OSMANLICA BELGELERDE KELİME ERİŞİMİ

WORD RETRIEVAL IN OTTOMAN DOCUMENTS

Damla Arifoğlu, Pınar Duygulu

Bilgisayar Mühendisliği Bölümü
Bilkent Üniversitesi
{arifoglu,duygulu}@cs.bilkent.edu.tr

ÖZETÇE

Bu çalışmada, Osmanlıca arşivlerinin analizi amacıyla, kelime erişimi problemi iki farklı resim eşleme yöntemi ile çözülmeye çalışılmaktadır. Bu amaçla (1) Dinamik Zaman Bükmesi (DZB) tabanlı kelime eşleme yöntemi [7] ve (2) Şekil İçeriği (shape context) tanımlayıcısı [10] Osmanlıca belgeler üzerinde uygulanmıştır. Öncelikle, verilen bir Osmanlıca belgedeki tüm alt-kelimeler bulunmuştur. Birinci yöntemde, her alt-kelime grubu için, üst ve alt kelime profili, siyah pikselden beyaz piksele geçiş sayısı ve dikey izdüşüm özniteliklerinden oluşturulmuş 4 parçalı öznitelik vektörü çıkartılmış, bu özniteliklerin birbirine olan uzaklığı DZB algoritmasıyla bulunmuştur. İkinci yöntemde ise, Şekil İçeriği tanımlayıcısı kullanılarak, alt-kelimelerin birbirine olan uzaklıkları hesaplanmıştır. Uygulanan yöntemler, Fuzuli'nin Leyla ve Mecnun divanının 10 sayfasından oluşan bir Osmanlıca veri kümesi üzerinde denenmiştir.

ABSTRACT

In this paper, two image matching methods are adapted to retrieve words in Ottoman documents. The first method is based on Dynamic Time Warping (DTW) method proposed in [7], while the second method is based on the Shape Context descriptor [10]. Firstly, all sub-words in a given Ottoman document are extracted. In the first method, a 4-variant feature vector (upper and lower word profiles, background to ink transition, vertical projection) is calculated for each sub-word and feature vectors' distance to each other is found by DTW algorithm. In the second method, shape context descriptor is used to calculate the distances of sub-word images. The methods are tested on an Ottoman data set, which consists of 10 pages of Leyla and Mecnun Divan of Fuzuli.

1. GİRİŞ

Osmanlı arşivleri bugün de dünyanın pek çok bölgesinden bir çok araştırmacının ilgisini çekmektedir. Bu belgelerin sayısal ortamda otomatik dizinlenmesi ve hızlı erişimi Osmanlıca belgeler üzerinde çalışan araştırmacıların işini kolaylaştırması açısından oldukça önemlidir. Bu amaçla, son zamanlarda Osmanlıca belgeler üzerinde yapılan çalışmalarda da artış gözlemlenmiştir [1-6].

Osmanlıca, Arapça'ya çok benzeyen bir dil olup, karakterleri birleştirilerek yazılmaktadır. Osmanlıca'da her boşluk bir kelime sınırını ifade etmeyebilir. Bir kelime bir karakter grubundan oluşabileceği gibi, bir çok ayrı karakter grubu da bir kelimeyi oluşturabilir (Şekil 1). Ayrıca, Osmanlıca'da karakterler, bulunduğu pozisyona (kelimenin başı, sonu, ortası) göre farklı şekiller almaktadır. Bir çok karakter birbirinden sadece nokta, zigzag gibi karakterin altına ve üstüne konan özel işaretlerle ayrılmaktadır. Tüm bu

özellikler, Osmanlıcanın tanınması ve de eşleştirilmesini zorlaştırmaktadır.

Osmanlıca belgeler sadece yazı içermekte olmayıp, bu belgelerin bir çoğu ayrıca ferman, tuğra, portre, minyatür ve figür gibi tarihsel önemi olan resimsel öğeler de içermektedir. Osmanlıca belgelerin çoğunluğu arşivlerde resim formatında saklanmaktadır ve bu nedenle Osmanlıca belgeleri birer resim gibi ele almak daha uygun olacaktır. Bu noktadan yola çıkarak, bu çalışmada Osmanlıca belgelerde kelime erişimi problemi resim eşleme problemi olarak düşünülmektedir.

Bu yaklaşım yakın zamandaki bazı çalışmalarda da öne çıkmakta, yazılı belgeler resim olarak ele alınmakta, resimler üzerinde kullanılan benzerlik bulma yöntemleri, yazılı belgeleri tanımak için de kullanılmaktadır [1-4,7]. Bu konuda Osmanlıca belgeler üzerinde yapılan çalışmalardan birinde [1], Ataer ve Duygulu Osmanlıca belgeleri resim olarak ele alarak dikey izdüşüm profili tabanlı kelime eşleştirmesinin mümkün olduğunu göstermişlerdir. Yazarlar, başka bir çalışmalarında ise [2] görsel öğeler kümesi (bag-of-visual-words) yöntemini Osmanlıca belgelerde kelime eşlemesi için kullanmışlardır.

Bu bildiride, daha önce özellikle George Washington'un el yazmaları üzerinde kelime erişimi amacıyla kullanılmış olan Dinamik Zaman Bükmesi (Dynamic Time Warping) [7] ve karakter eşlemede kullanılmış olan Şekil İçeriği (Shape Context) [10] yöntemlerinin, Osmanlıca belgeler üzerinde de kelime eşleştirilmesi ve erişimi için kullanılabileceği gösterilmektedir.

Aşağıda, ilk olarak uygulanan iki yöntemin detaylı açıklaması yer almaktadır. Daha sonra veri kümesi hakkında bilgi verilecek ve bu küme üzerindeki çeşitli deneylerden bahsedilecektir.

2. ÖNERİLEN YÖNTEMLER

2.1. Resimlerin ön-işlemesi ve alt-kelimelerin bulunması

Tüm resimler, öncelikle Otsu yöntemi kullanılarak ikilileştirilmiştir [9]. Veri kümesindeki tüm sayfalar yatay izdüşüm yöntemiyle satırlara bölütlenmiştir. Daha sonra, tüm resimlerdeki siyah piksel gruplarından oluşan bağlı bileşenler (connected components) bulunmuştur. İki nokta, üç nokta veya zig zaglardan oluşan küçük özel ayırıcılar da ait olduğu büyük bileşene eklenip oluşan bu grupların her birine *alt-kelime* adı verilmiştir. Örneğin, Şekil 1-a'daki Osmanlıca kelimede 4 bağlı bileşen varken, çizgiyle ayrılmış yalnızca 2 tane alt-kelime bulunmaktadır.

2.2. Dinamik zaman bükmesi yöntemi ile kelime erişimi

Dinamik Zaman Bükmesi (DZB) yöntemi, ilk olarak konuşma tanımada kullanılmak üzere önerilmiştir [8]. Bu yöntemde,

zaman sıralı iki sinyalden alınan örnekler arasındaki uzaklık devingen programlama ile bulunmaktadır. Bu yöntem Manmatha ve ekibi tarafından el yazısı belgelerdeki kelimelerin erişimi için uyarlanmıştır [7]. Aşağıda öncelikle [7]'de kullanılan özneliliklerin nasıl çıkartıldığından bahsedilecek, sonra da *DZB* algoritmasının yine [7]'de kelime erişimi problemi için nasıl uyarlandığı anlatılacaktır.

2.2.1. Öznelilik çıkartılması

Bulunan her alt-kelimeden 4 parçalı öznelilik vektörü oluşturmak için aşağıdaki öznelilikler çıkartılmıştır.

- *Üst kelime profili* : Kelime resminin her sütunu için, o sütundaki en üst siyah pikselin, en üst kelime sınıra uzaklığı
- *Alt kelime profili* : Kelime resminin her sütunu için, o sütundaki en alt siyah pikselin, en alt kelime sınıra uzaklığı
- *Dikey İzdüşüm* : Kelime resminin her sütunu için, o sütundaki siyah piksel sayısı
- *Siyah pikselden beyaza geçiş sayısı* : Kelime resminin her sütunu için, kaç kez siyah sütundan beyaz sütuna geçiş olduğu sayısı

Şekil 1.b-e'de verilen grafikler, Şekil 1-a'daki kelime için bu öznelilikleri göstermektedir.

2.2.2. Öznelilik Uzaklıklarının Bulunması

Öznelilik çıkarma adımı ile her alt-kelime için boyutu $C \times 4$ olan (C : kelimenin sütun piksel sayısı), her sütunu bir özneliliği temsil eden 4 parçalı bir öznelilik vektörü oluşturulmuştur. Verilen herhangi iki öznelilik vektörü arasındaki uzaklık ise, formül 1'deki gibi, devingen programlama kullanılarak hesaplanmıştır. Formüldeki $d(x_i, y_i)$, verilen iki A ve B kelime resimlerinin, i 'inci sütunlarındaki öznelilikleri arasındaki uzaklıktır ve bu uzaklık formül 2'deki gibi hesaplanır. Formül 2'de, i : resimlerin sütun numarasını, k ise öznelilik numarasını ifade etmektedir. Bu durumda örneğin; $A(1,1)$: A resminin 1. sütunun, 1 numaralı öznelilik değeridir. Elde edilen $D(i,j)$ değeri dinamik bükme yolunun uzunluğuna bölünerek normalize edilmiştir ve bu miktar iki alt-kelimenin birbirine olan uzaklığı olarak kullanılmıştır.

$$D(i, j) = \min \left\{ \begin{array}{l} D(i, j-1) \\ D(i-1, j) \\ D(i-1, j-1) \end{array} \right\} + d(x_i, y_i). \quad (1)$$

$$d(x_i, y_i) = \sum_{k=1}^4 (A(i, k) - B(i, k)) \quad (2)$$

2.3. Şekil içeriği yöntemi ile kelime erişimi

İlk olarak cisim tanımak üzere, Belongie ve Malik tarafından [10] öne sürülen bu yöntemde, şekillerin kontürleri üzerinden rastgele n nokta seçilir. Seçilen her nokta, diğer noktalar ile birleştirilerek $n-1$ tane vektör oluşturulur. Bu vektörlerin uzunluğu ve açısının sıklık grafiği çıkartılır.

Elde edilen bu grafik, o noktanın *Şekil İçeriği* tanımlayıcısı olarak kullanılır. Verilen iki resimdeki noktalar arasındaki birbirine en çok benzeyen nokta ikilileri, bu

noktaların *Şekil İçeriği* tanımlayıcıları kullanılarak eşleştirilmeye çalışılır. Noktalar arasından oluşturulan, birbirine en çok benzeyen nokta ikilileri arasındaki uzaklıklar toplanarak, bu iki şekil arasındaki uzaklık bulunmuş olunur (Daha fazla bilgi için çalışma [10]'a bakabilirsiniz). Bizim çalışmamızda, alt-kelime şekilleri arasındaki uzaklık bu şekilde bulunduktan sonra elde edilen uzaklıklar kelimelerin eşleştirilmesi sırasında kullanılmıştır.

Bu yöntem, çalışma [11]'in yazarları tarafından George Washington'ın eserlerinin dizinlenmesi için de denenmiştir, ancak yazarlar bu yöntemin kendi çalışmalarında başarılı sonuçlar vermediğini belirtmişlerdir.

Şekil 1: Bir Osmanlıca kelime “beyle” (böyle) ve ona ait Bölüm 2.2.1.’deki sıra ile 4 öznelilik değerlerinin grafiği

3. DENEYLER

3.1. Veri Kümesi

Uygulanan iki yöntemi test etmek üzere bir Osmanlıca veri kümesi oluşturmak için Osmanlıca bir kitaptan [12], 10 sayfa taranmıştır. Fuzuli'nin, *Leyla ile Mecnun* divanından seçilen sayfalardan oluşturulan bu veri kümesinde toplamda 1122 tane kelime vardır. Bunların 693 tanesi veri kümesi içinde tekrar eden kelimelerdir. Bu veri kümesi Osmanlıca bilen bir kişi tarafından kelimelerine ayrılmış ve her kelimenin Türkçe karşılığı etiketlenmiştir. Oluşturulan bu veri kümesi üzerinde aşağıdaki deneyler uygulanmıştır.

3.2. Sorgu kelimesinin kelimelerine ayrılmış veri kümesinde aranması

Bu deneyde, veri kümesindeki her kelime sorgu kelimesi olarak kullanılıp sorgu kelimelerine benzeyen kelimeler, kelimelere ayrılmış veri kümesinde, yani doğru bölütlenmiş kelimeler arasında aranmıştır.

Diyelim A ve B kelimeleri, $A = \{a_n, \dots, a_2, a_1\}$ ve $B = \{b_n, \dots, b_2, b_1\}$ şeklinde sıralı a ve b alt-kelime gruplarından oluşsun. A ve B kelimelerinin birbirine olan uzaklığı formül 3'teki gibi bulunur. $d(a_k, b_k)$ alt-kelimelerin birbirine olan uzaklığıdır. Alt-kelime sayısı birbirine eşit olmayan kelime ikilileri eşleştirilmiştir. Eğer $d(A, B)$ miktarı, deneysel olarak belirlenen bir eşik değerinden az ise verilen bu iki kelime eşleşmiş olarak kabul edilmiştir.

$$d(A, B) = \sum_{k=1}^n d(a_k, b_k) \quad (3)$$

Veri kümesindeki tüm kelimelerin sorgu kelimesi olarak kullanıldığı bu deneyde, farklı eşik değerleriyle yapılan deneyler ve çıkan geri getirme (recall) ve kesinlik (precision) oranları, formül 4 ve 5'teki gibi hesaplanmış olup, *DZB* ve *Şekil İçeriği* yöntemi için Tablo 1'de bu sonuçlar gösterilmiştir. Görüldüğü gibi, eşik değeri arttıkça, geri getirme oranı artıp kesinlik oranı azalmıştır. Çünkü, eşik değerinin artması doğru kelimelerle beraber, yanlış eşleşen kelimelerin sayısını da arttırmıştır. Kullanılan öznitelikler karakterlerin şekillerini iyi temsil ettiğinden, *DZB* yönteminde geri getirme, yani doğru kelimeleri yakalama oranı *Şekil İçeriği* yöntemine göre daha yüksektir.

$$\text{Geri getirme} = \frac{\text{Elde edilen doğru kelime sayısı}}{\text{Veri kümesindeki kelime sayısı}} \quad (4)$$

$$\text{Kesinlik} = \frac{\text{Elde edilen doğru kelime sayısı}}{\text{Erişilen kelime sayısı}} \quad (5)$$

Şekil 2'de *DZB* yöntemiyle, "*Felek*" sorgu kelimesi için 0.4 eşik değeri ile erişilen kelimeler gösterilmektedir. Birinci kelime sorgu kelimesi olup diğerleri eşleştirilen doğru kelimelerdir. Şekil 3'te, eşik değeri 0.4 iken *DZB* yöntemiyle yanlış eşleşen kelimelere örnekler verilmiştir. İlk sıradaki kelimeler sorgu kelimeleri olup her kelime alttakiyle eşleşmiştir. Görüldüğü gibi yanlış eşleşen kelimelerin bile şekilleri sorgu kelimelerine çok benzemektedir.

Şekil 4'te *Şekil İçeriği* yöntemiyle, 0.4 eşik değeri ile erişilen kelimelere örnekler verilmektedir. İlk kelime sorgu kelimesi olup "+" ile gösterilen kelimeler erişilen doğru kelimeler, "-" ile gösterilenler erişilen yanlış kelimelerdir. Yanlış eşlenen kelimelerin şekilleri yine sorgu kelimesine çok benzemektedir.

Tablo 1: *DZB* ve *Şekil İçeriği* yöntemleriyle elde edilen sonuçlar

Yöntem	Eşik Değeri	Geri Getirme	Kesinlik
Dinamik Zaman Bükmesi	0.2	0.80	0.82
	0.4	0.90	0.76
	0.6	0.93	0.73
	0.8	0.94	0.72
Şekil İçeriği	0.2	0.59	0.92
	0.4	0.73	0.83
	0.6	0.79	0.74
	0.8	0.80	0.67

فلک	ملك	فلک
فلک	فلک	فلک

Şekil 2: *DZB* yöntemiyle erişilen doğru kelimelere örnekler

حال	فلک	معنى
حمال	لعلک	منى

Şekil 3: *DZB* yöntemiyle yanlış eşleşen kelimelere örnekler

حال ⁺	حال ⁺	حال ⁺
غزل ⁻	خیال ⁻	بنای ⁻

Şekil 4: *Şekil İçeriği* yöntemiyle erişilen kelimelere örnekler

3.3. Verilen kelimenin, kelimelerine ayrılmamış veri kümesinde aranması

Bu deneyde diğerinden farklı olarak, veri kümesi kelimelerine ayrılmadan verildiğinden, aday kelimeler otomatik olarak bilinmemektedir. Bu yüzden verilen sorgu kelimesi veri kümesinde aranırken sayfaların satırlarındaki alt-kelimeler kayan pencere yöntemi ile taranmış, doğru aday kelimeler bilinmediğinden her olası kelimeye bakılmıştır. Bu pencere, piksel yerine, alt-kelimeler üzerinden kaydırılmıştır. Bu pencerenin genişliği, sorgu kelimesindeki alt-kelime sayısı kadar alınmakta olup taranan her pencerede alt-kelimelerden oluşan sahte veya doğru kelime, aday kelime olarak kabul edilmiştir. Sorgu kelimesi ile aday kelimeler arasındaki uzaklık ilk deneydeki gibi bulunmuştur. Yine elde edilen uzaklık belirlenen eşik değerinin altındaysa, aday kelime ve sorgu kelimesi eşleşmiş olarak kabul edilmiştir.

Tablo 2'de, her iki yöntem için de, farklı eşik değerleriyle yapılan deney sonuçları gösterilmiştir. Bu deneyde görüldüğü gibi kesinlik oranları daha düşüktür, çünkü sorgu kelimesinin arandığı olası kelime sayısı ve dolayısıyla

yanlış kelime sayısı artmıştır. Tablo 2’de, *Şekil İçeriği* yönteminde 0.2 eşik değeri düşük bir değer olup, bu değerle daha az kelime yakalandığından ve de erişilen kelimeler doğru olduğundan kesinlik değerleri yüksektir. Ancak 0.8 eşik değeri yüksek bir değer olduğundan, bu değerle, bir çok olası yanlış kelime grubu da doğru gibi alınmıştır ve kesinlik değeri çok düşmüştür.

Tablo 2: DZB ve *Şekil İçeriği* yöntemleriyle elde edilen sonuçlar

Yöntem	Eşik Değeri	Geri Getirme	Kesinlik
Dinamik Zaman Bükmesi	0.2	0.80	0.39
	0.4	0.90	0.31
	0.6	0.93	0.26
	0.8	0.94	0.25
Şekil İçeriği	0.2	0.59	0.72
	0.4	0.73	0.45
	0.6	0.79	0.25
	0.8	0.80	0.20

Şekil 5’te her iki yöntemle erişilen kelimelere örnekler verilmiştir. İlk kelimeler sorgu kelimesi olup “+” ile gösterilen erişilen doğru kelimeleri, “-” ile gösterilenler erişilen yanlış kelimeleri göstermektedir.

+ ايله	+ ايله	+ ايله
- افغا	- اعتبا	- ايستر
+ غزل	+ غزل	+ غزل
- كونلر	- جور	- خوش

Şekil 5: Birinci grup DZB yöntemi ile, ikinci grup Şekil İçeriği yöntemiyle erişilen kelimelerdir.

4. ÖZET VE TARTIŞMA

Bu çalışmada, *Dinamik Zaman Bükmesi* ve *Şekil İçeriği* yöntemleri kullanılarak Osmanlıca belgelerde, kelime erişiminin mümkün olduğu gösterilmiştir. Sonuçlar hem kelime tabanlı bir arama üzerinden, hem de verilen sorgu kelimesinin tüm olası kelime adayları üzerinden aranması deneyiyle gösterilmiştir. DZB yöntemi, bir kelimenin doğru eşlemelerini yakalamakta başarılıdır, ancak eşik değeri arttıkça, yanlış eşlemeleri eleyememektedir. Uyarlanan DZB yöntemi, *Şekil İçeriği* yönteminin aksine ölçeklendirme, kayma ve ilgin dönüşümü değişmezlik özelliklerine sahip değildir. Ancak, bu çalışmada kullanılan veri kümesi sadece matbu sayfalardan oluştuğu için kelimeler arasında aç ve büyüklükte farklılıklar yoktur, o yüzden bu özelliklerin olmaması çok da önemli değildir. Fakat farklı yazarlar

tarafından yazılmış belgelerden oluşan bir veri kümesinde, DZB yöntemi ve kullanılan bu öznelilikler *Şekil İçeriği* yöntemine göre daha kötü sonuçlar verecektir.

İleriki çalışmalarımızda, farklı yazarlardan tarafından oluşturulan bir veri kümesinde, Osmanlıcanın karakteristik özelliklerini daha iyi temsil edebilecek başka öznelilikleri denemeyi hedefliyoruz.

5. TEŞEKKÜR

Bu çalışma kısmen TÜBİTAK 109E006 nolu proje tarafından desteklenmiştir. Ayrıca, kullanılan Osmanlıca veri kümesi için doğrulamayı sağlayan Bilkent Üniversitesi Türk Edebiyatı Bölümü’nden Meriç Kurtuluş’a ve *Şekil İçeriği* yönteminin kodunu paylaştıkları için Belongie ve ekibine teşekkür ederiz.

6. KAYNAKÇA

- [1] Ataer, E. and Duygulu, P., “Retrieval of Ottoman documents”, *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, p 155-162, 2006.
- [2] Ataer, E. ve Duygulu, P., “Matching ottoman words: An image retrieval approach to historical document indexing”, *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, p 341-347, 2007.
- [3] Yalnız, I. Z. and Altıngöve, I. S. and Gudukbay, U. and Ulusoy, O., “Ottoman archives explorer: A retrieval system for digital Ottoman archives”, *J. Comput. Cult. Herit.*, 2 (3) : p 1-10, 2009.
- [4] Yalnız, I. Z. and Altıngöve, I. S. and Gudukbay, U. and Ulusoy, O., “Integrated segmentation and recognition of connected Ottoman script”, *Optical Engin.*, 48 (11), 2008.
- [5] Kılıç, N. and Gorgel P. and Ucan O. N. and Kala A., Multifont Ottoman character recognition using support vector machine, *EEE International Symposium on Control Communication and Signal Processing (ISCCSP'08)*, p 328-333, 2008.
- [6] Ozturk, A. and Gunes S. and Ozbay Y., Multifont Ottoman character recognition, *7th IEEE International Conference on Electronics Circuits and Systems (ICECS)*, p 945-949, 2000.
- [7] Toni M. Rath and R. Manmatha, “Word Image Matching Using Dynamic Time Warping”, *Proceedings of CVPR-03 conference 2002*, 2: p 521-527, 2002.
- [8] Kruskal, J.B., Liberman M., “The symmetric time-warping problem: from continuous to discrete”, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, p 125-161, 1983.
- [9] Otsu, N., “A Threshold Grey Scale Histogram”, *IEEE Trans. on Syst. Man and Cyber.*, p 62-66, 1979.
- [10] Belongie S. and Malik, J. and Puzicha, J., “Shape Matching and Object Recognition Using Shape Contexts”, *IEEE Trans. Pattern Anal. Mach. Intell.*, 24 (4): p 509-522, 2002.
- [11] Rath, T., Kane, S., Lehman, A., Partidge, E., Manmatha R., “Indexing for a Digital Library of George Washington’s Manuscripts: A Study of Word Matching Techniques”, *CIIR Technical Report*, 2002.
- [12] Doğan, M. N., “Mecnun ve Leyla Dilinden Şiirler”, *Enderun Kitabevi*, 1997.