http://fbe.trakya.edu.tr/tujs
Trakya Univ J Sci, 10(1):97-102, 2009
ISSN 1305–6468
DIC: 264MŞET1010906090709

Araştırma Makalesi / Research Article

BİRLEŞİK VE EĞİK TÜRKÇE EL YAZISI TANIMADA *K-NN* SINIFLAMA YÖNTEMİ VE SÖZLÜK KULLANIMI

Murat ŞEKERCİ*[1], Rembiye KANDEMİR[2]

[1] Trakya Üniversitesi Fen Bilimleri Enstitüsü, EDİRNE

[2] Trakya Üniversitesi Mühendislik-Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü, EDİRNE

*e-mail: muratsekerci@hotmail.com

Alınış: 11 Şubat 2009
Kabul Ediliş: 12 Haziran 2009

Özet: *Bu çalışma, henüz tam olarak çözülememiş problem olan birleşik ve eğik Türkçe el yazısı üzerinedir. El yazısı tanımadaki zorluk, kişiden kişiye yazım farklılıkları göstermesi ve harflerin birbirine bitişik yazılmasından kaynaklanmaktadır. Ayrıca Türkçe'nin eklemeli kelime yapısına sahip olması da bu zorluğu arttırmaktadır. Tanıma sisteminde, küçük harflerle yazılmış el yazısı kullanılmıştır. Karakter tanıma aşamasında, sınıflama için k-NN' den yararlanılmıştır. Kelimelerin tanınmasında, sözlük ve karakterlerin bölütlenmesi birlikte kullanılmıştır. Sözlük kullanımı ile kelime doğrulama aşamasında anlamsız harflerin seçilmesi engellenmiş ve yanlış tanınan kelimelerin düzeltilmesi sağlanmıştır. Çalışmadaki karakter tanıma performansı %90.5 iken kelime tanıma performansı %84 olarak elde edilmiştir. Elde edilen kelime tanıma performansının daha düşük olması çalışmada kullanılan sözcükteki kelime sayısının sınırlı olmasından kaynaklanmaktadır.*
Anahtar Kelimeler: *Türkçe el yazısı tanıma, birleşik yazı, karakter tanıma, k-NN sınıflama, sözlük*

Touching-Sloping Turkish Handwriten Text Recognition Using K-Nn Classification Method and Lexicon

Abstract: *This study is dealt with Turkish handwritten touching-sloping text recognition. The difficulty of handwritten recognition depends on changing of handwritten person by person and touching-sloping written characters. Also, agglutinative word structure of Turkish language increases difficulty of recognition. It was used lowercase handwritten for recognition system. It was used k-NN for character recognition stage. Character segmentation and lexicon were used together for word recognition. It was blocked choosing incorrect letters using lexicon and corrected recognition of incorrect words. In the study, while performance of character recognition was obtained 90.5%, performance of word recognition was obtained 84%. The lower value of performance of word recognition obtained depends on restricted word in lexicon used for the study.*
Keywords: *Turkish handwiten recognition, touching handwriting, character recognition,k-NN classification, lexicon*

Giriş

Karakter Tanıma Sistemi (KTS), makineyle veya elle yazılmış yazıların bilgisayar yardımıyla tanınması işlemi olarak adlandırılmaktadır. KTS uygulamaları veri girişinin türüne göre çevrimdışı veya çevrimiçi olarak işlenmektedirler. Çevrimdışı işlem sırasında veri girişi olarak yazılı belgenin sayısal imgesi verilirken; çevrimiçi işlemde ise yazım esnasında elde edilen bilgiler kullanılmaktadır.

Son yıllarda, yazı tanıma konusunda yapılan çalışmalarda büyük ilerleme kaydedilmiştir. Bu sayede özellikle temiz ve okunabilir bir zemin üzerinde matbaa ürünü daktilo veya bilgisayarda yazılmış yazıları otomatik olarak tanıyan programlar hayatımızın içine girmeye başlamıştır. Tanıma oranı yüksek olan bu programlar birçok kurum ve şirkette maliyeti düşürmekte ve hayatı oldukça kolaylaştırmaktadır. Bu gelişmelerin yanı sıra, şu anki teknoloji ile el yazısı tanıma problemi hala tam olarak çözülmüş değildir. El yazısı tanımadaki zorluk, çok fazla sayıda değişik yazı karakteri olması ve kişiden kişiye farklılıklar göstermesinin yanında harflerin birbirine bağlı yazılmasından da

kaynaklanmaktadır. Yazı sitili, duruma ve kullanılan kalem ya da kağıda göre de farklılıklar gösterebilmektedir. Kişilerin yazı yazma tarzlarına ve hızlarına bağlı olarak harfler çok değişik şekil ve büyüklükte olabilmektedir. İnsan görme sistemi harflerin büyüklük ve yön farklılıklarından etkilenmemekte, oysa otomatik bir sistemde bunlar büyük sorunlar oluşturmaktadır (Şekerci, 2007).

Bununla beraber İngilizce, Çince, Arapça gibi bazı dillerde, oldukça iyi düzeyde çalışan el yazısı tanıma sistemleri geliştirilmiştir (Senior ve Robinson, 1998 – Shridhar, Houle ve Kimura, 1997 – Gunter ve Bunke, 2005 – Verma, Blumenstein ve Kulkarni, 1998). Fakat Türkçe el yazısı üzerine yapılan çalışmalar sınırlı sayıdadır. Yanıkoğlu ve Kholmatov (2003) sınırlı sayıda kelime gurubu üzerinde yaptıkları çalışmalarında, ana sözlük yöntemine göre ve yazma tipine göre %53 ile %57 arasında değişen kelime tanıma performansı elde etmişlerdir. Çapar ve Ark. (2003), çeşitli öznitelik bulma ve sınıflandırma teknikleri kullanarak büyük harflerle yazılmış Türkçe el yazısı tanıma üzerine çalışmışlar ve SDNLL (Size Dependent Negative-Log-Likelihood) sınıflandırıcısı ile %93.6 başarım oranı elde etmişlerdir. Vural ve Ark. (2004) çevrimiçi Tablet PC üzerinde, az sayıda kelimeden oluşan bir sözlük listesini Gizli Markov Modelleri kullanarak Türkçe'yi tanıyan bir prototip uygulama sunmuşlardır. Erdem ve Uzun (2005), yapay sinir ağlarını kullanarak Türkçe times new roman, arial ve ayrık yazılmış el yazısını tanıma üzerine çalışmışlardır.

Bu çalışmada, çevrimdışı olarak çalışan birleşik ve eğik olarak yazılan Türkçe el yazısını tanıyan sistem Visual Basic.Net ortamında geliştirilmiştir. Bu sistem, yazı görüntüsü üzerinde gürültülerin temizlenmesi, satırların ve kelimelerin ayrıştırılması, kelime eğiminin düzeltilmesi, karakterlerin ayrıştırılması, karakterlerin ve kelimelerin tanınması aşamalarından oluşmaktadır. Karakter tanıma aşamasında korelasyon yöntemi kullanılarak, *k*-NN (*k*-en yakın komşuluk) sınıflama yöntemi ile tanıma oranı arttırılmıştır. Kelime tanıma aşamasında farklı türlere sahip kitaplardan rastgele seçilmiş 2500 kelime içeren bir sözlük kullanılarak anlamsız harf gruplarının seçilmesi engellenmiş ve yanlış tanınan kelimeler düzeltilmiştir.

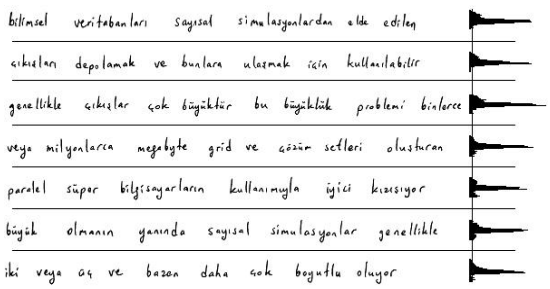
Veri Toplama ve Karakterlerin Standart Hale Getirilmesi

Yazı tanıma işleminde kullanılmak üzere farklı kişilerden el yazısı metin örnekleri toplanmıştır. Metin örnekleri 300 dpi çözünürlüğünde örüntü dosyalarına dönüştürülmüştür. Karakterleri tanıma uygulaması için de 172 kişiden 29 adet küçük harf olmak üzere toplam 4988 harften oluşan havuz oluşturulmuş ve bu harflerden 2900 adedi eğitim amaçlı, 2088 adedi test amaçlı kullanılmıştır.

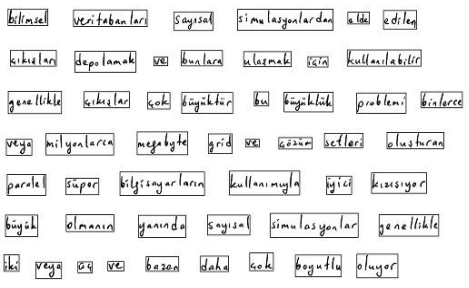
Farklı kişilerden alınan örnek harfler aynı standartta yazılmadığından bu karakterleri standart hale dönüştürmek için karakter görüntüsü, birbirine eşit 15 yatay ve 15 dikey dilime bölünerek 225 alan meydana getirilmiştir. Her bir alan incelenerek içinde siyah piksel bulunduran alan 1, hiç siyah piksel bulundurmayan alan ise 0 ile ifade edilmiştir. Elde edilen bu karakter görüntüleri, karakter tanıma işleminde kullanılmaktadır.

Satırların ve Kelimelerin Ayrıştırılması

İkili seviyeye indirgenen ve gürültüleri temizlenen görüntünün histogram bilgisinden yararlanarak ilk önce satırlar daha sonra da kelimeler ayrıştırılmaktadır (Şekerci ve Kandemir, 2006). Çalışmada bir metindeki satırların ayrıştırılması Şekil 1'de, kelimelerin ayrıştırılması Şekil 2'de gösterilmektedir.



Şekil 1. Satırların bulunması



Şekil 2. Kelimelerin ayrıştırılması

Kelime Eğiminin Düzeltilmesi

Eğim düzeltme işlemi, ön işlemenin önemli basamaklarından birisidir. Tasarlanan el yazısı tanıma sisteminde, eğimsiz kelimeler üzerinde dilimleme işlemi yapıldığından, kelimelerin eğimlerinin düzeltilmesi gerekmektedir.

Kelime eğimlerinin düzeltilmesi işleminde aday kelimeye, (1) denklemi kullanılarak -45 ile +45 derece arası 5 derecelik artımlar ile eğim düzeltme işlemi uygulanmaktadır.

$$\begin{pmatrix} x' = \\ y' \end{pmatrix} \quad \begin{pmatrix} 1 & -\tan\beta \\ 0 & 1 \end{pmatrix} \tag{1}$$

Her bir basamakta üretilen kelime görüntüsünün dikey doğrultudaki histogram grafiği çıkartılmaktadır. Üretilen histogramlardan en çok vadi sayısına sahip olan bulunarak bu histogramın hangi açı ile eğim düzeltme işlemine tabi tutulduğu saptamaktadır. Bu açı, aday kelimenin eğim açısı olarak kabul edilmekte ve kelime düzeltilmektedir. Aday kelimenin eğimi düzeltilmiş durumu Şekil 3 'de görülmektedir.

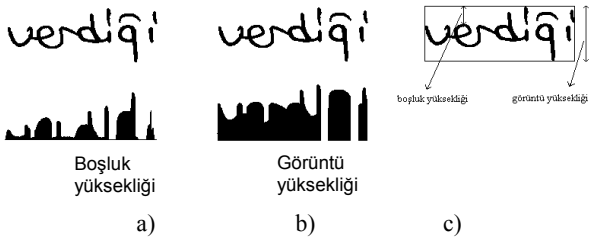
Şekil 3. a) orijinal kelime, b) eğimi düzeltilmiş kelime

Karakter Ayrıştırması

Karakterleri ayrıştırmak için eğimi düzeltilen kelimeler dilimleme işlemine tabi tutulmaktadır (Shi ve Govlndaraju,1997). Çalışmada karakter ayrıştırma işlemi için iki tip histogram grafiği kullanılmaktadır. Bu histogramlar (2) ve (3) formülleri ile elde edilmektedir. Şekil 4.'te aynı kelime ve iki farklı histogram tipi için elde edilen grafikler verilmektedir.

Histogram 1 için :
Sütun yüksekliği = ilk ve son siyah piksel arasındaki piksel mesafesi (2)

Histogram 2 için :
Sütun yüksekliği = Görüntü yüksekliği – Boşluk yüksekliği (3)



Şekil 4. Histogram tipleri. a) Histogram-1, b) Histogram-2, c) Seçilen kelime

Tanınacak kelimeye ilk önce Histogram-2 ile dilimleme işlemi yapılmaktadır. Dilimleme sonrasında (4) formülü yardımı ile dilimlerin bir karakter veya birden fazla karakter içerip içermediği belirlenir. Eğer dilimler bir karakter genişliğinden fazla ise dilimler histogram-1 kullanılarak tekrar dilimleme işlemine tabi tutulmaktadır.

Muhtemel karakter genişliği = (Üst Taban Çizgisi – Alt Taban Çizgisi) * 2.5 (4)

Fakat çalışmadaki uygulama sonuçları sadece bir eşik değeri belirlenerek yapılan dilimleme işleminin güvenilir olmadığını göstermektedir. Çünkü, kelimenin yazım şekli kişiden kişiye değişmektedir. Bir yazardan alınan kelimeyi doğru dilimleyen eşik değerinin başka bir yazardan alınan kelimeyi doğru dilimleyememektedir. Bu problemi çözmek için geliştirilen uygulamada 10 farklı eşik değerinden yararlanılmaktadır. Dilimleme işlemlerinde kullanılan eşik değeri ve bu eşik değerlerine göre hangi histogramların kullanılacağı Tablo 1'de verilmektedir.

Tablo 1. Farklı eşik değerleri ve kullandıkları histogramlar

Eşik Değeri	**Histogram -1**	**Histogram -2**
5		√
10	√	√
10	√	
Üst Taban Çizgisi		√
Üst Taban Çizgisi	√	√
Üst Taban Çizgisi – 10		√
Üst Taban Çizgisi – 10	√	√
Üst Taban Çizgisi – 20		√
Üst Taban Çizgisi – 20	√	√
Görüntü Yüksekliği – 20	√	

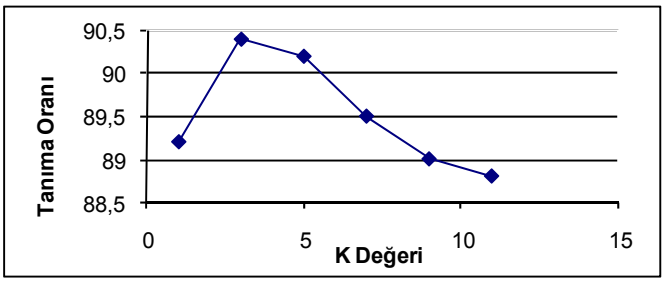
Histogram-2 ve Histogram-1 kullanılarak **"edirnede"** kelimesinin nasıl dilimlendiği Şekil 5.'te gösterilmektedir. Histogram-2 kullanılarak dilimleme işlemi sonucu üç dilim üretmektedir. Fakat her bir dilim, sadece bir karakteri ifade etmemektedir. Eğer her bir dilimin bir karaktere karşılık geldiği kabul edilirse yanlış sonuç üretilmiş olur. Verilen örnekte 1. ve 3. dilimler belirlenen eşik değerinin üzerinde kaldıkları için bu dilimler histogram-1 grafiği ile tekrar dilimleme işlemine tabi tutulmakta ve sözlük yardımı ile doğru sonuç üretilmektedir.

Şekil 5. Histogram-1 ve Histogram-2'nin beraber kullanılması ve sistemin ürettiği sonuç

Karakter tanıma sisteminde, öznitelik vektörünün elde edilmesinde karşılaştırılan ikili kodlardaki eşleşen bit sayısından yararlanılmaktadır. $X = \{ x_1,\ldots,x_i,\ldots,x_n \}$ ve $X' = \{ x_1',\ldots,x_i',\ldots,x_n' \}$ ikili görüntü değerleri olmak üzere (5) denklemi ile her bir görüntü için benzeme katsayısı hesaplanmaktadır.

$$\sum_{i=1}^{i=n} b_i = \begin{cases} 0, & \text{eğer } x_i \neq x_i' \\ 1, & \text{eğer } x_i = x_i' \end{cases} \quad (5)$$

Karakter tanıma işlemine tabi tutulan aday karakterin ikili kodu, veri tabanında bulunan kayıtlarla karşılaştırılmakta, benzeme değeri en yüksek olan karakterin bulunduğu sınıf aday karakterin de sınıfı olarak kabul edilmektedir. Ancak, en çok benzeme değerini veren sınıf doğru sınıf olmayabilir. Bu nedenle karakter tanıma işlemi *k-NN* sınıflama yöntemi ile kuvvetlendirilmektedir. K-NN (k-en yakın komşuluk) sınıflama yönteminde tanıma; aday öznitelik vektörüne, öznitelik vektör uzayında en yakın öznitelik vektörünün bulunmasına dayanmaktadır. K değeri, 1'den büyük ve genelde tek olarak seçilen bir tam sayıdır. Çalışmada uygun *k*'yı bulmak için 172 kişiden alınan karakterler ile testler yapılmıştır. Yapılan testler sonucunda Şekil 6'da görüldüğü gibi küçük harfler için *k* değerinin 3 olduğu durumda %90.5 ile en iyi sınıflama sonucunun üretildiği tespit edilmiştir [Şekerci ve Kandemir, 2006].

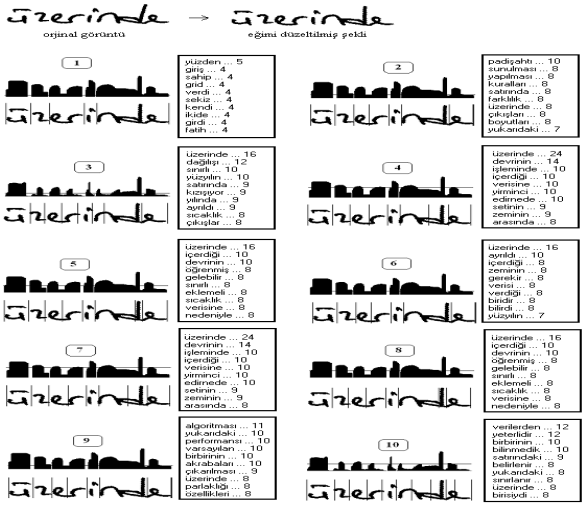


Şekil 6. *k* değerlerine göre küçük harf tanıma oranları

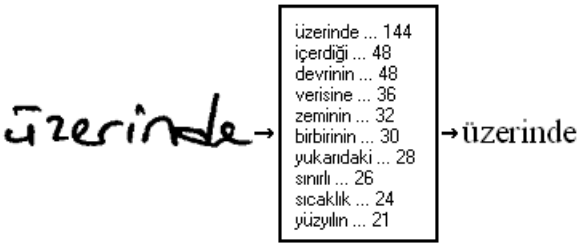
Kelime Tanınma

Tasarlanan sistemde, birleşik kelimelerin tanınması işlemi, dilimleme ve sözlük kullanımı aşamalarının beraber kullanımıyla gerçekleştirilmektedir. Karakter tanıma sisteminin ürettiği karakterlerden oluşan kelimeye en çok benzeyen sözlükteki ilk on kelime seçilerek birbirlerine benzeme değerleri saptanmaktadır. On farklı eşik değeri için yapılan dilimleme işlemi sonunda ilk ona giren aday kelimeleri içeren on adet tablo elde edilmektedir. Bu tablolar birleştirilerek en çok benzeyen kelime bulunmaktadır. Böylece kelime tanıma işlemi tamamlanmaktadır.

Kelime tanıma için yapılan işlemlerin **"üzerinde"** kelimesine uygulanarak elde edilen sonuçları Şekil 7'de gösterilmektedir. Tablo 1'de verilen 10 farklı eşik değerine göre dilimleme işlemi uygulanarak her bir dilimleme işlemi sonrasında kelime tanıma sistemine göre ilk ona giren kelimeler benzeme değerlerine göre sıralanmaktadır. Elde edilen 10 adet tablo birleştirilerek Şekil 8'de verildiği gibi yeniden bir sıralama işlemi yapılmakta ve bu sıralamaya göre ilk on kelime seçilmektedir, en çok benzeme değerine sahip olan kelime ise tanınan kelime olarak üretilmektedir.



Şekil 7. Dilimleme çeşitleri ile ilk ona giren kelimeler ve benzeme değerleri



Şekil 8. Orijinal görüntü, ilk ona giren kelimeler ve sistemin ürettiği sonuç

Sonuç ve Tartışma

Çalışmanın uygulaması 172 kişiden alınan küçük harflerle yazılmış farklı metinler üzerinde test edilmiştir. Türkçemiz eklemeli kelime yapısına sahip olduğu için günlük Türkçe'de kullanılan kelime sayısı milyonları bulmaktadır. Bu nedenle, geliştirilen tanıma sisteminde farklı türlere sahip kitaplardan rastgele seçilmiş 2500 kelime ile sınırlı bir sözcük listesi kullanılmıştır. Çalışmadaki, karakter tanıma performansı %90.5 iken kelime tanıma performansı %84 olarak elde edilmiştir. Elde edilen kelime tanıma performansının daha düşük olması çalışmada kullanılan sözlükteki kelime sayısının sınırlı olmasından kaynaklanmaktadır. Türkçe el yazısı tanıma üzerine yapılan çalışmaların sınırlı olması ve elde edilen tanıma sonuçları açısından çalışmamız umut vericidir.

Çalışmanın sonraki aşamalarında, sözlükteki kelime sayısı artırılarak son işlem aşaması geliştirilmeye çalışılacaktır. Bununla beraber sözlükteki kelime sayısının artması sistemin kullanacağı zamanı da artacaktır. İşlem süresinin azaltılması üzerindeki iyileştirici çalışmaların yapılması gerekmektedir. Zaman performansını yükseltmek için farklı yöntemler üzerine çalışmalar sürecektir.

Kaynaklar

A.Çapar, K.Taşdemir, Ö. Kılıç, M.Gökmen "A Turkish Handprint Character Recognition System" 18th International Symposium on Computer and Information Sciences (ISCIS'03), Antalya TURKEY, 2003

Erdem O. A., Uzun E. "Yapay sinir ağları ile Türkçe times new roman, arial ve el yazısı karakterleri tanıma", Gazi Üniv. Müh. Mim. Fak. Der. Cilt 20, No 1, 13-19, 2005

Gunter Simon, Bunke Horst, "Off-line cursive handwriting recognition using multiple classifier systems—on the influence of vocabulary, ensemble, and training set size", Optics and Lasers in Engineering 43 (2005) 437–454

Senior Andrew W., Robinson Anthony J., "An Off-Line Cursive Handwriting Recognition System", IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 3, march 1998

Shi Z., Govlndaraju V. "Segmentation and recognition of connected handwritten numeral strings", Pattern Recognition, Vol. 30, No. 9, pp. 1501 1504. 1997

Shridhar M., Houle G., Kimura F., **"**Handwritten Word Recognition Using Lexicon Free and Lexicon Directed Word", Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR), Pages: 861 – 865, 1997

Şekerci M., "Birleşik ve Eğik Türkçe El Yazısı Tanıma Sistemi", T.Ü.Fen Bilimleri Enst, tez, Mart 2007, Edirne

Şekerci M., Kandemir R., "Sözlük Kullanarak Türkçe El Yazısı Tanıma", Elektrik–Elektronik Bilgisayar–Mühendisliği Sempozyum (ELECO 2006), Aralık 2006, BURSA

Verma, B., Blumenstein M. & Kulkarni S.,"Recent Achievements In Off-Line Handwriting Recognition Systems," International Conference on Computational Intelligence and Multimedia Applications (ICCIMA '98), Melbourne, Australia, pp. 27-33, 1998.

Vural E., Erdoğan H., Oflazer K., Yanıkoğlu B., "Türkçe İçin Tablet PC Ortamında Çevrimiçi Yazı Tanıma Sistemi", 12.th Signal Processing and Communication Applications Conference, April 2004, Kuşadası

Yanıkoğlu B., Kholmatov A. 'Turkish handwritten text recognition: A case of Agglutinative Languages**",** Proceedings of SPIE, January 2003