OSMANLICA BELGELERİN SATIRLARA BÖLÜTLENMESİ
LINE SEGMENTATION OF OTTOMAN DOCUMENTS

Hande Adıgüze1, Pınar Duygulu Şahin

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

Mehmet Kalpaklı

Tarih Bölümü
Bilkent Üniversitesi
kalpaklı@bilkent.edu.tr

ÖZETÇE

Osmanlıca arşivler dünyanın pek çok yerinden tarihçilerin ve konuyla ilgilenen araştırmacıların ilgisini çekmektedir. Bu arşivlerin elle çevirisi için uzman kişilerin yardımı gerekmektedir ve bu işlemin zaman ve maddi açıdan uygulanabilirliği düşüktür. Bu sebeple, sayısallaştırılmış bu belgelerin otomatik çevirisi gerekmektedir. Bu çalışmada Osmanlıca el yazmalarının etkin ve kolay erişimini sağlayacak kelime erişimi sisteminin ilk aşaması olarak sayısallaştırılmış belgelerin ön işlemesi ve satırlara bölütlenmesi konusunda çalışmalar yapılmıştır. Basılı belgelerde kullanılan geleneksel satırlara bölütleme yöntemleri tarihi belgelerde çok başarılı sonuçlar getirememektedir. Bu nedenle daha gelişmiş çözümler üzerine yoğunlaşarak satır bölütlemede izdüşüm tabanlı bir yöntem geliştirilmiştir. Osmanlıca taş baskı belgelerin 120 sayfası üzerinde yapılan çalışmalarda elde edilen sonuçlar uygulanan yöntemin başarılı olduğunu göstermektedir.

ABSTRACT

Many researches and historians from all around the world are interested in historical Ottoman archives. However, translation of these documents requires competent historians which is not a feasible method in terms of time and cost. Thus, automatic translation of these documents are required. In this paper, preprocessing steps of accessing the Ottoman manuscripts with a word based search engine is studied. These preprocessing steps are binarization and line segmentation of digitalized documents. The traditional line segmentation methods applied to printed documents do not yield to satisfactory results for historical and handwritten documents. Due to this fact, more complex line segmentation techniques must be used. In this study, we developed a projection profile based method for line segmentation and local binarization is used. The experiments are conducted on a 120 page Ottoman archive and the results show that the proposed system is successful.

1. GİRİŞ

150 milyondan fazla tarihi belge içeren Osmanlıca arşivleri dünyanın değişik yerlerinden araştırmacıların ilgi alanına girmektedir. Fakat Osmanlıca belgelerin elle dizilenmesi ve etiketlenmesi zaman ve emek açısından oldukça zor bir problemdir ve araştırıcı tarafından bütün belgelerin teker teker incelenmesi gerekmektedir. Bu sebepten ötürü etiketleme ve çevirinin otomatik yapıldığı bir sistem geliştirilmelidir. Son dönemlerde, el yazması, taş baskı ve matbu biçiminde olan belgelerin sayısallaştırılmasıyla, hızlı ve kolay erişimlerini sağlamak için yöntemler geliştirilmeye başlanmıştır. Ancak tarihi belgelerin eski ve yıpranmış olmaları ve çoğu zaman el yazısı ile yazılmış olmaları bölütleme ve eşleme problemlerini daha da zorlaştırmaktadır.

Bu çalışmada amaç, büyük önem taşıyan Osmanlıca belgelerin kolay erişimini sağlayacak olan kelime erişimi sisteminin[1,2] önemli bir ön aşaması olan satırlara bölütleme konusunda tarihi belgelere uygun bir yöntem geliştirmektir.

Osmanlıca belgelerde satırlara bölütleme başka çalışmalarda da uygulanmıştır [3,4]. [3] çalışmasında belgelerin dikey izdüşümleri alınmıştır. Önceden tanımlanmış eşik değerlerine göre izdüşümdeki tepe noktaları hesaplanarak satırlar belirlenmiştir. Bir diğer çalışmada [4] ise her satıra birimler atanmıştır ve bu birimlere uygulanan çekici ve itici kuvvetler tanımlanmıştır. Birimlerin koordinatları, uygulanan kuvvetlere göre tekrarlamalı olarak kuvvetlerin toplamı yerel minimuma erişene kadar değişmektedir.

Geleneksel satırlara bölütleme yöntemleri belgelerin yapısı hakkında bazı varsayımlarda bulunup ona göre bölütleme yapmaktadırlar. Osmanlıca belgelerin taş baskı veya el yazması biçiminde olması ve yıpranmış bir yapısının olması bu varsayımların dışına çıkılmasına neden olmaktadır. Baskı biçiminde olan belgelerde satırlar arası uzaklıklar genelde sabittir ve satırlar birbirine çok yaklaşmaz. Fakat Osmanlıca da satırlar arası uzaklıklar çok daralırken, aynı sayfa içinde bile farklılıklar göstermektedir. Aynı zamanda, yıpranmış yapısından ötürü kırık harfler içermektedir ve Osmanlıca alfabesinde boyutu nokta kadar küçük olan çok fazla bileşen kullanılmaktadır. Bu sayılan özellikler geleneksel satırlara bölütleme yöntemlerinde sorunlara yol açmaktadır.

Bu çalışmada bu problemlerle başa çıkabilecek izdüşüm tabanlı bir yöntem sunulmaktadır. Dikey izdüşüm yöntemi geleneksel satırlara bölütleme yöntemlerinden biri olmasına rağmen bu çalışmada geliştirilerek tarihi belgelerin içerdiği problemleri çözecek şekilde değiştirilmiştir. Aynı zamanda bu çalışmada, dikey izdüşüm grafiklerindeki tepe noktalarını bulmak için de yeni bir yöntem sunulmaktadır.

2. ÖNERİLEN YÖNTEM

Resmin ikilenmesinde izlenilen yöntem Bölüm 2.1'de, kırık piksellerin düzeltilmesi için uygulanan morfolojik dönüşümler Bölüm 2.2'de verilecektir. Bölüm 2.3'de iz

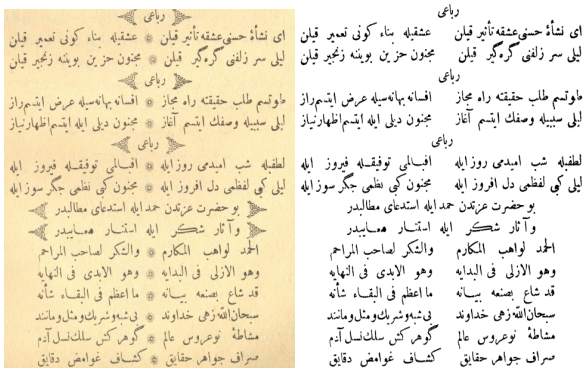
978-1-4673-0056-8/12/$26.00 ©2012 IEEE

düşüm tabanlı satırlara bölütleme yöntemi detaylıca açıklanacaktır.

2.1. Uyarlamalı İkileme

Sayısallaştırılmış belgeler önce gri tonlu biçime dönüştürülür. Sayısı az miktarda olan harf harici süslemeler elle işaretlenerek çıkarılır. Sonradan, belgeler uyarlamalı ikileme yöntemi ile ikili görüntü biçiminde saklanır (Şekil 1). Bunun için ilk olarak gri ölçekli görüntüler resmin büyüklüğüyle orantılı parçalara ayrılır ve sonraki aşamada her parçaya bağımsız olarak tekrarlamalı genel eşikleme uygulanır[5]. Parçaların sayısı arttıkça eşiklemenin daha iyi sonuç verdiği fakat daha yavaş çalıştığı gözlemlenmiştir.

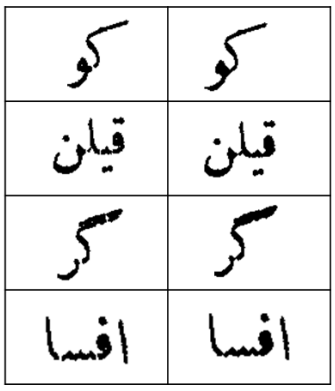
Uyarlamalı ikileme yerine tek bir eşik değeri kullanılarak genel ikileme de yapılabilir fakat uyarlamalı ikilemenin yıpranmış, tarihi belgelerde daha başarılı sonuçlar çıkardığı gözlemlenmiştir.



Şekil 1: Osmanlıca belgelerden bir örnek ve ikileme uygulanmasından sonraki görüntü.

2.2. Morfolojik Dönüşümler

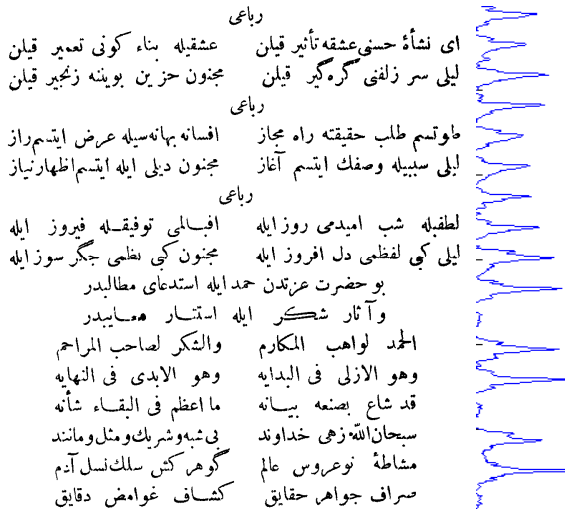
Yıpranmışlıktan kaynaklanan kırık harfleri düzeltmek için basit morfolojik dönüşümler uygulanmıştır. İkilenmiş görüntülere görüntülerin boyutuyla orantılı küçük yapılar ile kapama uygulanmıştır (Şekil 2).



Şekil 2: Kırık bileşenler ve morfolojik dönüşümler sonrasında düzeltilen bileşenler.

2.3. Satırlara Bölütleme

Yaklaşık sonuçları elde etmek için, ikilenmiş görüntülerin dikey izdüşümleri alınır. Şekil 3'te de görüldüğü gibi izdüşüm grafiklerinde vadiler satırlar arasındaki boşlukları, tepeler ise satırları temsil etmektedir. Öncelikle, bu izdüşüm grafiklerindeki değeri sıfır olan bölgeler bulunur. İki satırın arasında küçük bileşenler yoksa bu sıfır bölgeleri kesin olarak satırlar arasındaki boşluğu bulduğumuzu göstermektedir.



Şekil 3: İkili Osmanlıca görüntü ve onun dikey izdüşüm grafiği.

Yaklaşık biçimde satırları bulmuş olsak da yukarıda bahsettiğimiz sebepler yüzünden sonuçlarda birçok yanılgılar ortaya çıkmaktadır (Şekil 4). Bu yanılgıları düzeltmek için, hatalı saptanan satırlar iki kategoriye ayrılmıştır. Birinci kategori birden fazla satır içeren bölgelerin tek satır olarak algılanması, ikinci kategori ise satırlar arasındaki küçük bileşenlerin herhangi bir satıra ait olmadan tek başına bir satır olarak algılanmasıdır. Şekil 4-a'da iki kategoriye düşen satırlar gözlemlenebilir.

Birinci kategorideki hatalı saptanan satırların bulunması için belgedeki bulunan satırların yüksekliklerinin sıklık grafiği çıkarılarak dağılımına bakılır. Birden fazla satır içerip tek satır olarak algılanan hatalı satırların sayısının doğru saptanan satırlardan daha az olacağı varsayımında bulunarak sıklık grafiğindeki en kalabalık bölgenin doğru bulunan satırları temsil ettiğini söyleyebiliriz. Buradan yola çıkarak, satır yüksekliklerinin dağılımındaki aykırı değerlere sahip olan satırlar önceden tanımlanmış bir eşik değeri kullanılarak yanlış saptanan satırlar olarak otomatik biçimde işaretlenir. Fakat doğru tespit edilen satırlardan yüksekliği çok olan satırlar da bu aralığa düşebildiği için bu aykırı satırlar sonradan yine incelenmek üzere yanlış satır adayı olarak işaretlenir.

Şekil 4-a: Dikey izdüşümden elde edilen yaklaşık satırlar ve *b:* yanıltıcı satırların düzeltilmesinin ardından elde edilen sonuçlar.

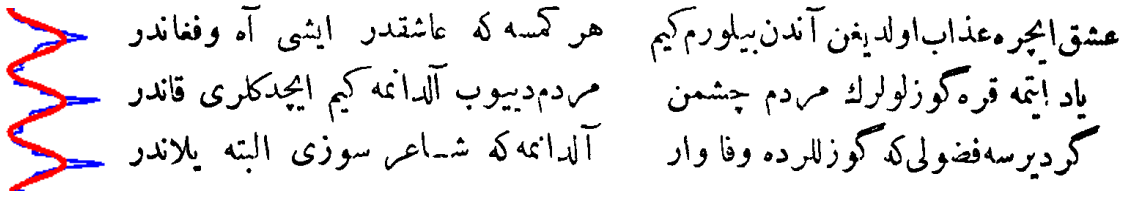
Benzer bir şekilde ikinci kategoriye düşen, yüksekliği çok düşük olan satırlar da saptanan satırların yükseklik dağılımına bakılarak tespit edilir ve işaretlenir.

Yanlış satır adaylarının başlangıç ve bitiş koordinatları kullanılarak belgeden çıkarılır ve tek satır olarak tespit edilip birden fazla satır içeren bu bölgenin dikey izdüşümü bulunur. Sonradan, izdüşümüne Şekil 5'te görüldüğü gibi Fourier eğrisi oturtulur[6]. Bu aşamada polinomsal eğri kullanılırsa yanlış tespit edilen satırların sayısına göre eğrinin derecesini tayin etmek gerekmektedir. Fourier eğrilerde ise herhangi bir parametreye ihtiyaç duyulmamaktadır bu sebeple Fourier eğrisi oturtulması uygun görülmüştür.

Hesaplanan Fourier eğrisindeki yerel maksimumlar satırları temsil etmektedir ve bir tane yerel maksimum tespit edilmesi durumunda yanlış adayı olarak işaretlenen satırın doğru tespit edildiği anlaşılır. Birden fazla satır içeren sonuçların birbirinden ayrılması için yerel maksimum noktaları arasında kalan dikey izdüşümündeki minimum noktanın koordinatı kullanılır. Eğri oturtma yöntemiyle ikiden fazla tek satır olarak yanlış tespit edilmiş satırların ayrılması problemi de çözülmüş olur.

Sonraki aşamada, ikinci kategoriye düşen yükseklikleri çok küçük olup küçük bileşenler içeren satırların doğru satır ile birleştirilmesi vardır. Satırların yükseklik dağılımından önceden tespit edilen bu satırlar en yakın olduğu satıra eklenir.

Son olarak, bütün bağlantılı bileşenlerin en yakın olduğu satırlar hesaplanır ve bileşenler bu satırlara atanırlar. Satırların tespit edilen koordinatlarını kullanarak direk ikili resimden kesmek bazı bileşenlerin kesilmesine yol açmaktadır. Bu sebepten ötürü bağlantılı bileşenlerin en yakın olduğu satırlar saptanır ve bileşenler bu şekilde satırlara gruplanmış olur (Şekil 6).



Şekil 5: Tek satır olarak algılanan satırların dikey izdüşümü ve izdüşümüne oturtulan Fourier eğrisi.

3. DENEYLER

3.1. Veri Kümesi

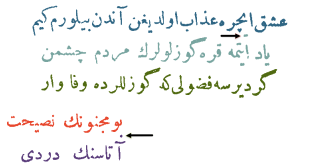
Önerilen yöntemleri test etmek için 120 sayfalık Osmanlıca bir veri kümesi oluşturulmuştur. Belgelerde yüksek derecede eğimli taranma gözlemlenmemiştir. Aynı zamanda, önerilen yöntem belli bir miktar açıya kadar satırları doğru hesaplamaktadır. Bu sebeplerden ötürü, olası dönmelere karşı belgelerin düzeltilmesine gerek görülmemiştir. Veri kümesi oluşturulurken, satırlara ayrılması kolay olan belgelerin yanında satıları birbirine yakın ve açılı satırları bulunan belgeler de seçilmiştir.

3.2. Satırlara Bölütleme Sonuçları

Veri kümesindeki belgelerde toplam 3210 satır vardır ve elde edilen sonuçlarda da 3210 satır bulunmuştur. Veri kümesinin 20 sayfalık kısmı elle bölütlenmiş biçimdedir. Referans alınan bu kümede ise 539 satır vardır. Elde edilen sonuçlar referans kümesiyle karşılaştırıldığında tespit edilen satır sayısı aynıdır. Veri kümesindeki belgelerde satırlar arası boşluklar değişiklik göstermesine rağmen sonuçlarda tek satır olarak tespit edilen birden fazla satır veya satır olmayan yerlerde satırlar saptanmamıştır. Örnek bir satırlara bölütleme Şekil 6'da görülmektedir.

Şekil 6: Tespit edilen satırlar.

Fakat saptanan satırların %4 lük kısmında ufak hatalar vardır. Hataların büyük kısmı küçük boyuttaki bileşenlerin yanlış satırlara atanmasından kaynaklanmaktadır (Şekil 7). Fakat bu küçük bileşenlerin hangi satıra ait olduğunu Osmanlıca dilinden bağımsız olarak insan gözüyle bile ayırmak zor bir işlemdir.



Şekil 7: Yanlış satırlara atanan bileşenler.

Hataların düzeltilmesi için, bileşenleri en yakın olduğu satıra atamak yerine Osmanlıca diline özgü bir metrik kullanılabilir. Başka bir yöntem ise, en baştan küçük bileşenleri resimden ayıklayıp satırları onlar olmadan bölütlemektir. Daha sonradan kelime eşleme aşamasına geçerken bölütlenen satırların yüksek oranda doğruluğunu sağlamak için Osmanlıca bilen bir araştırıcı tarafından ayıklanan küçük bileşenler ait oldukları satırlara atanabilir.

4. ÖZET VE TARTIŞMA

Sunulan yöntemin genel başarısını yüzdelerde ifade etmek gerekirse, referans alınan 20 sayfalık veri kümesinin %96'sı başarıyla satırlara bölütlenmiştir. Hatalı tespit edilen satırlarda bulunan bileşenlerin çoğu doğru satırlarda saptanmıştır, sadece bazı satırlar için fazladan veya eksik bileşenler vardır.

Satırlara bölütlemede Osmanlıca veri kümesindeki problemler dikkate alınarak dikey izdüşüm tabanlı bir yöntem geliştirilmiştir. Satırların birbirine çok yakın olması ve çok fazla küçük boyutta bileşen içermesi sebebiyle ortaya çıkan yanıltıcı satırlar sonradan işlenerek düzeltilmiştir. Sonuçlarda elde edilen hatalı satırların Osmanlıca bilen kişiler tarafından düzeltilmesi ve kelime eşleme seviyesinde analizi için başarı oranının %100'e çıkarılması oldukça kolaydır.

Aynı zamanda, satırlara ayrılması daha zor olan Osmanlıca veri kümeleri için sunulan yöntem geliştirilerek kullanılabilir. Önerilen yöntemlerde Osmanlıca diline bağlı bir kural kullanılmadığından sadece Osmanlıca belgelerde değil başka dillerdeki tarihi belgelerde de benzer sonuçlar vermesi ön görülmektedir. Gelecekteki çalışmalarda önerilen yöntemin el yazması belgeler için geliştirilmesi planlanmaktadır.

5. TEŞEKKÜR

Bu çalışma TÜBİTAK 109E006 nolu proje tarafından desteklenmiştir. Ayrıca, kullanılan Osmanlıca veri kümesini satırlara ayıran Bilkent Üniversitesi Türk Edebiyatı Bölümü'nden Meriç Kurtuluş'a teşekkür ederiz.

6. KAYNAKÇA

[1] Arifoglu, D.; Duygulu, P.; , "Word retrieval in ottoman documents," IEEE 19th Conference on Signal Processing and Communications Applications, vol., no., pp.526-529, 20-22 April 2011.

[2] Ataer, E.; Duygulu, P.; , "Matching Ottoman Words," IEEE 15th Signal Processing and Communications Applications, vol., no., pp.1-4, 11-13 June 2007.

[3] E. Ataer, P. Duygulu, "Retrieval of Ottoman documents", Proceedings of the 8th ACM international workshop on Multimedia information retrieval, October 26-27, 2006.

[4] E. Öztop, A.Y. Mülayim, V. Atalay, F. Yarman-Vural, "Repulsive attractive network for baseline extraction on document images" , Signal Processing, Volume 75, Issue 1, 5 January 1999.

[5] Kavallieratou E., Stathis S. "Adaptive Binarization of Historical Document Images", IEEE proceedings of 18th International Conference on Pattern Recognition (ICPR'06), pp. 742-745, 2006.

[6] Bochner S., Chandrasekharan K., "Fourier Transforms", Princeton Univ. Press, Princeton, 1949.