Matching Ottoman Words: An image retrieval approach to historical document indexing

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

Large archives of Ottoman documents are challenging to many historians all over the world. However, these archives remain inaccessible since manual transcription of such a huge volume is difficult. Automatic transcription is required, but due to the characteristics of Ottoman documents, character recognition based systems may not yield satisfactory results. It is also desirable to store the documents in image form since the documents may contain important drawings, especially the signatures. Due to these reasons, in this study we treat the problem as an image retrieval problem with the view that Ottoman words are images, and we propose a solution based on image matching techniques. The bag-of-visterms approach, which is shown to be successful to classify objects and scenes, is adapted for matching word images. Each word image is represented by a set of visual terms which are obtained by vector quantization of SIFT descriptors extracted from salient points. Similar words are then matched based on the similarity of the distributions of the visual terms. The experiments are carried out on printed and handwritten documents which included over 10,000 words. The results show that, the proposed system is able to retrieve words with high accuracies, and capture the semantic similarities between words.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*document analysis*; I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement—*feature representation*

General Terms

Documentation, Experimentation

Keywords

word-image matching, bag-of-features, indexing

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

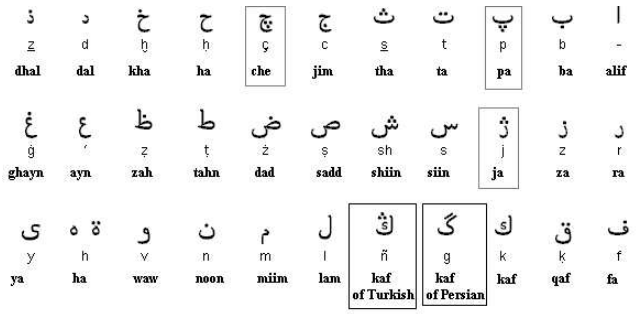

Figure 1: Characters in Ottoman Alphabet. Ottoman alphabet has 5 more additional characters than Arabic alphabet which consists of 28 basic characters. Ottoman characters which are different from Arabic are indicated with bounding rectangles.

1. INTRODUCTION

Large archives of historical documents remain inaccessible due to the huge amount of work required for manual annotation or transcription and the difficulty in building automatic systems.

Ottoman archives, as being one of the largest collection of historical documents, hold over 150 million documents ranging from military reports to economic and political correspondence, belonging to the Ottoman era. Large number of researchers from around the world are interested in accessing the archived material [13]. However, many documents are in defective editions or in manuscript format and manual transcription and indexing of Ottoman texts require a lot of time and effort, causing most of these documents inaccessible.

The Ottoman script is a connected script based on Arabic alphabet with additional vocals and characters from Persian and Turkish languages [8] (Figure 1) and therefore shares the difficulties faced in Arabic [4]. Similar to Arabic, in Ottoman scripts each character can have four different forms according to the position of the character in the word (beginning, middle, end and isolated). Another common property of Ottoman and Arabic is that there are only a few vowels. Therefore, transcription of a word strongly depends 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.

Although character recognition is a well studied area [9, 27, 26], there are not many studies on recognition of Arabic characters [1, 4, 2, 10] and recognition or retrieval of Ottoman documents is almost untouched other than a few studies [16, 20, 22, 3]. Recogni-

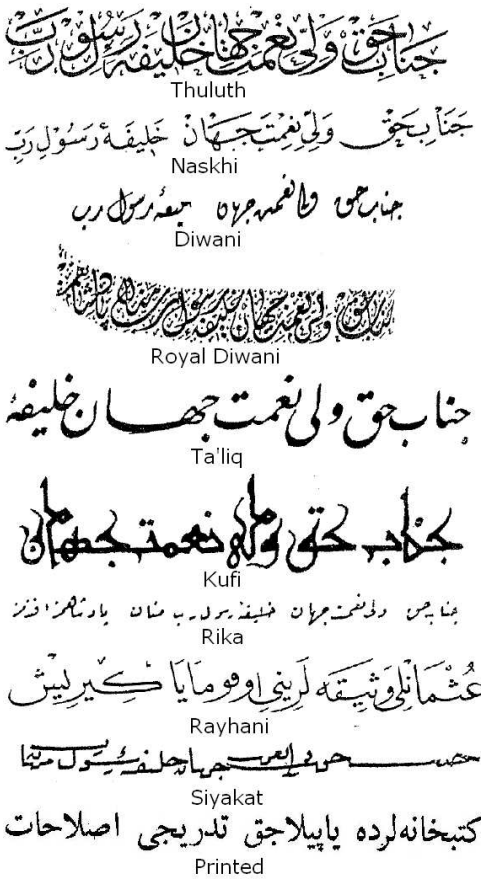

Figure 2: Some calligraphy styles in Ottoman script.

tion of Ottoman characters is difficult, if not impossible, with the traditional character recognition systems, due to the skewed and elongated characteristics of writing.

The main difficulty of Ottoman documents is due to the different calligraphy styles heavily used in writing. Ottoman calligraphy was a respected and encouraged art during Ottoman Empire and therefore many calligraphy styles were founded and improved by Ottomans [6]. Some examples of different calligraphy styles are shown in Figure 2.

In Ottoman documents, different types of drawings, such as miniatures and signatures are also widely used. Some royal documents which are called as Firmans [15] are shown in Figure 3. As can be seen, the documents are very complicated and they look like drawings rather than simple writings as aimed in calligraphy.

These characteristics of Ottoman documents suggest us to treat the problem as an image retrieval problem rather than a traditional character recognition problem. With the proposed method, the documents are first segmented into words, which are considered as images rather than collection of characters, and then words are retrieved based on their visual descriptors using the image matching techniques.

With the assumption that the curvature or connection points, or the dots which are commonly used in Ottoman characters, are important visual features, we make use of the salient points which are shown to be successful in representing such distinctive areas and heavily used for matching. Specifically, we used SIFT operator for detecting and representing salient points [11]. Then, we use the

Figure 3: Example Firmans (royal documents)

bag-of-words approach which is recently adopted from statistical text analysis literature for classifying objects and scenes [24, 17, 23]. The descriptors extracted from the salient points are vector quantized to have a set of visual terms and then the words are represented as collections of these visual terms. Then, the similarity of words are found based on the similarity of the visual term distributions.

In the following, first some related studies on character recognition and document retrieval are reviewed. After describing the data set, and explaining the preprocessing methods for obtaining words, the proposed method for matching words will be discussed. Finally, detailed experimental results will be presented and the results will be compared with Dynamic Time Warping approach which is used for a similar problem.

2. RELATED WORK

In [21], matching on bywords is proposed by using different combinations of five feature categories: angular line features, cocentric circle features, projection profiles, Hu's moment and geometric features. Another study [7] make use of atomic segments and Viterbi algorithm in order to create an Arabic text recognition system.

Edwards *et al.* [5] 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.* [4] 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.* [20] used the idea of compression for content-based retrieval of Ottoman documents. They create a code book, for the characters and symbols in the data set 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.

Recently, Rath and Manmatha [18] 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. They make experiments with seven different methods, which showed that their projection based DTW method results in

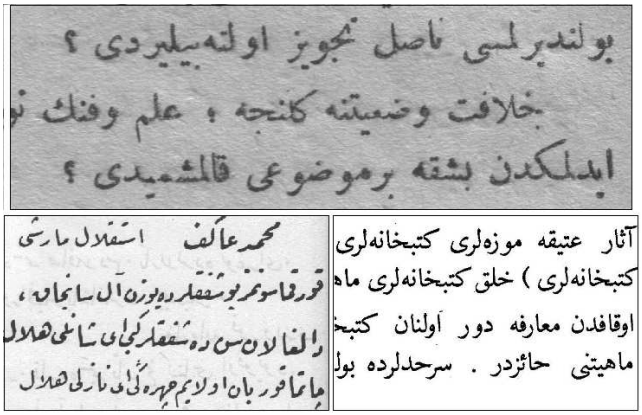

Figure 4: Example documents used in the study: top: large-printed, bottom left: rika, bottom right: small-printed.

Figure 5: Some examples of word extraction errors. Red boxes are the wrong extractions and green boxes are the correct extractions. One reason for the segmentation errors is the phrases with very small gaps in between the words. This situation is seen in the phrase at the upper left corner, meaning *Public Libraries*, which could not be split into two words because of the tails of the letters near spaces. Similarly, the words at the bottom left are also unextracted due to elongated letters. On the other hand, isolated format of some consecutive characters may result in large gaps causing over-segmentation as seen in the words at the right.

the highest precision [12]. In their another study [19] they used a Hidden Markov Model based automatic alignment algorithm in order to align text to handwritten data. Experiments are done on the same dataset as before and they claim that this algorithm outperforms the previous DTW approach.

Srihari *et al.* [25] proposed a system using word matching idea after a prototype selection step. They make use of 1024 binary features in word matching step and acquired promising results for a dataset with various writers.

In our previous study [3], Ottoman words are matched using a sequence of elimination techniques which are mainly based on vertical projection profiles.

3. DATA SETS

In this study, we used three different sources to construct our data sets (Figure 4). Two of these sources are in printed form. The first one, referred as **small-printed**, is a relatively small data set consisting of 6 documents about arrangements of the libraries. This data set, consisting of 823 words, is manually transcribed for automatic evaluation. The second data set, referred as **large-printed**, consists of 9524 words extracted from the first 25 pages of a book. This data set is used to show that the proposed method scales to large volumes. Although, both of these data sets are in printed form, the scale of the characters slightly differ. The third data set, referred as **rika**, consists of handwritten words which are written with a widely used calligraphy style named Rika, which is especially used in documents about governmental issues. It consists of 3 pages with 257 words and is also manually annotated for evaluation.

4. WORD SEGMENTATION

For extracting words from documents, simple and commonly used techniques are adopted. First, documents are binarized using the OTSU method [14]. Then, lines and words are segmented using smoothed horizontal and vertical projection profiles respectively. Since documents are carefully scanned, and the writing styles used are mostly written on straight lines, rectification is not required.

With the evaluations on small-printed data set, 100% line segmentation and 82% word segmentation performances are obtained. Figure 5 shows some errors in word segmentation. Only on rika data set, after automatic binarization and line segmentation steps, manual word segmentation is performed due to the difficulty of the data set.

We should note that better methods could be applied for preprocessing, but our focus is on representation of words after segmentation, and therefore in this stage we choose the simplest methods with the knowledge that better segmentation would result in better retrieval performance.

We should also mention that word segmentation errors could be tolerated with the proposed approach. For example, if a single word is wrongly segmented into two parts, it is likely that the subparts will be matched with the original word with relatively large scores since the proposed approach is able to capture the semantic relations between words which have common parts.

5. WORD MATCHING

Rather than further segmenting the words into characters and matching words based on consecutive characters as in traditional methods, in this study the words are considered as images and directly matched using visual features extracted from the entire words.

In this study, we represent the words using SIFT descriptors of the keypoints [11]. As Figure 6 shows, the keypoints usually correspond to distinctive areas on the characters, such as dots, high curvature or connection points.

Then, rather than matching words based on similarity of individual keywords, we prefer to use **bag-of-words** approach where the images are treated as documents with each image being represented by a histogram of visual terms. The visual terms, usually referred as **visterms**, are obtained by vector quantization of the feature vectors. We use k-means for vector quantization and represent the segmented words with the normalized distribution of visterms. The similarity of two words are then found by using symmetric KL-divergence of the two distributions.

In the following, we discuss a method to generate better visterms by making use of a codebook.

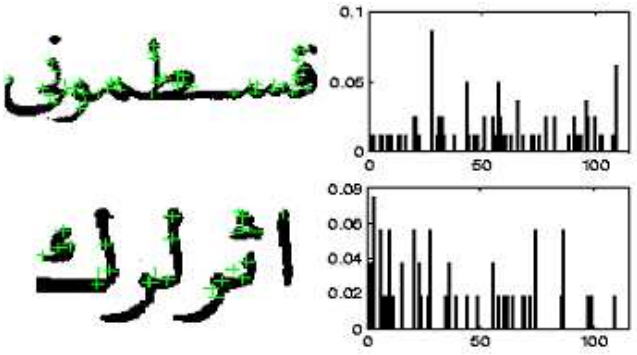

Figure 6: Example words with extracted keywords and visterm distributions.

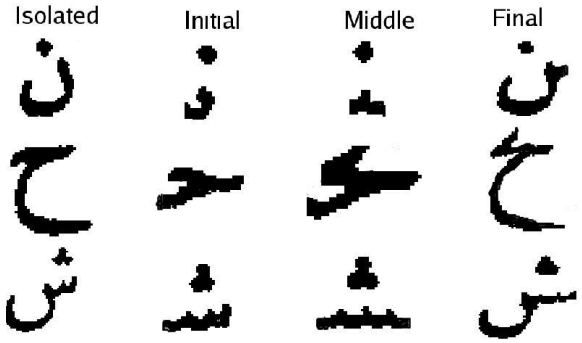

Figure 7: Four different forms of three example letters, nun, ha and shin respectively.

6. VISTERM GENERATION

Characters in Ottoman alphabet may have four different forms according to the position in the word: initial, middle, final, and isolated (see Figure 7). Usually, these different forms of the same character are very similar. Also, although some characters may include some common parts, the similarity among these different forms is higher than the similarity to other characters.

Based on these observations, we force the keypoints in different forms of the same character to be in the same cluster and the keypoints of different characters to be in separate clusters, by computing an error measure for different k values in k-means while obtaining the visterms.

For this purpose, a codebook of 117 elements which includes up to four different forms of 31 characters in the alphabet is created by also considering different connections to other characters. This codebook is then used to choose the best number of clusters, k, in k-means by minimizing the following error:

$$error = 1/C \sum_{i=1}^{C} c_i + 1/M \sum_{j=1}^{M} m_j \quad (1)$$

Here C is the number of characters in the alphabet, c_i is the number of clusters that a character c appears, M is the number of clusters, and m_j is the number of different characters that the cluster j includes. The error measure corresponds to the sum of the average number of clusters for a character with the average number of characters for a cluster. We assume that a character is in a cluster if a keypoint belonging to any form of that character is in that cluster.

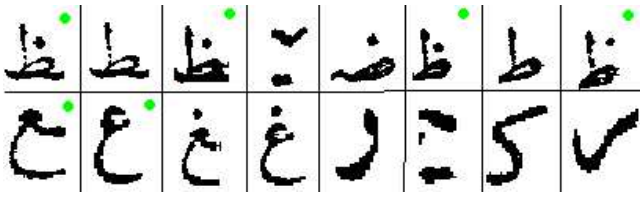

Figure 8: Example query results on codebook.

Using 1256 keypoints from the codebook, by incrementing k values by 10 between 20 and 200, the minimum error is achieved for $k = 110$.

In order to test the effectiveness of this method, we use the clusters obtained by the codebook and perform queries on each element in the codebook using the proposed method. Figure 8 shows two examples of character queries. As can be seen, different forms of the same character can be successfully matched, and the wrong matches are usually due to similar sub-patterns.

Although, this codebook could also be used for initializing the k-means, since the results were not very different than using random initialization, we prefer not to use the codebook further. In the experiments, only k=110 value is used to set the number of clusters, but the clusters are randomly initialized using the elements of the working data set.

7. EXPERIMENTAL RESULTS

The experiments are carried out on all of the three data sets. Since, small-printed and rika sets are manually annotated, quantitative results are obtained on these data sets in the form of mAP (mean Average Precision) values. In order to test the effect of mixed writing styles we also build another data set, which we refer as **combined**, by combining small-printed and rika data sets. Due to difficulty of annotating large-printed data set consisting of over 9500 words only qualitative results are given on this data set.

Figure 9 shows example query results for three words on large-printed data set and Figure 10 shows example query results for two words on rika data set. As figures show, other instances of the query words are correctly retrieved within the top matches.

Due to characteristics of Turkish language new words can be generated from a common stem using suffixes. Therefore, it is important to also match these words which are semantically similar. As can be seen from the figures, the proposed system is also able to capture the semantic similarities.

Figure 11 shows retrieval results for some selected words on small-printed and large-printed data sets. In this experiment, the words which are semantically similar are also considered as correct matches. As can be seen, the black dots, corresponding to relevant images, are mostly on the left, showing that most of the words are matched either with other instances of the same word or with semantically similar words with high accuracies.

Table 1 shows the mAP values on small-printed, rika and combined data sets. Clusters are obtained on each data set separately. Each word in a data set is used as a query and all the other words in that data set are ranked according to their similarity to the query word. The query word is always found as the first match, therefore the results when we use all words as query are higher (the number of words which appear only once is 472 for small-printed, 174 for rika, and 648 for combined). Note that, not the similar words but only the exact matches are considered for quantitative results.

There are only 10 words which are common in small-printed

Figure 9: Example query results for the first 15 matches on large-printed data sets. Exact matches are shown by green colors, and matched similar words are shown in blue colors. The first query word `olan` means `existed`, the second word `ordu` means `army` and the third word `milletin` means `nation's`. Note that, in the second query, a similar word `orduya` meaning `to army` is also retrieved.

Table 1: mAP results using proposed method. The results are reported on small-printed, and rika data sets together with the combined data set which is the combination of these two data sets. Three cases are evaluated: all words: all of the words are used as query, frequent the ones which appear only once are skipped and the rest is used for query, common: only the words which are common in small-printed and rika are used for query.

	small-printed	rika	combined
all words	0.84	0.91	0.81
frequent words	0.62	0.71	0.54
common words	0.55	0.61	0.30

and rika, most of which are stop words, such as `one, each, what, which, like, all`. Since, most of these words are short words, the performance is low and when the data sets are mixed mAP values decrease further. However, as Figure 12 shows, we are still able to capture the similarities even for different writing styles. Therefore, we believe that the proposed system can be used to retrieve words written by different people in very different writing styles.

8. COMPARISON WITH DTW METHOD

The most similar work to ours is the study of Rath and Manmatha [18], where Dynamic Time Warping (DTW) is used for matching words in single-author historical documents. This method is used for comparison to evaluate the success of our proposed method. Vertical Projection Profiles (VPP) are obtained for each word in small-printed and rika data sets, and then DTW is used for matching.

With DTW method we obtain mAP values 0.94 for all words and

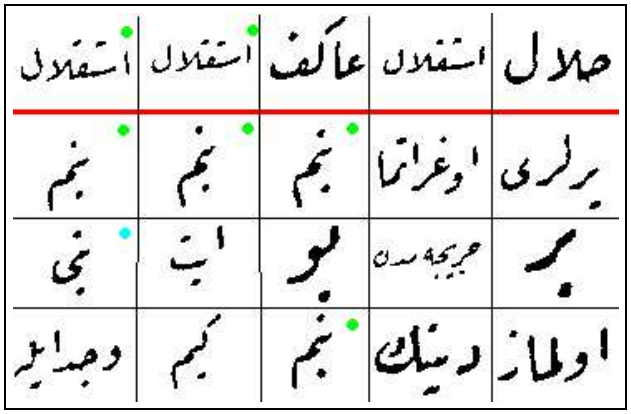

Figure 10: Example query results on rika data set. Exact matches are shown by green colors, and matched similar words are shown in blue colors. The first query word `istiklal` means `independence`, and the second word `benim` means `my`. Note that, in the second query, a similar word `beni` meaning `me` is also retrieved.

0.86 for words appearing more than once. When compared with Table 1, the results of DTW method is better when mAP values are used for comparison. It is an expected result, since the DTW method is very successful in matching the exact instances of the words.

However, as discussed above, in Turkish new words are generated from a common stem using suffixes. For example, the words meaning `libraries, to library, to libraries, from library, at library` are all derived from a single stem meaning `library`. For a better retrieval of Ottoman documents these words should also be considered but DTW method is unable to capture these similarities as shown in Figure 13.

9. CONCLUSION

In this study, we propose a method for retrieval of Ottoman documents without requiring character recognition. We show that, large number of historical documents can be indexed by using the image retrieval techniques. The proposed system matches the words with high accuracies and also captures the semantic similarity of words. Since words are treated as images, the proposed system can be extended to match words in different documents with different writing styles and authors. Although this system is evaluated on Ottoman documents, it could be adopted to other languages.

10. REFERENCES

[1] B. Al-Badr and S. A. Mahmoud. Survey and bibliography of arabic optical text recognition. *Signal Processing*, 41(1):49–77, 1995.

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

[3] E. Ataer and P. Duygulu. Retrieval of ottoman documents. In *8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[4] 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)*, 2006.

Figure 13: A word `kutuphane`, meaning `library`, is queried both using DTW approach (top) and our system (bottom). Top 12 matches are shown. Green dots indicate correct matches and blue dots indicate the semantically similar words, such as `libraries, to the library, public library` etc. Note that, all of the results are related to the query when the proposed system is used, while DTW is only able to capture 5 of the related words.

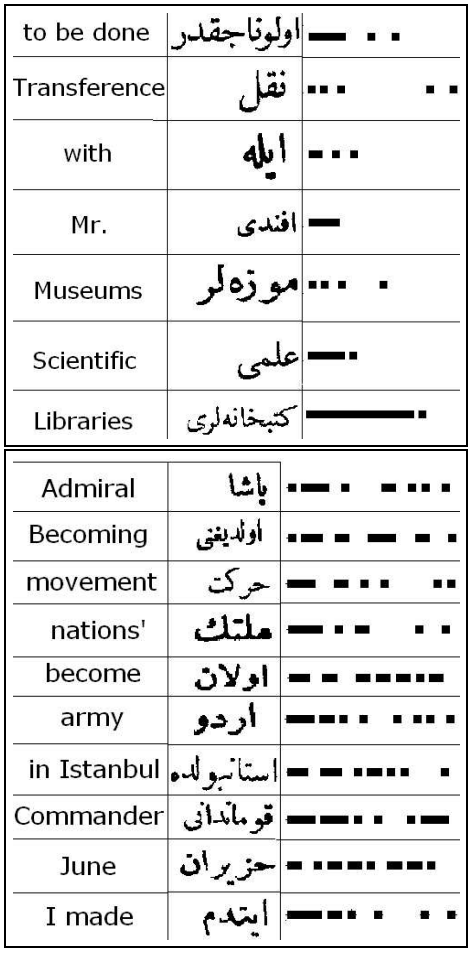

Figure 11: The ranked retrieval results for some selected words are shown on small-printed (top) and on large-printed (bottom) data sets. X axis shows the order of relevant documents retrieved in top 25 matches and Y axis represents the query words. In the ideal case, we expect all of the black dots on the left.

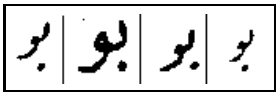

Figure 12: Example query results on combined data set, for one of the common words `bu` meaning `this`. The first one is the query word written in rika, and the others are retrieved words with rankings 12, 14 and 27. Note that second word which is in printed form is retrieved before the third and the fourth words which are in rika form, showing that the words written in different writing styles could also be matched. Note also the differences, especially in position of dots, for the words in rika, showing that slight differences are tolerated.

[5] J. Edwards, Y. W. Teh, D. A. Forsyth, R. Bock, M. Maire, and G. Vesom. Making latin manuscripts searchable using ghmms. In *Neural Information Processing Systems (NIPS)*, 2004.

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

[7] A. Gillies, E. Erlandson, J. Trenkle, and S. Schlosser. Arabic text recognition system. In *Proceedings of the Symposium on Document Image Understanding Technology*, 1999.

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

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

[10] L. M. Lorigo and V. Govindaraju. Offline arabic handwriting recognition: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(5):712–724, 2006.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal on Computer Vision*, 60(2), 2004.

[12] R. Manmatha and T. Rath. Indexing of handwritten historical documents - recent progress. In *Proc. of the Symposium on Document Image Understanding Technology (SDIUT)*, pages 77–85, 2003.

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

[14] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9, 1979.

[15] Ottoman web page, http://www.osmanli700.gen.tr/.

[16] A. Ozturk, S. Gunes, and Y. Ozbay. Multifont ottoman character recognition. In *The 7th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2000.

[17] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV)*, 2005.

[18] T. Rath and R. Manmatha. Word image matching using dynamic time warping. In *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2003.

[19] J. L. Rothfeder, R. Manmatha, and T. M. Rath. Aligning transcripts to automatically segmented handwritten manuscripts. In *Document Analysis Systems*, pages 84–95, 2006.

[20] 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, 2004.

[21] S. A. Shahab, W. G. Al-Khatib, and S. A. Mahmoud. Computer aided indexing of historical manuscripts. In *Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation*, 2006.

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

[23] J. Sivic, B. Russell, A. A. Efros, A. Zisserman, and B. Freeman. Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV)*, 2005.

[24] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2003.

[25] S. N. Srihari, H. Srinivasan, P. Babu, and C. Bhole. Spotting words in handwritten arabic documents. In *Proc. Document Recognition and Retrieval XIII (SPIE)*, 2006.

[26] 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)*, 1980.

[27] J. R. Ullmann. Advance in character recognition. In *Application of Pattern Recognition*, 1982.