Osmanlıca Kelimeleri Eşleme
Matching Ottoman Words

Esra Ataer, Pınar Duygulu

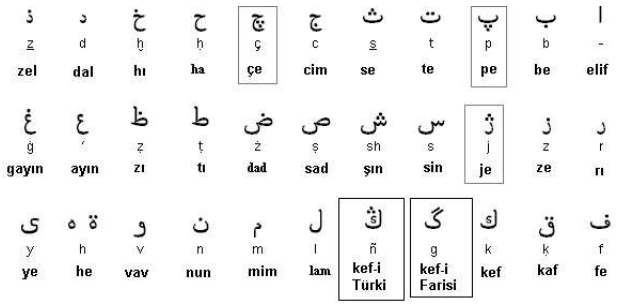
Bilgisayar Mühendisliği Bölümü
Bilkent Üniversitesi, Bilkent, Ankara
{ataer, duygulu}@cs.bilkent.edu.tr

Özetçe

Osmanlı arşivleri dünyanın pek çok yerinden araştırmacının ilgi alanına girmektedir. Fakat bu belgelerin elle çevirisi zor bir iş olduğu için, bu arşivler kullanılamaz durumdadır. Otomatik çeviri gerekmektedir, fakat Osmanlıca'nın yazma özelliklerinden dolayı karakter tabanlı tanıma sistemleri istenen başarıyı gösterememektedir. Ayrıca, belgeler minyatür ve tuğra gibi önemli kısımlar içerdiği için, imge formatında saklanmaları gerekmektedir. Bu nedenle, bu çalışmada Osmanlıca kelimeleri imge olarak görerek probleme imge erişim problemi olarak yaklaşıldı ve kelime eşleme tekniği üzerine bir çözüm önerisinde bulunuldu. Nesne tanımada başarılı olan görsel öğeler kümesi (bag-of-visterms) tekniği kelime eşleme işlemine uyarlandı ve böylece her kelime imgesi taç noktalarından çıkarılan SIFT özelliklerinin vektör nicemlemesiyle sembolize edildi. Benzer kelimeler görsel öğelerin dağılımına göre eşlendi. Deneyler 10,000 kelimenin üzerindeki matbu ve elyazması belge üzerinde yapıldı. Sonuçlar sistemin benzer kelimeleri yüksek doğrulukla eşlediğini ve anlamsal benzerlikleri bulduğunu gösteriyor.

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.



Şekil 1: Osmanlı alfabesindeki harfler. Osmanlıca Arapçadaki 28 harften farklı olarak 5 harf daha içermektedir, bunlar şekilde çerçeve içine alınmıştır.

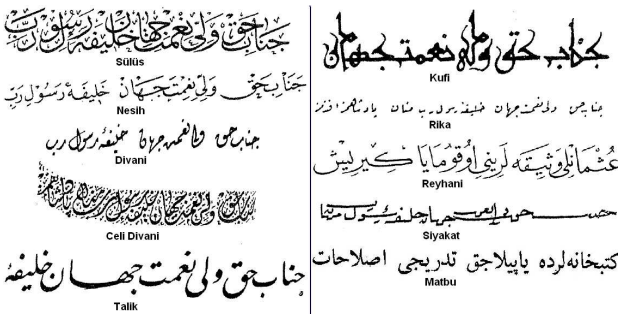
1. Giriş

Dünyanın pek çok yerinden araştırmacının ilgi alanına giren Osmanlı arşivleri Osmanlı dönemine ait askeri, politik ve ekonomik belgeler içeren 150 milyondan fazla belge içermektedir. Elle etiketlemenin ve çevirinin ve bunları otomatik yapacak bir sistem oluşturmanın zor olmasından dolayı bu arşivler rahatlıkla kullanılamamaktadır. Ayrıca belgeler minyatür ve tuğra gibi çizimler içerdiği için bu belgelerin imge formatında saklanmaları gerekmektedir ve eskiyen bu belgelerin okunması giderek zorlaşmaktadır.

Osmanlıca, Arap alfabesinin harflerine Farsça ve Türkçeden bazı seslilerin de ilavesiyle oluşturulmuş birleşik bir yazı stilidir (Şekil 1) ve Arapça belge analizindeki zorluklar Osmanlıca için de geçerlidir. Osmanlı alfabesindeki harfler Arapça harfler gibi kelime içindeki yerine göre dört farklı formda bulunabilir (Başta, ortada, sonda ve ayrı). Osmanlıca ve Arapçanın ortak bir diğer özelliği az sesli harfe sahip olmalarıdır. Bu nedenle bir metnin çevirisi okuyucunun kelime dağarcığına ve metnin içeriğine göre değişebilmektedir.

Osmanlı yazı sanatı olan hat sanatı imparatorluk tarafından teşvik edilmekle beraber Türkler tarafından bir çok yazı çeşidi kullanılmıştır (Şekil 2). Yazı çeşitlerinde kullanılan uzatmalar ve çizimler probleme karakter tanıma olarak yaklaşılmasını zor kılmaktadır. Şekil 3'de görüldüğü gibi Osmanlıca belgeler yazıdan ziyade resme benzemektedir.

Arapça karakter tanıma üzerine çok çalışma olmasına rağmen [7, 8] Osmanlıca belgelerin erişim ve tanıması prob-

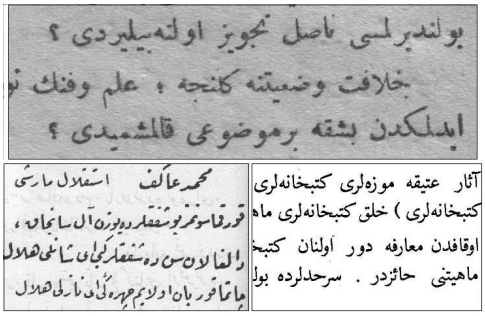


Şekil 2: Osmanlı hat sanatındaki bazı yazı çeşitleri.

Şekil 3: Örnek fermanlar.

lemi bir kaç bildiri dışında [5, 6, 9] fazla çalışılmamıştır. Osmanlıca'nın yukarıda sayılan özelliklerinden dolayı bu çalışmada probleme belge çözümleme problemi olarak değil, imge eşleme problemi olarak yaklaşıldı. Bu çalışmada Osmanlıca belgeler öncelikle kelimelere ayrıldı ve kelimelere harfler bütünü olmaktan çok imge olarak bakıldı. Her kelime imgesi görsel tanımlayıcılarla sembolize edilip imge eşleme tekniği kullanıldı. Kelime imgeleri üzerindeki kıvrım ve bağlantı noktalarının karakteristik noktalar olduğunu varsayarak, bu ayırt edici bölgeleri sembolize etmede başarılı olan taç bölgeleri kullanıldı. Bu bölgeleri bulmak ve tanımlamak için SIFT tanımlayıcılarından [1] yararlanıldı. Sonra bu bölgeler üzerinde, nesne tanımada kullanılan görsel öğeler kümesi yöntemi uygulandı. Taç bölgelerden çıkarılan SIFT tanımlayıcıları görsel birimler oluşturmak için vektör nicemlendi ve her kelime imgesi bu görsel birimlerle sembolize edildi. Görsel birimlerin o kelime imgesindeki dağılımına göre benzer kelime imgeleri eşlendi.

Bildirinin kalan kısmı şöyle tasnif edilmiştir: Bölüm 2 de çalışmamızda kullandığımız veri kümelerinden bahsedilecek. Bölüm 3 de deneyler ve sonuçları üzerinde durulup, görsel öğe oluşumu bölüm 4 de anlatılacak. Deneyler ve sonuçlar bölüm 5 de verilerek ve bölüm 6 de sonuçlar bu alandaki benzer bir yöntem olan Dinamik Zaman Bükmesi (DZB) (Dynamic Time Warping) [3] yöntemiyle kıyaslanacaktır. Sonuçlar üzerindeki özet ve sonuç bölüm 7 de verilecektir.



Şekil 4: Kullanılan örnek belgeler: **üst:** büyük-matbu, **alt sol:** rika, **alt sağ:** küçük-matbu.

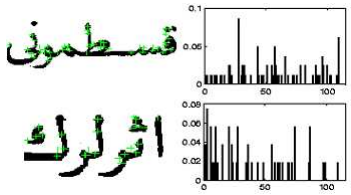
2. Veri Kümeleri

Bu çalışmada üç farklı veri seti kullanılmıştır (Şekil 4). Bunlardan biri küçük-matbu adını verdiğimiz, Türkiye Cumhuriyeti'nin ilk dönemlerinde kütüphaneler hakkındaki düzenlemelerden bahseden 6 matbu belgeden oluşmaktadır. Otomatik değerlendirme için bu belgelerden çıkarılan 823 kelime imgesi elle etiketlendi. Diğer veri seti, büyük-matbu, Mustafa Kemal Atatürk'ün Nutuk adlı eserinin ilk 25 sayfasından çıkarılan 9524 kelime imgesinden oluşuyor. Üçüncü veri seti, rika, İstiklal Marşı'nın rika yazı tipi ile yazılmış halinden çıkarılan 257 kelimeden oluşuyor. Rika, Osmanlı Devleti'nde özellikle resmi yazışmalarda kullanılan bir yazı çeşididir ve rika kümesindeki kelimeler de elle etiketlenmiş durumdadır. Küçük ve büyük matbu veri kümesindeki belgeler aynı yazı tipinde olsa da karakter büyüklükleri farklılıklar arzetmektedir.

3. Kelime Eşleme

Bu çalışmada kelime eşleme üzerinde yoğunlaşıldığı için bölütleme işlemi basit ve hızlı tekniklerle yapıldı. Otsu yöntemiyle [2] temizlenen belgelerden yatay ve dikey izdüşüm profilleriyle önce satırlar sonra kelimeler bölütlendi. Matbu metinlerde yazı tipinin düzgünlüğünden dolayı otomatik bölütleme yapılırken, rika için yazının düzensizliğinden dolayı satırlarda otomatik, kelimelerde elle bölütleme yapılabilmiştir.

Bu çalışmada kelime imgeleri harfler bütünü olmaktan çok imge olarak görülmüştür ve bu imgelerin tanımlanmasında anahtar noktalardan çıkarılan SIFT özellikleri kullanılmıştır. Şekil 5'de görüldüğü gibi bu anahtar noktalar genellikle kıvrım ve bağlantı bölgeleri ya da noktalar gibi ayırt edici bölgelerde bulunmaktadır. Kelime imgelerini sadece bu anahtar bölgelerindeki özellikleri kullanarak eşlemek yerine, görsel öğeler kümesi yaklaşımı kullanıldı [4]. Bu yaklaşımla her kelime imgesi görsel öğeler içeren bir doküman olarak görüldü. Görsel öğeler öznitelik vektörlerinin k-means yöntemiyle nicemlenmesiyle oluşturuldu ve her kelime imgesi bu görsel öğelerin düzgelenmiş dağılımıyla sembolize edildi. Kelimeler arasındaki eşlemeler bu dağılımlar arasındaki kl-ıraksay uzaklığına göre yapıldı.



Şekil 5: İki örnek kelime imgesinin anahtar noktaları ve görsel öğe dağılımı.

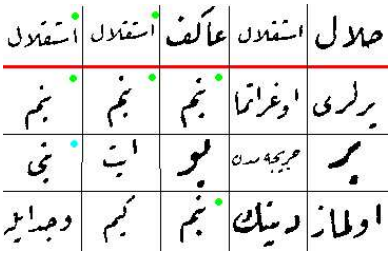
Şekil 6: Büyük-matbu veri kümesindeki bazı sorgularda erişilen ilk 15 kelime. Tam uyumlu kelimeler yeşil, benzer kelimeler mavi noktalarla göserilmiştir. Birinci sorgu kelimesi `olan`, ikinci sorgu kelimesi `ordu` ve üçüncü sorgu kelimesi `milletin`'dir. İkinci sorguda `orduya` kelimesine de erişildiğine dikkat ediniz.

4. Görsel Öğelerin Oluşturulması

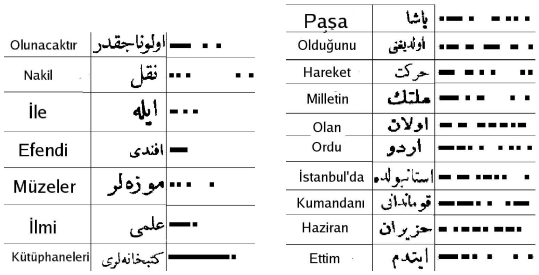
Osmanlı alfabesindeki harfler başta, ortada, sonda ve ayrı olmak üzere dört farklı formda olabilir. Her harfin farklı formları birbirine benzerdir ve bazı farklı harfler ortak kısımlar içerse de, bir harf en çok kendisinin farklı bir formuna benzerdir. Bu gözlemlerden yararlanarak k-means gruplandırmasında aynı gruptaki anahtar noktalarının mümkün olduğunca az harfe ait olması ve bir harfin farklı formlarında bulunan toplam grup sayısının mümkün olduğunca az olması gerektiğini düşünerek k-means gruplandırmasında farklı k sayıları için aşağıdaki hata oranını hesapladık:

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

Burada C alfabedeki harf sayısını, c_i c harfinin içerdiği görsel öğe gruplarının sayısı, M grup sayısı ve m_j j'inci grubun içerdiği toplam harf sayısını göstermektedir. Her harfin farklı formlarının bulunduğu 117 elemanlık bir kod tablosunda bu hata oranı 10-200 arasındaki 10'un katlarındaki sayılar için ölçüldü. Bu hata ölçüsünü optimize eden k değeri 110 olarak hesaplandı. Böylece görsel öğeleri oluşturmak için en uygun k sayısı seçilmiş oldu. Bu grup sayısını k-means algoritmasını ilklendirme için de kullanabilecekken, rast gele ilklendirmenin verdiği sonuç da bundan çok farklı olmadığı için kod tablosundan gelen gruplama merkezleri asıl veri setlerindeki deneylerde kullanılmadı.



Şekil 7: Rika veri kümesindeki bazı örnek sorgu sonuçları. Tam uyumlu kelimeler yeşil, benzer kelimeler mavi noktalarla gösterilmiştir. Birinci sorgu kelimesi `istiklal`, ikinci sorgu kelimesi `benim`dir. İkinci sorguda benzer bir kelime olan `beni` kelimesi de yakın sonuçlar arasındadır.



Şekil 8: Sıralanmış sonuçların küçük-matbu (solda) ve büyük-matbu (sağda) veri kümelerindeki gösterimi. X ekseni doğru sonuçların sırasını, Y ekseni kelimeleri göstermektedir. En ideal durumda bütün noktaların sol tarafta olmasını bekleriz.

5. Deneyler ve Sonuçlar

Deneyler bütün veri setlerinde denendi. Ancak etiket bilgisinin olduğu küçük-matbu ve rika setlerinde mAP formatında sayısal sonuçlar elde edilmiştir. Etiketlemenin zor olduğu büyük-matbu veri kümesinde ise bazı nitel sonuçlar verilmiştir. Farklı yazı tiplerindeki başarıyı görmek amacıyla rika ve küçük-matbu kümelerini birleştirerek oluşturulan bileşik veri kümesinde sayısal sonuçlar elde edilmiştir.

Şekil 6'de büyük-matbu veri kümesindeki ve Şekil 7'de rika kümesindeki 2 sorgu sonucu görülüyor. Şekillerde görüldüğü gibi ilgili imgeler erişilen sonuçların ilk sıralarında görülmektedir.

Türkçenin sondan eklemeli ve türemiş kelime içeren bir dil olmasından dolayı bazı kelimeler diğer kelimelerin içinde bulunabilmektedir. Sistem bu kelimeleri de ilk sıralarda bularak, anlamsal yakınlığı olan kelimeleri de bulmuş olmaktadır.

Şekil 8 küçük ve büyük matbu veri kümelerinde bazı örnek sorgu sonuçları görülmektedir. Bu sonuçlarda anlamsal benzerliği olan kelimeler de doğru sonuç olarak kabul edildi. Şekilde görüldüğü üzere siyah noktalar sola doğru daha yoğundur, bu da tam doğru veya anlamsal doğruluğu olan kelimelere ilk sıralarda erişildiğini göstermektedir.

Tablo 1 küçük-matbu, rika ve bileşik veri setlerindeki sayısal sonuçları göstermektedir. Gruplama her kümede ayrı olarak yapıldı ve kümedeki her kelime sorgu kelimesi olarak kullanılarak doğruluk değerlerinin ortalaması alındı. Sorgu ke-

Tablo 1: mAP sonuçları. Bütün kelimeler sorgu olarak kullanıldığı gibi, sadece birden fazla geçen kelimelerin sorgu olarak kullanıldığı sonuçlar da verilmiştir. Ortak kelimeler rika ve küçük-matbu veri kümelerinin ikisinde de olan kelimelerdir.

	küçük-matbu	rika	bileşik
bütün kelimeler	0.84	0.91	0.81
sık geçen kelimeler	0.62	0.71	0.54
ortak kelimeler	0.55	0.61	0.30

limesine hep ilk sırada erişildi, bu nedenle bütün kelimeleri sorguladığımızda elde ettiğimiz doğruluk değeri, sık geçen kelimelerin sorgularının doğruluk değerinden daha yüksek olmaktadır. Bu nedenle sık geçen kelimelerin sorgulanmasındaki doğruluk değerleri de verilmiştir. Bu sonuçlarda benzer kelimeler doğru erişim olarak alınmamıştır.

Rika ve küçük-matbu kümelerinde ortak olan 10 kelime var ve bunlar `bir, her, ne, hepsi` gibi kısa kelimeler olduğu için başarı bileşik sette diğerlerinden daha düşüktür. Fakat Şekil 9 de görüldüğü gibi bileşik set için de doğru sonuçlara ilk sıralarda erişilebilmektedir. Bu nedenle kullanılan yöntemin farklı yazı tipi içeren veri kümelerinde erişim yapmak için uygun olabileceği düşünülmektedir.

6. DZB Yöntemiyle Kıyaslama

Bu çalışmaya en çok benzeyen çalışmalardan biri Rath ve Manmatha'nın dinamik zaman bükmesi tekniğidir. Bu teknik kelime imgeleri arasındaki benzerlikleri, uzunluklar arasında fark olsa bile birbirine uydurmaya çalışarak yakalamayı hedeflemektedir. Dinamik zaman bükmesi tekniği küçük-matbu veri kümesinde bulunan kelime imgelerinin dikey izdüşüm profilleri çıkarılarak, bunlar üzerinde denendi. DZB tekniğinin doğruluk oranı bütün kelimelerde 0.94, sık geçen kelimelerde 0.86 olarak ölçülmektedir. DZB tekniği tam uyumları bulmakta başarılı olduğu için bu sonuç beklenen bir sonuçtur ancak bu çalışmada kullanılan yöntem anlamsal benzerlikleri de ortaya çıkardığı için Şekil 10'de görüldüğü gibi DZB yönteminden daha başarılı sorgu sonuçları gözlenebilmektedir.

7. Sonuç

Bu çalışmada karakter tanıma gerektirmeden Osmanlıca belgelerin erişimini sağlayan bir sistem sunuldu. Önerilen sistem benzer kelimeleri yüksek doğrulukla eşlemekte ve ayrıca diğer çalışmalardan farklı olarak anlamsal benzerlikleri de bulabilmektedir. Bu teknik sayesinde farklı yazarlardan ve farklı yazı tiplerinden oluşturulan veri kümelerinde dizgeleme yapılabilir. Önerilen yöntem sadece Osmanlıca belgelerde denenmiş olsa da, diğer dillere de uyarlanabileceği düşünülmektedir.

8. Kaynakça

[1] David G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal on Computer Vision, Vol. 60, p 91-110, 2004.

[2] N. Otsu, "A Threshold Selection Method from Gray Level Histograms", IEEE Trans. Systems, Man and Cybernetics, Vol. 9, p 62-66, 1979.

[3] T. Rath and R. Manmatha, "Word Image Matching Using Dynamic Time Warping", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, p 521-527, 2003.

[4] P. Quelhas and F. Monay and J.-M. Odobez and D. Gatica-Perez and T. Tuytelaars and L. Van Gool, "Modeling Scenes with Local Descriptors and Latent Aspects", IEEE International Conference on Computer Vision (ICCV), Vol. 1, p 883-890, 2005.

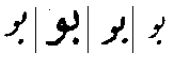
[5] A. Ozturk and S. Gunes and Y. Ozbay, "Multifont Ottoman Character Recognition", IEEE International Conference on Electronics, Circuits and Systems (ICECS), Vol. 2, p 945-949, 2000.

[6] E. Saykol and A. K. Sinop and U. Güdükbay and Ö. Ulusoy and A. E. Çetin, "Content-based retrieval of historical Ottoman documents stored as textual images", IEEE Transactions on Image Processing, Vol. 13, p 314-325, 2004.

[7] L.M. Lorigo and V. Govindaraju, "Offline Arabic Handwriting Recognition: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, p 712-724, 2006.

[8] J. Chan and C. Ziftci and D. Forsyth, "Searching Off-line Arabic Documents", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, p 1455-1462, 2006.

[9] E. Ataer and P. Duygulu, "Retrieval of Ottoman Documents", ACM SIGMM International Workshop on Multimedia Information Retrieval, p 155-162, 2006.



Şekil 9: Bileşik veri kümesinde `bu` kelimesi için sorgu sonuçları. Birinci kelime rika formatındaki sorgu kelimesi, diğerleri 12, 14 ve 27. sırada erişilen doğru sonuçlar. İkinci doğru kelime matbu formunda olup, diğer rika formlarından daha önce erişilmiştir.

Şekil 10: `Kütüphane` kelimesinin DZB yöntemi (üst) ve bizim tekniğimizdeki (alt) sorgu sonuçları. İlk 12 sonuç gösterilmiştir ve yeşil noktalar tam doğru, mavi noktalar benzer kelimeleri göstermektedir. Bizim sistemimizdeki sorgu sonuçlarında hep ilgili kelimelere erişildiğine dikkat ediniz.