Ankara, TURKEY
ataer@cs.bilkent.edu.tr

Pinar Duygulu
Bilkent University
Department of Computer Engineering
Ankara, TURKEY
duygulu@cs.bilkent.edu.tr

ABSTRACT

There is a growing need to access historical Ottoman documents stored in large archives and therefore managing tools for automatic searching, indexing and transcription of these documents is required. In this paper, we present a method for the retrieval of Ottoman documents based on word matching. The method first successfully segments the documents into word images and then uses a hierarchical matching technique to find the similar instances of the word images. The experiments show that even with simple features promising results can be achieved.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*document analysis*

General Terms

Documentation, Experimentation

Keywords

word-image matching, projection profiles

1. INTRODUCTION

Large archives of historical documents are now available online with the developments in electronic imaging. For these documents to be efficiently and effectively accessible to the scholars, creation of managing tools for automatic indexing, searching and transcription is important.

Ottoman Empire, which had lasted over six centuries and covered a large area including many different cultures, has left a large collection of valuable documents interesting to historians from all over the world [12]. However, access to important Ottoman documents is very limited, since many documents are in defective editions or in manuscript format and manual transcription and indexing of Ottoman texts requires a lot of time and effort. Therefore, it is important to build automatic systems to search and transcribe these documents.

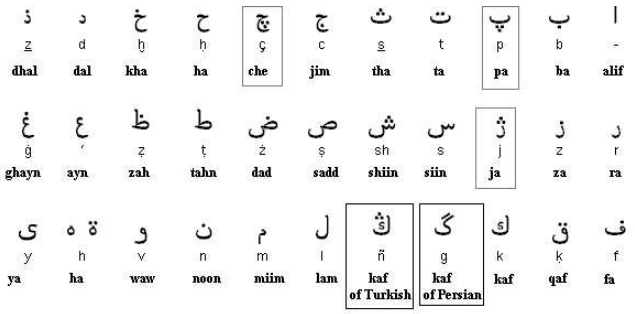

Figure 1: Characters in Ottoman Alphabet. Ottoman alphabet has 5 more additional characters than Arabic alphabet which consists of 28 basic characters. Ottoman characters which are different from Arabic are indicated with bounding rectangles.

In this paper, a method for searching Ottoman documents based on word matching is presented. We used printed type of Ottoman scripts, which is more regular and justified than the other styles. First, we segment the documents into lines and then into words. Line extraction is performed by finding the baselines using horizontal projection profiles and then by specifying the upper and lower limits for the characters relative to the baselines. Then, words are extracted using the vertical projection profiles. After the word extraction part, we perform a hierarchical matching technique for retrieving similar words. The matching method is composed of four consecutive tests, namely (i)length similarity; and similarity of quantized vertical projection profiles (ii) for the entire words, (iii) for the ascender part of the words and (iv) for the descender part of the words. The experimental results on printed documents show that most of the words can be retrieved correctly with the proposed approach.

The paper is organized as follows. The characteristics of Ottoman scripts are described in Section 2. Section 3 presents an overview of the related studies. Then, in Section 4 proposed approach is explained. Experimental results are provided in Section 7. Finally, we conclude the paper and discuss possible future directions in Section 8.

MIR'06, October 26–27, 2006, Santa Barbara, California, USA.

2. CHARACTERISTICS OF OTTOMAN SCRIPTS

The Ottoman script is a connected script based on Arabic alphabet with additional vocals and characters from Persian and Turkish languages [9] (see Figure 1). Similar to Arabic, in Ottoman scripts each character can have four different formats according to the position of the character in the word (beginning, middle, end and isolated). Another common property of Ottoman and Arabic is that they include only a few vowels. Therefore, transcription of a word is strongly based on the context of the document and vocabulary of the reader. Sometimes two different words can be written as the same, but suitable one is selected according to the context of the document.

Ottoman calligraphy was a respected and encouraged art during Ottoman Empire and therefore many calligraphy styles are found and improved by Ottomans [2, 3, 8, 19]. Some examples of different calligraphy styles are shown in Figure 2. As can be seen, some letters can be skewed or elongated in different writing styles making the segmentation and the recognition process more complicated.

3. RELATED WORK

Although character recognition is a well studied area [5, 10, 20, 18, 17], there are not many studies on recognition of Arabic characters [1, 6, 4, 3] and recognition or retrieval of Ottoman documents is almost untouched other than a few studies [13, 15, 16, 21].

Recently, Rath and Manmatha [14] proposed a word-image matching technique for retrieval of historical documents by making use of dynamic time warping and show that the documents can be accessed effectively without requiring recognition of characters with their word spotting idea. They use intensity, background-ink transition, lower and upper bound of the word as the features for matching process.

Edwards et al. [7] described a generalized HMM model in order to make a scanned Latin manuscript accessible to full text search. The model is fitted by using transcribed Latin as a transition model and each of 22 Latin letters as the emission model.

Chan et al. [6] presented a segmentation based approach that utilizes gHMMs with a bi-gram letter transition model. Their lexicon-free system performs text queries on off-line printed and handwritten Arabic documents.

Saykol et al. [15] used the idea of compression for content-based retrieval of Ottoman documents. They create a code book, for the characters and symbols in the dataset and processed the queries by the help of this codebook. Scale invariant features named distance and angular span are used in the formation of the codebook.

4. PROPOSED METHOD

Our method is composed of two main stages: segmentation and matching. Segmentation aims to extract the words from the document, and includes line segmentation and word segmentation steps. In matching, each word is queried and relevant images are retrieved in ranked order according to a set of similarity criteria.

Since the experiments are carried out on printed, relatively clean documents, a binarization step based on simple thresholding can produce acceptable results for further pro-

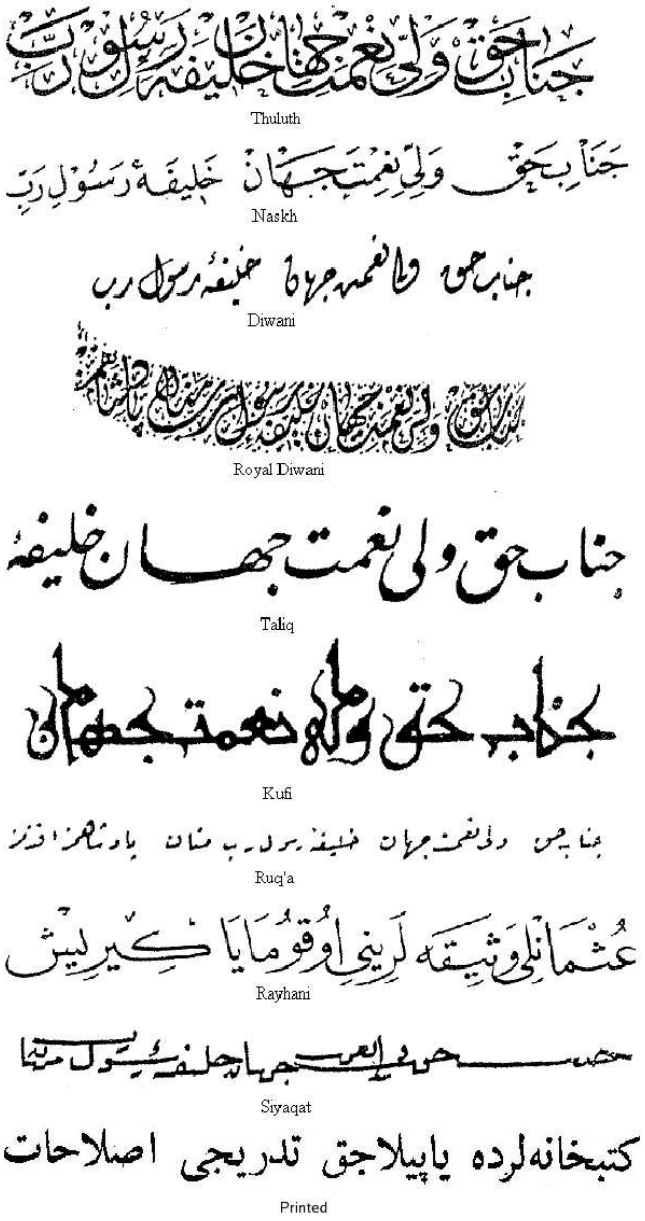

Figure 2: Some Calligraphy Styles in Ottoman Script. The last row is a printed (matbu) type writing style, which we used in our experiments.

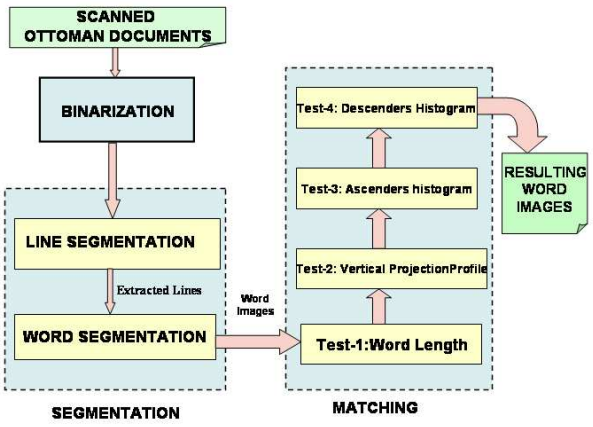

Figure 3: System Overview.

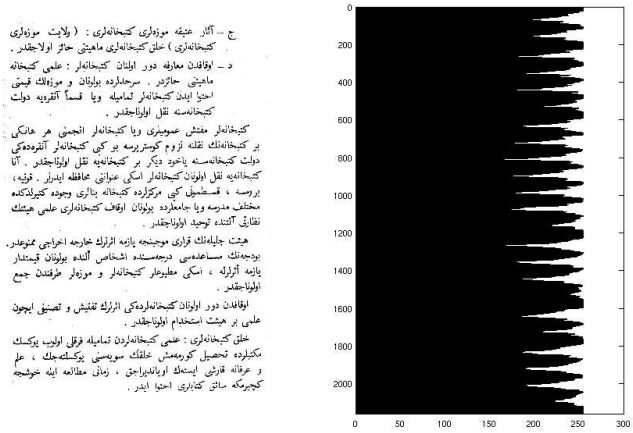

Figure 4: Horizontal projection profile of a document

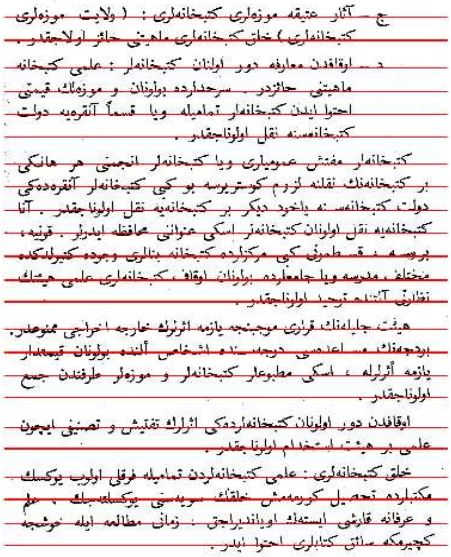

Figure 5: Computed baselines of the document

cessing. Also, rectification is not necessary since the documents are scanned carefully.

The system overview is shown in Figure 3. In the following sections, the segmentation and matching steps will be described in detail.

5. SEGMENTATION

Segmentation is performed in two consecutive steps: line segmentation and word segmentation. Both steps make use of the projection profiles.

5.1 Line Segmentation

Each line has a baseline on which most of the characters lies. For extracting lines in Ottoman scripts, we observed that finding positions of the baselines and then separating the lines based on the character sizes is better than finding the spaces between lines for separation. That is due to the characteristics of Ottoman alphabet having some letters with long descender and ascender parts causing the space between two lines in variable length.

Baselines have black pixels more than the other rows and the number of black pixels on the baselines are almost equal to the width of the document. With these assumptions, we first get the horizontal projection profiles and find the lines which have black pixels greater than some threshold. This process allows us to find a set of rows which are candidate positions for a baseline. Then, among the candidate positions for the baselines, the local maximum having the largest number of black pixels is chosen as the final baseline.

In the experiments, the documents are selected from a single source and therefore the characters have fixed sizes throughout all documents allowing us to set a maximum height for the characters both for the ascending and the descending part of the baseline (48 and 37 pixels respectively for our dataset). These values are used to determine the upper and lower limits of a line and separate it from the others according to the position of the baseline.

Due to thresholding, we can only find the complete lines -which starts at the rightmost part and continues until the leftmost part-, but not the lines which end in the middle. Therefore, those lines remain unextracted. As a postprocessing step, we apply another thresholding on unextracted parts of the document and check whether this part has sufficient length and black pixel ratio to be a line.

For an example document, horizontal projection profile is given in Figure 4 and computed baseline positions of that document are shown in Figure 5.

5.2 Word Segmentation

As the next step, the extracted lines are segmented into words. To find the boundaries between the words, we apply a threshold value on the length of the space in between the words. After finding the positions of the spaces between words we also eliminate the parts of the line segment, which do not include any letters such as the noisy areas at the beginning or at the end of a line. An example line segment and extracted words are shown in Figure 6. As a final step, we eliminate the white borders around each word image in order to make feature extraction more easier.

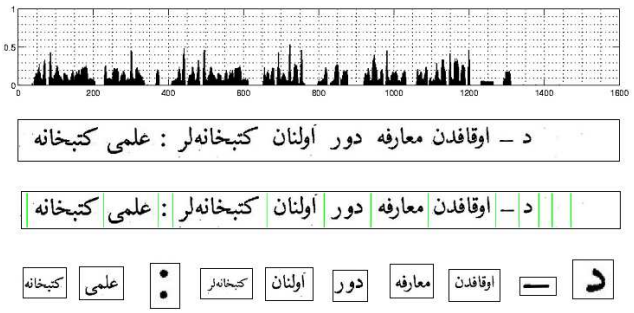

Figure 6: Word Extraction: First image is vertical projection profile of the line shown in second image. Third is the split points for that line. Last row is the word images that are extracted from that line.

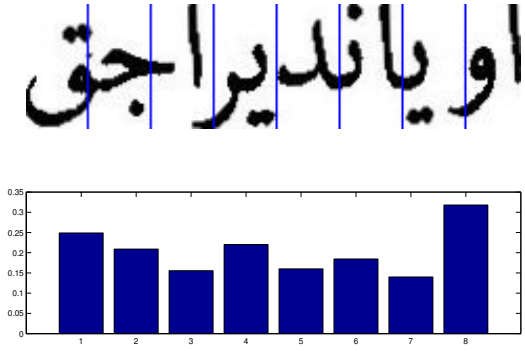

Figure 7: Quantized Vertical projection profile of the word *'uyandiracak'* **meaning** *will wake up*

6. MATCHING

For finding the similar instances of a word, the query word is matched with the other words in the data set. Word matching is performed through four consecutive tests:

1. length of the word,
2. quantized vertical projection profile,
3. quantized vertical projection profile of ascenders, and
4. quantized vertical projection profile of descenders.

In each test, some of the irrelevant images are discarded from the resulting set and a smaller set is formed for the next test. The final remaining set is then considered as the correctly retrieved set of images.

In test1, the word images which have a length difference more than a threshold value are discarded. The threshold is taken as 15 pixels, which is approximate length of a character in the alphabet. We used length of a word instead of scale ratio of the word, because we deal with the documents that have scripts of same size.

For test2, before extracting vertical projection profile (VPP) feature, first we filter the image with a Gaussian function and then downsample it to its half size in order to eliminate the noise on the image.

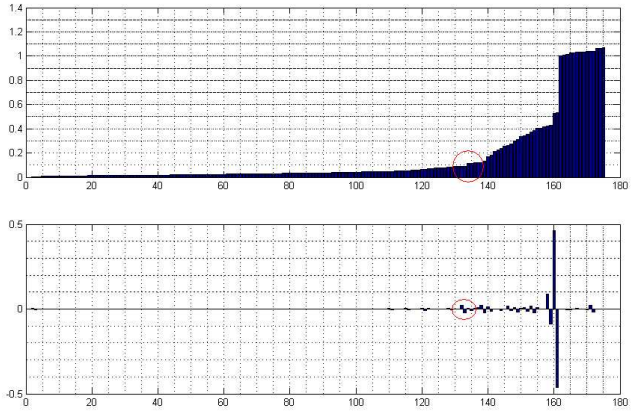

Figure 8: An example histogram of sorted distances between vertical projection profiles. Below is second derivative of first one. Peaks are indicated with red circles.

Then, we segment the word image into fixed size bins, and obtain a quantized vertical projection profile for these bins as the ratio of black pixels for each bin of the word image. In our system, we take the bin size as 15 pixels, which is the approximate length of a character in the alphabet. An example vertical projection profile is shown in figure 7.

During second test, Euclidean distance between the VPP of the query image and the remaining dataset after the first test is computed and sorted. The first largest difference of the sorted distances is taken as the threshold value which is then used to eliminate the words that do not have similar VPPs. We find the threshold value by looking at the second derivative of the sorted distances as in Figure 8.

Ascender is the part of the word, which remains in the upper part of the baseline, while descender is the lower part of baseline (see Figure 9). Similar to Test 2, in test3 and test4 we find the vertical projection profiles of ascender and descender parts as in Figure 9 and used them for further elimination.

7. EXPERIMENTAL RESULTS

The experiments are carried out on a dataset of six printed documents, which are the official letters about the arrangements of the government libraries in the early stages of the Turkish Republic [11]. An example document and its transcription is shown in Figure 10. Figure 11 shows the distribution of all the words in the dataset. Since the documents are on the same subject, some words have high occurrences.

7.1 Segmentation results

There are in total 99 lines in the entire dataset and the proposed line extraction method works with 100% accuracy as shown in Table 1. The results of word extraction process is shown in Table 2. In total, 823 word images out of 946 are extracted successfully, resulting in 82% average word extraction performance. The difference in the performances for different documents is due to the different noise levels of the documents.

Since Ottoman Turkish have some phrases, and the space in between the words of the phrases can be small, the thresh-

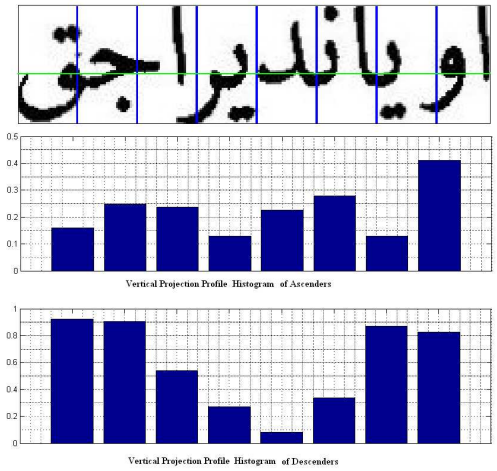

Figure 9: Baseline of the example word *'uyandiracak'* is seen in the first part. Horizontal green line indicates baseline of the word and vertical blue lines indicate the bins of projection profile. Second and third parts are the histograms for vertical projection profile of ascenders and descenders respectively.

هيئت جليلەنك قرارى موجبنجه يازمه اثرلرك خارجه اخراجى ممنوعدر.
بودجەنك مساعدەسى درجەسنده اشخاص ألنده بولونان قيمتدار
يازمه أثرلرله ، اسكى مطبوعلر كتبخانەلر و موزەلر طرفندن جمع
اولوناجقدر .

Heyet-i Celile' nin kararı mucibince yazma eserlerin harice ihracı memnudur.
Bütçenin müsadesi derecesinde eşhas elinde bulunan kıymetdar yazma
eserlerle, eski matbular kütüphaneler ve müzeler tarafından cem olunacaktır.

Figure 10: An example document and its transcription. The document is about the arrangement for printed and manuscript documents in the libraries.

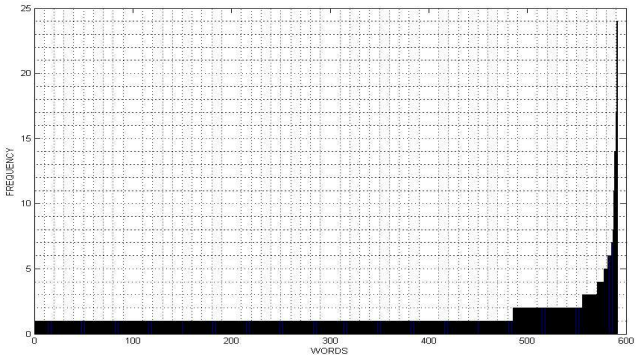

Figure 11: The frequency of the word images in the data set in sorted order.

Table 1: Line Segmentation Results

Doc. No.	Number of Lines	Number of Segmented Lines	Success
1	23	23	100%
2	17	17	100%
3	16	16	100%
4	12	12	100%
5	16	16	100%
6	15	15	100%
Total	99	99	100%

Table 2: Word Segmentation Results

Doc. No.	Number of Words	Number of Segmented Words	Error	Success
1	186	166	6	86%
2	129	118	7	86%
3	182	156	9	80%
4	110	99	9	82%
5	185	157	10	80%
6	154	127	9	77%
Total	946	823	50	82%

old value may not segment the phrase into words and may take the phrase as a single word. Besides, isolated format of some consecutive characters can result in large gaps in a word. Thus some words can be segmented wrongly. Another reason for the errors is the dots or tails of letters. In Figure 12, some words which are segmented wrongly are shown. Figure 13 shows the distribution of word lengths in pixels for the correctly segmented ones.

7.2 Matching results

Each word in the data set is queried using the proposed matching scheme. We use mean Average Precision (mAP) values for evaluating the performance of the word queries. Figure 14 shows mAP values for some selected queries. For all the words in the data set the average mAP value is obtained as 0.8524.

In Figure 15, the results for the retrieval of the most popular words are shown. The black dots indicate the relevant documents among all which are ranked according to the similarity of the documents. As can be seen, for most of the words all of the relevant documents are retrieved correctly. Some example retrieval results are shown in Figure 16. In the figure, the first one shows a successfull query of a word, while the second one shows a successfull query of a phrase, which is composed of two words. The third one is the query of the word *'libraries'* and the 8th and 10th images retrived means *'in the library'* and *'to the library'* showing that sometimes the words with having similar meanings can be retrieved. Fourth one is the retrieval of a stop word meaning *'and'*. Shortness of the word length causes irrelevant results for that query. The last one is an example query with many irrelevant results.

8. SUMMARY AND DISCUSSION

In this paper, we proposed a novel approach for searching Ottoman documents. Ottoman script is a connected script, which is difficult to segment and recognize.

In the proposed system, firstly the scanned documents

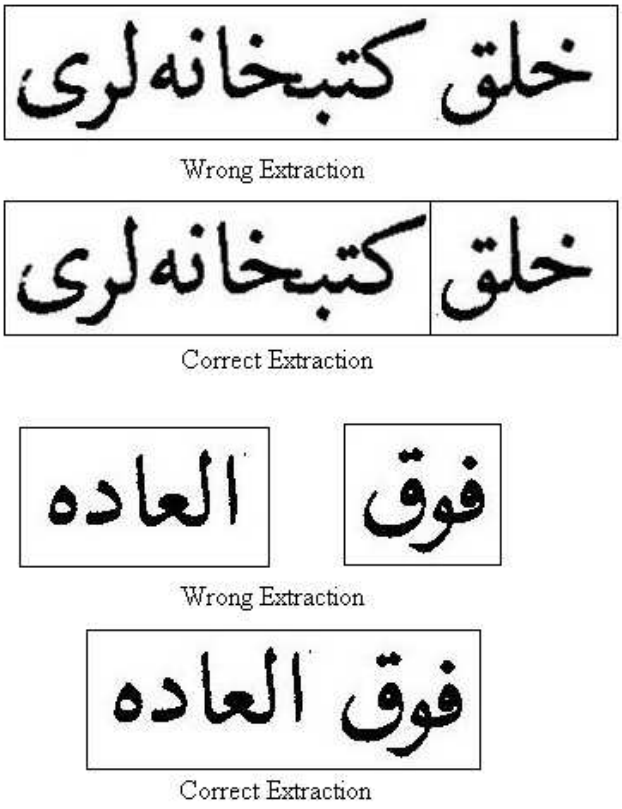

Figure 12: Two examples of word extraction errors. In the first one the phrase *'halk kutuphaneleri'*, meaning *Public Libraries*, could not be splitted into two words because of the tails of the letters near spaces. A word is splitted into two parts in the second example. The second error comes from the isolated letters that create a large gap.

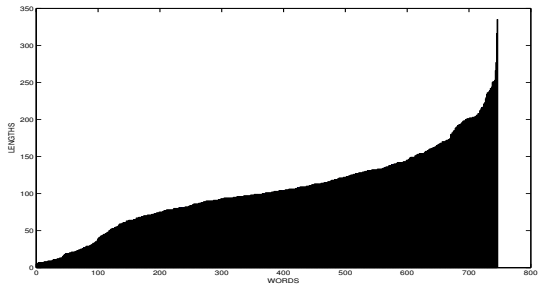

Figure 13: Distribution of word lengths in pixels for correctly segmented word images.

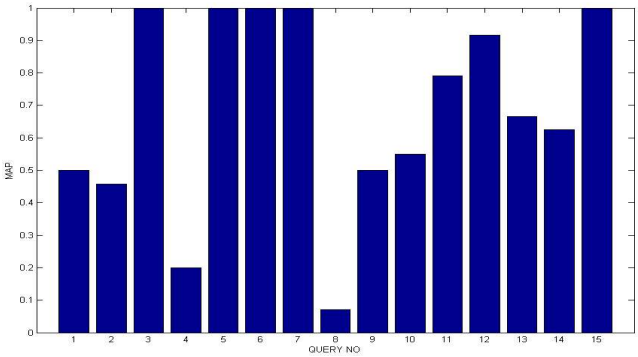

Figure 14: mAP values for the selected 15 queries. Smaller mAP values comes from the short length words like and, one etc.

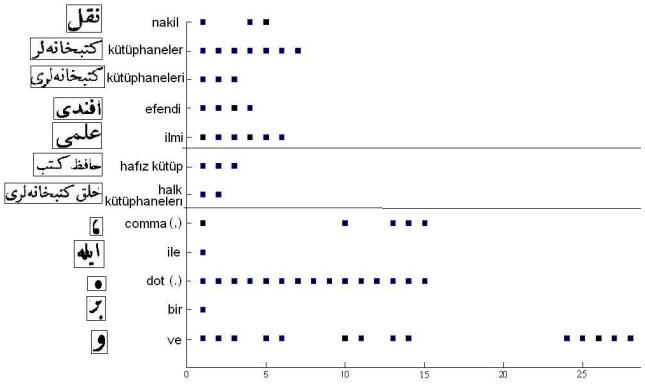

Figure 15: Retrieval results for the most popular words in the dataset. X axis shows the order of relevant documents retrieved and Y axis represents the words and their transcriptions. The words in the second part are phrases that are composed of two words. The words in the third part are stop words like and, with, dot and comma.

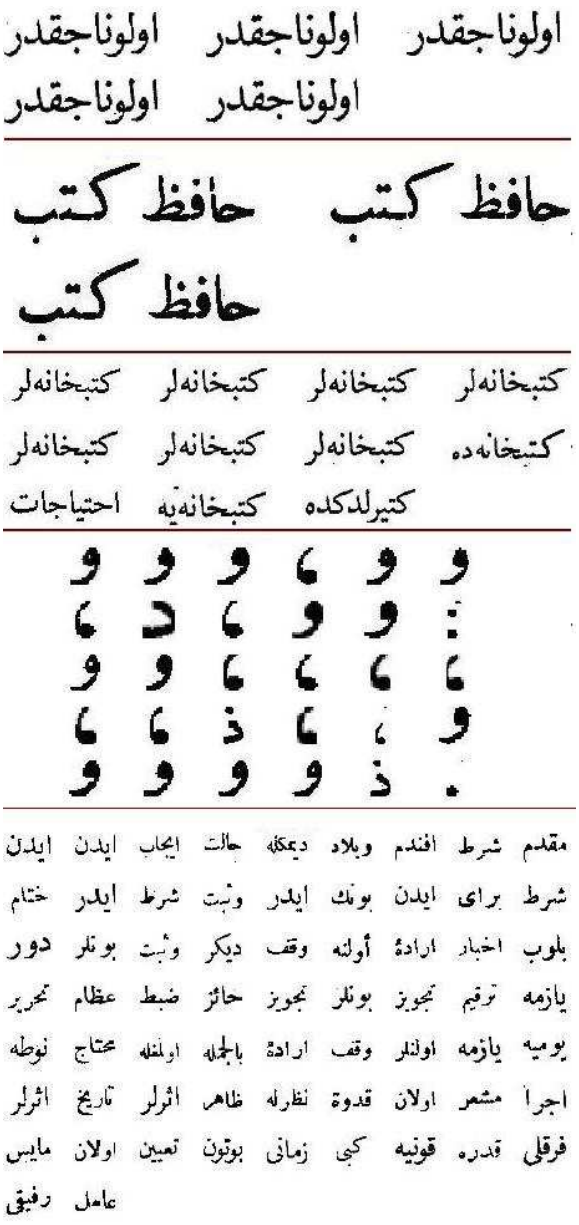

Figure 16: Some query results. The image in the upper left is the query image and the remaining ones are the resulting images.

are passed from a binarization process. Then, the system makes use of thresholding on horizontal and vertical projection profiles to segment lines and words recpectively. In the matching stage, segmented words are queried and retrieval is performed with the use of four distinctive features: word length, quantized vertical projection profile and quantized vertical projection profiles of ascenders and descenders. These four features are used in four consecutive tests and each test discards the dissimilar word images for the next test.

Although we have primitive features for matching, we acquire the relevant word images in the first orders of the ranking. This shows that, if we use better features for word matching, we would have more accurate results. As a future extension of this approach, we aim to have improved set of features by including some shape features and compute the occurrence of each word image in the dataset. Thus, we would match the popular word images in the dataset with the popular words in the transcribed document set resulting in the transcription of the most popular words in the dataset. Consequently, this research serves a basis for future researches about transcription of Ottoman scripts with its succesfull word matching results even if using some primitive features.

9. REFERENCES

[1] B. Al-Badr and S. A. Mahmoud. Survey and bibliography of arabic optical text recognition. *Signal Processing, v.41 n.1, p.49-77.*, Jan. 1995.

[2] A. Alparslan. *Osmanli Hat Sanati Tarihi*. Yapi Kredi Yayinlari, 1999.

[3] A. Amin. Off-line arabic character recognition: the state of the art. *Pattern Recognition*, 31(5):517–530, 1998.

[4] A. Amin, G. Masini, and J. Haton. Recognition of handwritten arabic words and sentences. In *ICPR84*, pages 1055–1057, 1984.

[5] C. S. Belongie. Shape matching and object recognition using shape.

[6] J. Chan, C. Ziftci, and D. Forsyth. Searching off-line arabic documents. In *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, June 17-22 2006.

[7] J. Edwards, Y. W. Teh, D. A. Forsyth, R. Bock, M. Maire, and G. Vesom. Making latin manuscripts searchable using ghmms. In *NIPS*, 2004.

[8] M. Eminoğlu. *Osmanli Vesikalarini Okumaya Giris*. Turkiye Diyanet Vakfi Yayinlari, 2003.

[9] N. Gök. Osmanlicayi herkes kolayca ogrenebilir mi. *Zaman*, 2004.

[10] S. Impedovo, L. Ottaviano, and S. Occhinegro. Optical character recognition - a survey. *International Journal of Pattern Recognition and Artificial Intelligence*, 5:1–24., 1991.

[11] Y. Kurt. *Osmanlica Dersleri 1*. Akcag Yayinlari, 2000.

[12] Otap: Ottoman text archive project, http://courses.washington.edu/otap/.

[13] A. Ozturk, S. Gunes, and Y. Ozbay. Multifont ottoman character recognition. In *The 7th IEEE International Conference on Electronics, Circuits and Systems (ICECS 2000)*, volume 2, pages 945–949, December 17-20 2000.

[14] T. M. Rath and R. Manmatha. Word image matching using dynamic time warping. In *CVPR (2)*, pages 521–527, 2003.

[15] E. Saykol, A. K. Sinop, U. Gudukbay, O. Ulusoy, and A. E. Cetin. Content-based retrieval of historical ottoman documents stored as textual images. *IEEE Transactions on Image Processing*, 13(3):314–325, 2004.

[16] A. Sisman and F. Yarman-Vural. Ottoman transcription system. In *ISCIS-IX*, 1996.

[17] S. Srihari. Off-line arabic character recognition: the state of the art. *Pattern Recognition*, 31(5):517–530, 1998.

[18] Suen, C. Y., Berthod, Marc, Mori, and Shunji. Automatic recognition of handprinted characters - the state of the art. In *Proceedings of the IEEE 68 (4)*, pages 469–487, 1980.

[19] M. Ülker. *Baslangictan Günümüze Türk Hat Sanati*. Turkiye Is Bankasi Kültür Yayinlari, 1987.

[20] J. R. Ullmann. Advance in character recognition. In *Application of Pattern Recognition*, pages 197–236, 1982.

[21] F. Yarman-Vural and A. Atici. A segmentation and feature extraction algorithm for ottoman cursive script. In *The Third Turkish Symposium on Artificial Intelligence and Neural Networks*, June 1994.