El Yazısı Belgelerde Kelime Tabanlı Arama

Word Spotting in Historical Documents

Ethem F. Can, Pınar Duygulu

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

Abstract

We present new methods to retrieve words in historical handwritten documents. With the assumption that the words can be seen as images, we used the word spotting idea and search for the words in the documents using image retrieval techniques. Specifically, we proposed two methods, one based on the histogram of gradient orientations and one based on the correlation coefficient. We also proposed a new method by combining these two methods. In the experiments the data set consisting of George Washington's handwritings is used.

Özetçe

Bu çalışmada el yazısı belgelerde arama yapabilmek için yeni yöntemler önerilmiştir. Bu çalışmadaki en temel varsayım ve yola çıkış noktası; her bir kelimenin resim gibi ele alınabileceği ve dolayısıyla resim arama teknikleri ile sorgulama yapılabileceğidir. Özel olarak resim üzerindeki kenar noktalarının eğimlerinin yönlerinin dağılımı ve korelasyon katsayısı tabanlı iki yöntem önerilmiş, ayrıca bu iki yöntemin nasıl birleştirilebileceği anlatılmıştır. Deneyler George Washington'un el yazmaları veri kümesi üzerinde yapılmıştır.

1. Giriş

El yazması belgelerin sayısal ortamlara aktarılmasıyla, bu belgelere hızlı ve kolay erişime olan ihtiyaç ön plana çıkmıştır.

Dökümanların elle dizilenmesi ve etiketlenmesi çok emek gerektiren ve zaman alan bir problemdir. Öte yandan karakter tanıma yöntemleri baskı belgelerde iyi çalışmasına rağmen, el yazmalarında istenilen sonuçları vermemektedir.

Son yıllarda, alternatif bir yöntem olarak el yazısı belgelerin resim gibi ele alınabileceği, ve kelimelerin resim arama teknikleri ile bulunabileceği gösterilmiştir [1,2,3,4,5].

Rath ve Manmatha [2] *'dinamik zaman bükmesi'* olarak adlandırılan yöntemi resim bazında kelime sorgusu için kullanmıştır. Bu teknik kelime imgeleri arasındaki benzerlikleri, uzunluklar arasında fark olsa bile birbirine uydurmaya çalışarak yakalamayı hedeflemektedir. Bu yöntem verdiği iyi sonuçlara rağmen, çok yavaş olmasından dolayı tercih edilmemiştir.

Ataer ve Duygulu [3,4] çalışmalarında dikey izdüşüm profilleri yöntemini kullanarak kelime eşleştirmesi yapmıştır ve sonuç olarak resmin basit özelliklerini kullanarak resim bazlı kelime sorgusu yapılabileceğini göstermişlerdir.

Bu noktadan yola çıkarak, bu çalışmada belge aramada resim bazlı kelime sorgusu yöntemi seçildi ve kelime sorgusu yapmak için iki değişik yöntem önerildi: eğim yönleri dağılımı tabanlı eşleme ve korelasyon katsayısı tabanlı eşleme. Ayrıca üçüncü bir yöntem olarak iki yöntem birleştirildi.

Aşağıda öncelikle önerilen yöntemlerin detayları açıklanacak, daha sonra benzer çalışmalarda kullanılmış olan George Washington'un el yazmaları veri kümesi [6] üzerindeki deneysel sonuçlar açıklanacaktır. Kullanılan yöntemler resmin basit özelliklerini kullanarak eşleme yapılması mantığı üzerine kurulmuş olup, ileride yapılacak daha karmaşık ve iyi özniteliklerin kullanılacağı çalışmalara temel oluşturmaktadır.

2. Önerilen Yöntemler

2.1. Eğim Yönleri Dağılımı Tabanlı Eşleme

Lowe ve ekibi [7] ile Mikolajczyk ve ekibi [8] çalışmalarında anahtar noktalar çevresindeki eğimlerin yönünü hesaplamış ve bunların dağılımını nesne eşleme ve tanıma konusunda öznitelik olarak kullanmışlardı.

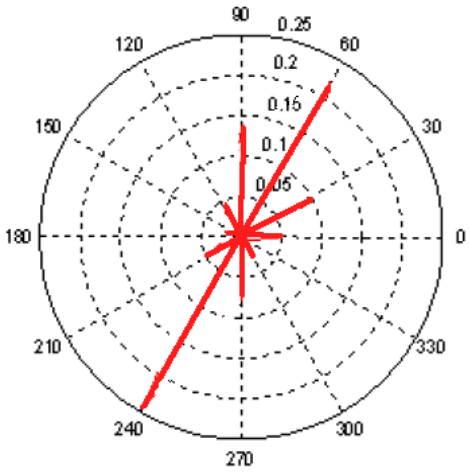
Bu fikirden esinlenerek, kelime eşlemede ilk yöntem olarak resimdeki kenar noktaları için eğim yönlerinin dağılımı bilgisi kullanıldı.

Bu amaçla öncelikle her bir kelime resmi üzerinde *Canny* kenar bulma algoritması uygulanarak kenar noktaları bulundu (Şekil 1). Daha sonra bu noktalara *Sobel* filtresi uygulanarak eğimler hesaplandı.

(a)

(b)

Şekil 1 – (a) Ayıklanmış kelime resmi (b) Resim üzerinde bulunan kenar noktaları



Şekil 2 – Eğim yönlerinin dağılımı

Bir sonraki aşamada her bir kelime resmi sabit uzunluk ve genişlik değerine sahip kutulara ayrıldı ve her kutu parçasındaki eğim yönlerinin dağılımı bulundu. Bu amaçla eğim yönleri 30'ar derecelik açılara denk gelicek şekilde birimlendiridi ve kutu içindeki eğim yönlerinin her birimdeki sayıları hesaplandı (Şekil 2).

Resimler arasındaki farklar hesaplanırken öncelikli olarak hiyerarşik bir düzenlemeyle eşit sayıda bölüme ayrılmış kelimeler seçildi. Daha sonra her bir bölümdeki dağılımlar ard arda eklenerek dağılımlar arasındaki farklar Öklid uzaklığına göre hesaplandı. Bu değer daha sonra resimlerin benzerliklerine gore sıralanması için kullanıldı.

2.2 Korelasyon Katsayısı

Bu method ilk defa Choo ve Kang [9] tarafından uygulanmıştır. Verilen iki resim arasındaki benzerlik aşağıdaki gibi hesaplanır.

$$\frac{\sum_m \sum_n (F_{mn} - \bar{F})(G_{mn} - \bar{G})}{\sqrt{[\sum_m \sum_n (F_{mn} - \bar{F})^2][\sum_m \sum_n (G_{mn} - \bar{G})^2]}} \quad (1)$$

F ve G resimleri, m ve n resimlerin en ve boylarını, ve $\bar{F}$ ve $\bar{G}$ de F, G resimlerinin ortalama gri seviyesini göstermektedir.

İki resim arasında hesaplanan korelasyon katsayısı 0 ile 1 arasındadır ve 0 benzerliğin olmadığını 1 ise benzerliğin mükemmel olduğunu göstermektedir.

Kullandığımız test setindeki yazıların hemen hemen aynı eğikliğe ve büyüklüğe sahip olması bu yöntemi bizim çalışmalarımızda da kullanılabilir hale getirdi. Bu amaçla kelime resimleri arasında elde edilen korelasyon katsıyısı değerleri 1'e en yakından en uzağa olmak üzere sıralandı ve en benzer resimler belirlendi.

2.3. İki Yöntemin Beraber Kullanılması

Çalışmamızın bu kısmında yukarıda bahsettiğimiz iki yöntem beraber kullanıldı. Eğim yönteminin yeterince iyi sonuçlar vermemesi ve korelasyon katsayısı yönteminin bazı kelimeleri kaçırması ve çok hızlı calışan bir yöntem olmaması bizi bu iki yöntemi beraber kullanmaya itti.

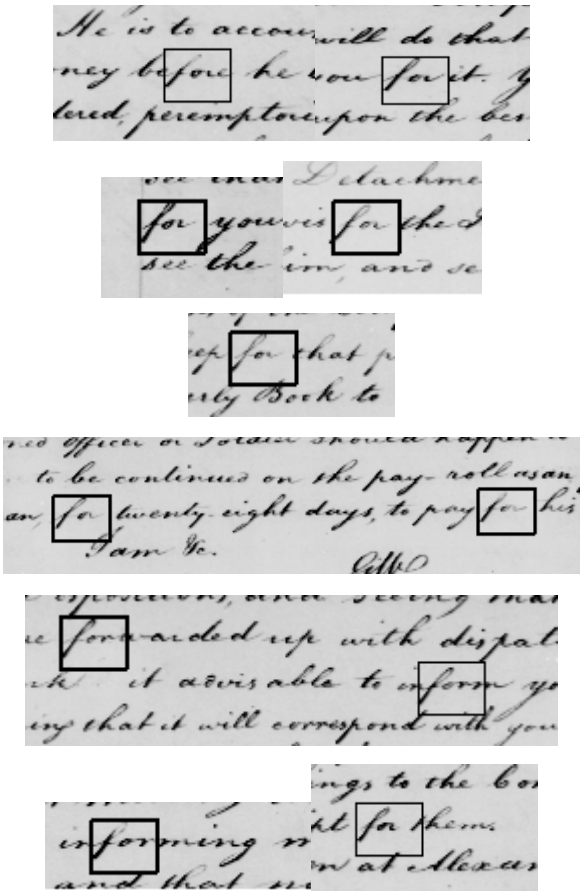
Öncelikle resimlerin boyutlarının çok büyük olması nedeni ile resimler 1/4 oranında küçültüldü. Bu amaçla Bilinear Interpolation yöntemi kullanıldı [10]. Aranılacak kelime için her bir sayfanın her bir noktası için korelasyon katsayısı hesaplandı ve belli bir eşik değerinin üstündeki noktalar bulundu. Bulunan bu noktaların orjinal resimlerdeki yerleri hesaplandı ve bu noktalarda yeniden korelasyon katsayısı hesaplandı. Son olarak belli bir değerin üzerindeki noktalar 2.1 bölümünde anlattığımız şekilde benzerliklerine göre sıralandı.

3. Deneyler

Deneylerde daha önce benzer çalışmalarda da kullanılmış olan George Washington'un el yazmaları veri kümesi [6] kullanıldı. Bu veri kümesi daha önceden ayrılmış 4680 kelime içermektedir.

Deneylerde 4680 kelimenin her biri için sorgu yapıldı ve sonuçlar hesaplandı. Bütün kelimelerde yapılan deneylerde elde edilen sonuçlara göre birinci yöntem %42.6 başarı , ikinci yöntem %48.6 başarı ve ikisinin beraber kullandığı yöntem ise %54.9 başarı verdi.

Örnek olarak; 'For' kelimesi için yapılan sorguda, iki yöntemin birlikte kullanıldığı deneylerde toplamda bulunan 104 tane 'for' kelimesinden (başlı başına kelime olanlar ve başka bir kelimenin bir parçası olanların toplamı), 64 tane doğru sonuçla beraber 8 tane 'from' kelimesi, 21 tane 'you' kelimesi ve 21 tane farklı yanlış sonuç dönmüştür. Bu sorgunun sonucundaki örnek doğru sonuçlar Şekil.3'te, yanlış sonuçlar ise Şekil.4'te gösterilmiştir.



Şekil 3 - 'for' kelimesi için doğru sonuçlar

Şekil 4- 'for' kelimesi için yanlış sonuçlar

Korelasyon katsayısı yöntemi ile bulunan yanlış sonuçlar, eğim yönlerinin dağılımı bilgisinin eklenmesi ile büyük oranda elenebilmiştir. Ancak hala bazı yanlış sonuçlar elenememektedir. Örnek olarak 'furnished' kelimesi iki yöntemde de yanlış sonuçlar arasında görülmüştür. Bu kelimenin 'for' kelimesinin boyutunu kapsayacak yere kadar olan kısmının sorgu yaptığımız kelime ile çok benzer olması sorguda yanlış sonuçlar içinde yer almasına neden olmuştur.

5. Özet ve Tartışma

Çalışmamızda, el yazısı dökümanlarda resim bazında kelime tabanlı arama için yeni bir yöntem önerildi. Bu amaçla resim üzerindeki eğim yönlerinin dağılımı ve korelasyon katsayısı tabanlı yöntemler karşılaştırıldı ve birleştirildi. George Washington el yazmaları veri kümesi üzerinde yapılan deneylerde umut verici sonuçlar elde edildi. Deneyler birleşik yöntemin diğer iki yöntemin ayrı ayrı kullanılmasına gore daha iyi sonuçlar verdiğini gösterdi.

Aynı veri kümesi üzerindeki çalışmalarında, Rath ve Manmatha [2] 15 kelime ile yapılan testlerde %90.72 ve 2372 kelime ile yapılan testlerde ise %71.11 başarı elde etmişlerdir. Kelime sayısı arttığında başarının düşmesi normaldir. Bu nedenle çalışmada elde edilen sonuçlar da tatmin edicidir. Ayrıca önerilen yöntem Rath ve Manmatha'nın yöntemine göre çok daha hızlı çalışmaktadır.

İlerki çalışmalarımızda eğim yönlerinin dağılımı yöntemini geliştirmeyi hedefliyoruz. Ayrıca eğim yönü bilgisi yerine çizgi ve eğim gruplarının sayılarının kullanılmasını planlıyoruz.

6. Teşekkür

Bu çalışma 104E077 ve 104E065 numaralı TUBİTAK projeleri tarafından desteklenmektedir.

7. Kaynakça

[1] R. Manmatha, C. Han, and E. M. Riseman. Word Spotting: A New Approach to Indexing Handwriting. In Proceedings of the IEEE Computer Vision and Pattern Recognition Conference, 1995.

[2] T. M. Rath and R. Manmatha. Word Image Matching Using Dynamic Time Warping. In CVPR (2), pages 521{527, 2003.

[3] E. Ataer and P. Duygulu. Matching Ottoman Words: An Image Retrieval Approach to Historical Document Indexing. In Proceedings of the International Conference on Image and Video Retrieval, 2007.

[4] E. Ataer and P. Duygulu. Retrieval of Ottoman Documents. In 8th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2006.

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

[6] Center for Intelligent Information Retrieval, Univ. of Massachusetts Amherst http://ciir.cs.umass.edu/downloads/

[7] D. G. Lowe. Distinctive Image Features From Scale-Invariant Keypoints. International Journal on Computer Vision, 60(2), 2004.

[8] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. IEEE Transactions on Pattern Analysis Machine Intelligence, 27(10): 1615-1630, 2005

[9] Y.J. Choo and B. S. Kang. "The characteristics of the particle position along an optical axis in partical holography. " Meas. Sci. Technol, 17, 761-770 (2006)

[10] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. Numerical Recipes in C. Cambridge University Press, Cambridge, NY, USA, 1992.