Redif Extraction in Handwritten Ottoman Literary Texts

Ethem F. Can[1], Pınar Duygulu[1], Fazlı Can[1], Mehmet Kalpaklı[2]
Department of Computer Engineering[1], Department of History[2]
Bilkent University
Ankara, Turkey
Email: {efcan, duygulu, canf }@cs.bilkent.edu.tr[1], kalpakli@bilkent.edu.tr[2]

***Abstract*—Repeated patterns, rhymes and *redifs*, are among the fundamental building blocks of Ottoman *Divan* poetry. They provide integrity of a poem by connecting its parts and bring a melody to its voice. In Ottoman literature, poets wrote their works by making use of the rhymes and redifs of previous poems according to the *nazire* (creative imitation) tradition either to prove their expertise or to show respect towards old masters. Automatic recognition of *redifs* would provide important data mining opportunities in literary analyses of Ottoman poetry where the majority of it is in handwritten form. In this study, we propose a matching criterion and method, *Redif* Extraction using Contour Segments (RECS) using the proposed matching criterion, that detects *redifs* in handwritten Ottoman literary texts using only visual analysis. Our method provides a success rate of 0.682 in a test collection of 100 poems.**

***Keywords*-Ottoman Manuscripts; Word-retrieval; Word-spotting;**

I. Introduction

As an alternative to OCR based studies which are likely to fail in handwritten historical documents, the word spotting techniques [1], [2], [3], [4], [5] are proposed with the idea of considering the words as a whole rather than a sequence of characters. This area of study has gained more interest with the study of Manmatha et al. [4] for accessing historical documents. The popular approach in word-spotting based studies is to segment the documents into word images, and then use the resulting word images for word retrieval and recognition purposes. Most of the studies in historical documents focus on Latin character-based languages such as English, leaving the retrieval and recognition of documents in other languages still as an open research area.

Ottoman documents, as being produced in more than six centuries over a large area in three continents, is one of the most challenging collections. However, there have been only a few number of studies on historical Ottoman language documents [6], [7], [2], [3], [8]. Ottoman (Divan) poetry, with around one hundred thousand of available online documents [9], [10] is being studied with many scholars over the world. However, the access to these documents is currently done manually, requiring a huge effort. Most importantly, analysis of these documents and extraction of valuable information is almost impossible when the huge amount of data is considered.

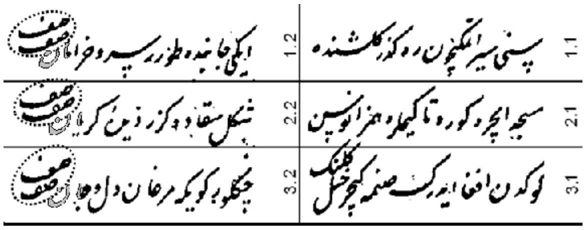
In this study we propose a method for word matching to retrieve and analyze historical documents. We represent the contour segments as sequence of code words obtained from line descriptors. The distances between the sequence of code words then determine the degree of similarity of the images. The proposed method, unlike most of the other word spotting techniques, does not require segmentation, and searches the unsegmented text as a whole. As a challenging application for word matching we focus on Ottoman literary texts and propose a method, *Redif* Extraction using Contour Segments (RECS), that extracts the *redifs*-repeated patterns-in handwritten Ottoman literary texts.

II. Motivation

In Ottoman (*Divan*) poetry, most of the poems are based on a pair of lines, i.e., distich or couple. A distich contains two hemistichs (lines). In poems, hemistichs of the same distich completes each other. The rhyme and *redif* are used to provide the integrity of the distichs of a poem and provide a melody to its voice. The *redif* can be explained as the repeated patterns following the rhyme in a poem. In Fig. 1, an original text in Ottoman script (the image on the top) and its transcription are given. The circled parts of the original text are the *redifs*, and the letters in gray are the rhymes. In the transcription the boxed words “saf saf,” are the *redifs*, and underlined letters “ân” are the rhymes. The words “saf saf” are not counted as rhyme since their meanings are the same in the hemistichs; however, the letters “ân” are the last two characters of different words.

The most common rhyme and *redif* schema in Ottoman poetry is that the rhyme followed by the *redif* in the last hemistichs of the distichs [11]. Most of the Ottoman literary texts contains *redifs* following rhymes. In Ottoman literature poets wrote poems using the same rhyme, and *redif*, and scheme of an already written poem either to challenge themselves and prove their expertise by showing that they could write as good as or better than the imitated poet or to express their respect towards old masters. This approach is called “*nazire* (creative imitation)” [12]. Among millions of un-transcribed handwritten Ottoman literary texts, *redif* extractions are almost impossible especially for non-linguists. Automatic extraction of *redifs* from handwritten documents will open Ottoman manuscripts which are among the most

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.478


1.2
1.1
2.2
2.1
3.2
3.1

1.1	Seni seyr itmek içün reh-güzer-i gülşende
1.2	îki cânibde durur serv-i hırâmân saf saf
2.1	Mescid içre göre tâ kimlere hemzânûsın
2.2	Şekl-i sakkada gezer dide-i giryân saf saf
3.1	Gökde efgân iderek sanma geçer hayl-i küleng
3.2	Çekilür kûyune mürgân-ı dil ü cân saf saf

Figure 1. Original text and its transcription.

important world literary heritages for wider and comprehensive studies.

The contributions of this study are the following. We present a pioneering image-based automatic *redif* extraction method, RECS, for handwritten Ottoman literary texts. It is based on a novel contour matching approach that can also be used in word-retrieval and -recognition. To the best of our knowledge, it is a first in the literature. We envision that Ottoman texts, which are among the world's foremost historical and cultural heritages, will become more accessible, better preserved, and open to further researches by studies like ours. Our work can be used in different ways and has several implications within the context of Ottoman literary studies, and in the implementation of various data mining methods for literary analysis of Ottoman poetry as well. For example, it can be used for detecting poems having the same *redifs* and eventually for identifying "*nazires*." Automatic identification of *nazires* can be used for detecting literary trends in Ottoman literature.

III. Word Matching

In order to match word images, we first binarize the original gray-scale images. The connected components are then found using eight-neighbors and contour segments are extracted from these connected components. Then, we approximate the points on the contour segments into lines using the Douglas and Peucker algorithm [13]. A contour segment C consists of a set of line descriptors $\{\ell_1, \ell_2, ..., \ell_n\}$ and each line descriptor ℓ contains position, length, and orientation information. We define a reference line descriptor $\ell^r = (p^r_m, \theta^r, \rho^r)$ which is the line descriptor having the closest distance to the center point (X, Y) of the contour segment. Then, we normalize each line descriptor depending upon the reference line descriptor, and describe a contour segment $C' = \{\ell'_1, \ell'_2, ..., \ell'_n\}$ using the normalized lines $\ell' = (p'_m, \theta', \rho')$. The normalization is performed in the following way.

$$\begin{gathered} X = \frac{\sum x^i_m}{n}, Y = \frac{\sum y^i_m}{n}, i = 1, 2, ..., n \\ x'_m = x_m - x^r_m,\ y'_m = y_m - y^r_m \\ \theta' = \theta - \theta^r, \rho' = \rho/\rho^r \end{gathered} \tag{1}$$

Based on the recent successes in object recognition we make use of codebook representation in our study and represent each line descriptor with its code from a codebook $B = \{b_1, b_2, ..., b_k\}$, where k is the number of labels in the codebook. In order to generate the codebook, we cluster all line descriptors from all contour segments using k-means, and represent each cluster with a single label. Having generated the codebook, each contour segment descriptor C' is described as a set of elements from codebook where each normalized line descriptor is replaced by its corresponding label from the codebook. A contour segment descriptor; thus, turns out to be $C' = \{b_{\ell'_1}, b_{\ell'_2}, ..., b_{\ell'_n}\}$, where $b_{\ell'_i} \in B$, $b_{\ell'_i}$ is the code of ℓ'_i, and $i = 1, 2, ..., n$. In order to compute the distance between two contour segment descriptors, C'_i, and C'_j, we find the amount of difference between sequences of codes of them. The difference is the sum of insertions, deletions, and substitutions of a single label in codes of the contour segment to transform codes of one descriptor to the other [14]. We use the distances between contour segment descriptors to rank the matching images for a given contour segment descriptor.

IV. *Redif* Extraction using Contour Segments (RECS)

Redif is a repeated pattern; therefore, the first to extract the *redifs* is to find the contour segments which are repeated, that is which are similar to each other. However, not all repeated sequences are *redifs*; thus, we add additional constraints, inspired by the definition of *redifs*, to differentiate the *redifs* from other repeated patterns. A *redif* must appear: (constraint 1) at the end of the second hemistich -line- of a distich -couple- and (constraint 2) in every distich.

According to constraint number 1, the x positions of the *redifs* should roughly be the same and they should be at the end of the last hemistichs. Besides, a contour segment and its matches are required to be vertically aligned to be counted as a *redif*. A contour segment is defined as being in the last part of the hemistich (last hemistichs of distichs) if its x position is less than $\alpha_1 \times w$ where w is the width of image of the poem and $\alpha_1 \in [0, 1]$. For a contour segment, we check each of its matches whether they are vertically aligned. Two contour segments are referred to as vertically aligned if the distance in x positions are less than $\alpha_2 \times w$ where w is the width of image of the poem and $\alpha_2 \in [0, 1]$. We perform experiments with different values of α_1 and α_2, and empirically determine the values of these parameters as 0.25 and 0.15 respectively.

Determination of the number of distichs in an image of a poem is a challenging task and left as a future work. Instead we use five as the minimum number of matches that should be extracted for a contour segment since five is the minimum number of distichs that a poem must have in Ottoman literature (in our collection, the poems have at least five distichs). Before applying the constraint two, we check the remaining contour segments and their matches in the case of two contour segments having the same match. In other words, we search for the contour segments that have one or more common matches and take the union of the matches of those contour segments, and we perform this operation until any pair of contour segments has a common match. We take the union of the contour segments and matches in order not to extract the same contour segment as *redif* more than once. Finally, we check the remaining contour segments whether they have a minimum of five matches or not. In the case a contour segment does not have more than four matches, we eliminate the contour segment. The remaining contour segments are extracted as *redifs*.

In order to understand the proposed method let's consider a poem with ten contour segments $\{C_1, C_2, ..., C_{10}\}$. Assume that after eliminating the contour segments not satisfying the constraint 1 only four contour segments are left and they are $\{C_1, C_6, C_7, C_9\}$ and their matches are as follows: for C_1: (C_1, C_3, C_5, C_7), for C_6: (C_6, C_4), for C_7: (C_7, C_2, C_3, C_5), and for C_9: (C_9, C_{10}). The contour segments having one or more common matches are combined by taking the union of the matches. The resulting matches turn out to be $(C_1, C_2, C_3, C_5, C_7)$, (C_6, C_4), and (C_9, C_{10}). Only C_1 has five matches, and the others has two. Depending upon the constraint 2, we eliminate the contour segment C_6, and C_9 since the number of matches are less than five. Finally, $(C_1, C_2, C_3, C_5, C_7)$ are extracted as the *redifs* in distichs. Among all the matches we select the one having minimum distance to the left border of the poem image, and return it as the *redif* of the poem.

V. Experimental Results

We constructed a collection that consists of 100 poems or part of poems from twenty different poets written in the period between the 15^{th} and 19^{th} centuries. The poems are written with different handwriting styles. The images of the poems are obtained from the "Turkey Manuscripts" web page of the Ministry of Cultures and Tourism of Turkey[10], and Ottoman Text Archive Project (OTAP)[9].

The correctness of the proposed method is computed by "extraction rate (ER)," which is the ratio of correct extractions and maximum value between number of extracted contour segments and contour segments in truth table.

In Fig. 2 we provide eight sample extraction results (Poet (century) information for the images: (a) Hamza (18-19), (b) Hayâlî (16), (c) Nihânî (16), (d) Mihrî (16), (e) Nesîmî (15), (f) Ümîdî Ahmed (16), (g) Bâkî (16), and (h) Ümidî (17). The *redifs* are circled, and extracted *redifs* are in white boxes) In Fig. 2(a), 2(b), and 2(c) the *redifs* are extracted correctly; therefore, the extraction rates are 1.0. In Fig. 2(d) three segments of the *redif* extracted correctly while four should be extracted and the ER value is 0.75. In Fig. 2(e) and 2(f) *redifs* as well as one extra contour segment are extracted in which the ER values turn out to be 0.75. Two out of three are extracted correctly in Fig. 2(g) and in Fig. 2(h); therefore, the ER value is 0.67. For the images of Fig. 2 the overall (average) ER value is 0.82 which is the average of the above ER values. Note that the number of contour segments for the *redifs* in the collection is roughly the same. For this reason, we do not prefer a weighted average computation.

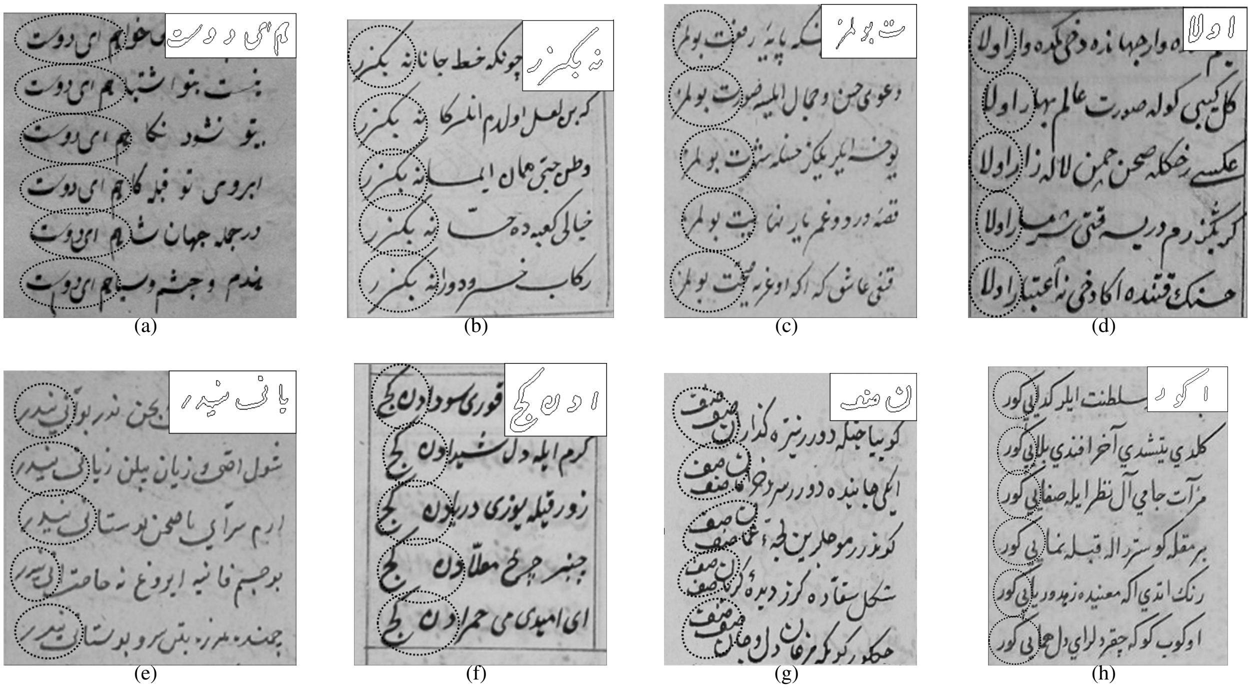
Considering the entire collection, our method extracts the *redifs* correctly with an ER score of 0.682. The score of 0.682 is obtained when k, the number of clusters in k-means, is set to 45 which is empirically found to be the best. The performance changes between 0.60 and 0.68 for different k values between 5 and 150. We observe that as the value of k increases, sensitivity of the method increases. In other words, for large values of k, our method is able to extract more complicated *redifs*; however, at the same time it misses more number of *redifs*. The experimental results imply that we have room for improvement.

As stated before, in constraint 2, the minimum number of matches of a contour segment that should be counted as *redif* is set to five. If it is decreased, the method extracts more number of correct contour segments as *redifs*; however, the number of false matches also increases. For higher values of the same parameter, the number of false matches decreases; however, in this case the number of misses increases. We also observe similar findings for α_1 and α_2.

VI. Conclusion and Future Work

We introduce an image-based automatic *redif* extraction method for handwritten Ottoman poems. Our method RECS is based on a novel contour matching approach that can also be used in word-retrieval and -recognition.

In the experiments, we obtain a success rate of 0.682. In our future research we plan to improve the performance by fusing the similarity values among the contour segments obtained by different k values (of the k-means clustering algorithm). By this way we plan to exploit the advantages provided by smaller and larger k values (remember that larger k values enable us to detect more complicated *redifs* but miss the simpler ones and we have the reversed observation with smaller k values). The results with smaller and larger k values can be fused to improve the system performance.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2. Some extraction results

Acknowledgment

This work is supported by Scientific and Technical Research Council of Turkey (TUBITAK) under grant number 109E006.

References

[1] T. Adamek, N. E. O'Connor, and A. F. Smeaton, "Word matching using single closed contours for indexing handwritten historical documents," *International Journal of Document Analysis and Recognition*, vol. 9, pp. 153–165, 2007.

[2] E. Ataer and P. Duygulu, "Retrieval of Ottoman documents," in *Proceedings of the 8th ACM SIGMM*, 2006, pp. 155–162.

[3] ——, "Matching Ottoman words: An image retrieval approach to historical document indexing," in *Proceedings of the 6th ACM CIVR*, 2007, pp. 341–347.

[4] R. Manmatha, C. Han, and E. M. Riseman, "Word spotting: A new approach to indexing handwriting," in *Proceedings of CVPR '96*, 1996, pp. 631–637.

[5] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal on Document Analysis and Recognition*, vol. 9, pp. 139–152, 2007.

[6] A. A. Atici and F. T. Yarman-Vural, "A heuristic algorithm for optical character recognition of Arabic scripts," *Journal of Signal Processing*, vol. 62, no. 1, pp. 87–99, 1997.

[7] E. Saykol, A. K. Sinop, U. Gudukbay, O. Ulusoy, and E. Cetin, "Content-based retrieval of historical Ottoman documents stored as textual images," *IEEE Trans. on Image Processing*, vol. 13, no. 3, pp. 314–325, 2004.

[8] I. Z. Yalniz, S. Altingovde, U. Gudukbay, and O. Ulusoy, "Ottoman archives explorer: A retrieval system for digital Ottoman archives," *ACM Journal on Computing and Cultural Heritage*, vol. 2, no. 3, pp. 8–20, 2009.

[9] "Ottoman text archive project (OTAP)," http://courses.washington.edu/otap/, January 2010.

[10] "T.C. Kültür ve Turizm Bakanlığı - Türkiye yazmaları," http://www.yazmalar.gov.tr, January 2010.

[11] W. G. Andrews, *An Introduction to Ottoman Poetry*. Minneapolis, USA: Bibliotheca Islamica, Inc, 1976.

[12] M. Kalpaklı, *Osmanlı Şiir Akademisi: Nazire*, ser. Türk Edebiyat Tarihi-C2. T.C. Kültür ve Turizm Bakanlığı Yayınları, 2006, pp. 133–137.

[13] D. Douglas and T. Peucker, "Algorithms for reduction of the number of points required to represent a digitized line or its caricature," *The Canadian Cartographer*, vol. 10, no. 2, pp. 112–122, 1973.

[14] I. V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.