Türkçe İçin Tablet PC Ortamında Çevrimiçi Yazı Tanıma Sistemi
An Online Handwriting Recognition System For Turkish

Esra Vural, Hakan Erdoğan, Kemal Oflazer, Berrin Yanıkoğlu *

Sabancı Üniversitesi
Orhanlı İstanbul 34956
esravural@su.sabanciuniv.edu
{haerdogan,oflazer,berrin}@sabanciuniv.edu

Özetçe

Bu çalışmada Tablet PC üzerinde, Türkçe için bir dinamik yazı tanıma sistemi geliştirilmiştir. Son yıllarda Tablet PC kullanımında büyük oranda artış gerçekleşse de, Türkçe'yi tanıyacak bir uygulama henüz mevcut değildir. Bu çalışmada böyle bir sistemin prototipi geliştirilmiştir: az sayıda kelimeden oluşan bir sözcük listesini Gizli Markov Modelleri kullanarak tanıyan bu prototip, kelime tanımada %90'ın üzerinde bir başarı göstermiştir.

Abstract

In this work an online handwritten text recognition system for Turkish has been developed using a Tablet PC as an interface. In recent years, although there has been great developments in the Tablet PC technology, still there are no applications for recognition in Turkish language. In this work, we have developed a prototype system using Hidden Markov Models which recognizes handwritten words from a small vocabulary list. This system has achieved a recognition rate over %90 percent.

1. Giriş

Elle veya makinayla yazılmış yazıların bilgisayar tarafından tanınması işlemine OCR adı verilmektedir. OCR uygulamaları girdinin türüne göre çevrimiçi ve çevrimdışı olmak üzere iki ayrı gruba ayrılır. Çevrimiçi uygulamalarda kalem basıncına karşı hassas bir tablet aracılığıyla alınan yazının tanınması amaçlanır. Çevrimdışı uygulamalarda ise sisteme sadece bir belgenin dijital imgesi verilir.

Bu calışmada amaç, çevrimiçi bir uygulama aracı olan Tablet PC'den alınan Türkçe yazının tanınmasını sağlamaktır. Tablet PC ortamı için çevrimiçi yazı tanıma sistemi 15 kadar Avrupa dili için mevcut olsa da, Türkçe için böyle bir sistem yoktur [5].

Kelime tanıma aşamasında genellikle sistemin performansını yükseltmek için, dokümanda karşılaşılabilecek kelimeleri içerdiği varsayılan bir sözlük kullanılır ve sistem bu kelimelerle sınırlanır. İngilizce dokümanlar için 30,000 kelimelik bir sözlük pek çok uygulama için yeterli olmaktadır. Ancak, Türkçe'nin eklemeli sözcük yapısı böyle sınırlı büyüklükte bir sözlük oluşturmaya imkan vermez. Türkçe yazı tanıma konusunda uygulamalar kadar akademik çalışmalar da azdır; Yanıkoğlu ve Kholmatov'un geniş dağarcıklı çevrimdışı yazı tanıma sistemleri [1] sözcük listesi kullanmadan, Oflazer ve grubunun geliştirdiği Türkçe önek tanıyıcısını kullanarak çalışır [6]. Bu çalışmada geliştirilen sistem ilk aşamada bir sözcük listesi kullanacak şekilde geliştirildiyse de, ilerde benzer şekilde genişletilecektir.

* Corresponding author.

2. Kullanıcı Arayüzü

Tablet PC'ye giriş, basınca duyarlı ekranın üzerine ve her örneklemede kalemin o andaki x,y koordinatları ve basınç bilgisi saklanır. Bu şekilde elde edilen herbir kelime için ortalama 300 noktada örnekleme yapılmaktadır. Geliştirdiğimiz sistemde Tablet PC'nin topladığı bu verilere ulaşmak için Tablet PC API'sini kullanan bir kullanıcı arayüzü geliştirilmiştir.

3. Sistem

Sistemin iki aşaması vardır: eğitme ve test. Eğitme aşamasında toplanılan veriler (30 farklı kişiden toplanan yaklaşık 800 kelime), yazı ve konuşma tanımada en çok kabul gören yöntem olan Gizli Markov Modelleri (GMM) 'ni eğitmek için kullanılır. Her harf veya her kelime için bir modelin eğitildiği bu aşamadan sonra, test olarak nitelendirilen normal kullanım aşamasında yazılan bir kelimenin tanınması en uygun modelin bulunmasıyla gerçekleştirilir.

3.1. Gizli Markov Modelleri

Tablet PC aracılığıyla toplanan el yazısı dataları gizli Markov (GMM) modeliyle eğitilmiştir. Gizli Markov modelleri dinamik bir değişkene bağlı olarak, durağan olmayan bir şekilde değişen öznitelikleri açıklamak için kullanıldığından bu problem için uygun bir modeldir. GMM'lerde sonlu sayıda durum (state) tanımlanır ve özniteliklerin bulundukları duruma göre sabit bir olasılık dağılımından üretildikleri varsayılır. Böylece özniteliklerdeki durağansızlık durum değiştirme yöntemiyle açıklanmış olur.

GMM'inde kullanılan bazı parametreler ve denklemler aşağıda açıklanmıştır.

N: Durum sayısı
$A = [a_{ij}]_{N \times N}$: Durum Geçiş Matrisi

0-7803-8318-4104/$20.00©2004 IEEE

$b_j(o_t) = P(o_t, s_t = j)$:$j$ Durumundaki Gözlem olasılıkları.
o_t: t zamanındaki Gözlem
s_t: t zamanındaki Durum
$\pi_i = P(s_1 = i)$ ise i durumunda başlama olasılığı

Gizli Markov Modellerindeki üç ana probleme lineer zamanlı etkin çözümler bulunmuştur[3]. Bunlardan bazıları şöyledir:

Belirli $O = [o_1 o_2 o_3 ....o_n]$ gözlem diziliminin belirli bir durum dizilimi $q = [s_1 s_2 s_3 ....s_n]$ tarafından üretilme olasılığı

$$P(O, q|\lambda) = P(O|q, \lambda)P(q|\lambda) = \prod_{i=1}^{T} b_{s_i}(o_i) \prod_{i=1}^{T-1} a_{s_i s_{i+1}} \quad (1)$$

şeklinde ifade edilir. Burada λ model parametresini göstermektedir.

Bütün olası durum dizileri için O gözlem diziliminin elde edilme olasılığı ise denklem 2'de gösterildiği gibi hesaplanır. En olası model, sistemin kullanılışı sırasında en olası kelime olarak seçilir.

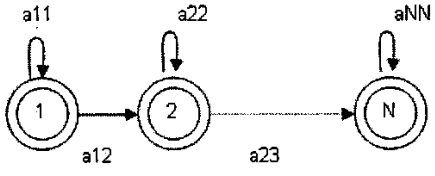
$$P(O|\lambda) = \sum_{\forall q} P(O|q, \lambda)P(q|\lambda) \quad (2)$$

Verilen bir gözlem dizilimi için en olası durum dizilimi ise aşağıdaki denklemle bulunur.

$$q_{max} = \text{argmax}_q P(O|q, \lambda)P(q|\lambda) \quad (3)$$

Bu çalışmada GMM modeli için Şekil 1'de görüldüğü gibi soldan sağa bir topoloji kullanılmıştır. Bu topoloji aynı zamanda ses tanımada kullanılan en basit ve başarılı topolojilerdendir. Bu modelde durum geçiş olasılıkları $a_{ij} = 0, j < i$, başlangıç durum olasılıkları ise $\pi_i = 0, i > 1$ şeklinde sınırlanır.

Gizli Markov Modelini gerçekleştirmek için HTK yazılımı kullanılmıştır [3]. Model olarak kelime bazlı ve harf bazlı eğitim modelleri denemiştir. Harf modeli kullanıldığında bütün harfler için, kelime modeli kullanıldığında bütün kelimeler için sabit sayıda durum kullanılmıştır. Gözlem olasılıkları için ise Gauss dağılımı kullanılmıştır. Harf modeli kullanıldığı zaman, bir kelimenin butun harflerine karşılık gelen harf modelleri peşpeşe dizilerek kelime modeli yaratılmış olur.



Şekil 1: Soldan sağa yapıya sahip olan GMM gösterimi

3.2. Öznitelikler

Bu çalışmada beş ana öznitelik hesaplanır: x ve y koordinatlarının birincil ve ikincil türevleri ve yüzde olarak basıncın değişimidir; x ve y koordinatlarının gerçek değerleri, imza yerinin kaymasından dolayı uyumsuzluk meydana gelebileceğinden kullanılmamıştır.

Koordinatların birinci ve ikinci dereceden türevleri sırasıyla (4) ve (5) no'lu denklemler kullanılarak hesaplanmıştır:

$$dx_t = \frac{\sum_{\theta=1}^{\Theta} \theta(x_{t+\theta} - x_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \quad (4)$$

$$ddx_t = \frac{(dx_{t+1} - dx_{t-1})}{2\Theta} \quad (5)$$

Burada x_t, x-koordinatının t anındaki değeri, Θ, gözlem penceresi genişliğidir. Pencere genişliği bu çalışma için 5 olarak seçilmiştir. y-koordinatının türevi ise benzer şekilde bulunur.

Basıncın değişim yüzdesi ise (6) nolu denklemle hesaplanmıştır:

$$dp_t = \frac{(p_{t+1} - p_{t-1})}{2p_t} \quad (6)$$

Burada da p_t, basıncın t anındaki değeridir.

3.3. Veritabanı

Sistem iki aşamada geliştirilmiş ve test edilmiştir. İlk aşamada elli farklı kelimeden oluşan bir sözcük dağarcığı belirlenmiş ve toplam 20 kişiden bu sözcük dağarcığındaki kelimelere karşılık gelen el yazısı verisi toplanmıştır. Bu şekilde elde edilen toplam veritabanı büyüklüğü 1000 kelimeden oluşmaktadır (50 kelime x 20 kişi) . Veri toplanırken önceden geliştirdiğimiz arayüz kullanılmıştır. Bu arayüzde, Tablet PC yazı uygulamalarında olduğu gibi, kullanıcıların düz bir çizgi üzerine kelimeleri yazması istenmiştir. Fakat kelimelerin yazımına dair başka hiçbir kısıtlama getirilmemiştir.

İkinci veritabanı için elektronik posta mesajlarından derlenen 1000 farklı kelimeden oluşan bir sözcük dağarcığı belirlenmiştir. Bu 1000 kelime, yüzer kelimeden oluşan on ayrı sözcük kümesine ayrılmış ve 30 farklı kişi kümelerin birindeki kelimeleri veri olarak yazmıştır. Sonuç olarak, her bir sözcük kümesi için üç farklı kişiden el yazısı örneği toplanmıştır. Toplam veritabanı büyüklüğü 3000 kelimedir (10 küme x 100 kelime x 3 kişi).

Her iki veritabanında da sözcük dağarcığının eşit dağılımlı olarak Türkçe karakterleri içermesine ve sık kullanılan kelimelerden seçilmesine dikkat edilmiştir.

4. Deneyler ve Çıkarımlar

İlk deneyde az kelime grubunu çok kişi ile denemek amaçlanmıştır. Bu deney için ilk veritabanı kullanılmıştır. Toplanan verinin 15 kişilik kısmı (750 kelime) eğitim, geriye kalan 5 kişilik kısmı (250 kelime) ise test için ayrılmıştır. Gizli Markov Modelini eğitmekte kullanılan kelime dağarcığı, testte kullanılan dağarcıkla aynı seçilmiştir.

Tablo 1'de bu veri setiyle denenen harf ve kelime modellerinin sonuçları gösterilmektedir. Harf modeli %97, kelime modeliyse %95 performans elde etmiştir. Genelde kelime modelinin daha iyi sonuç vermesi beklense de, kanımızca kelime

modeli için durum sayısı daha fazla artırılmadığı için bu şekilde bir sonuç elde edilmiştir. Sınırlı sözcük dağarcıklarında harf veya hece modeli ve kelime modeli arasında çok büyük bir farklılık olmamasına rağmen, eğitilmesi gereken kelime modeli sayısı sözcük sayısı kadar olduğundan, sözcük sayısı arttıkça uygun olmamaktadır.

Model	Yazar Sayısı	Kelime Sayısı	Başarı
harf	20	50	%97
kelime	20	50	%95

Tablo 1: Birinci Veritabanı üzerinde Harf ve Kelime modelleriyle Başarı Oranı

İkinci deneyde durum sayılarının değişimi ile başarı oranı arasındaki ilişki gözlenmiştir. Tablo 2'de görüldüğü gibi, harf modeline en uygun durum sayısı 20 olarak bulunmuştur. Tablo 3'te ise kelime modelinin durum sayısıyla olan ilişkisi gösterilmektedir. Buna göre durum sayısı kelime modeli için artmaktadır, ve ortalama 70 durumla kelime modeli ifade edilebilir.

Model	Durum Sayısı	Kelime Sayısı	Başarı
harf	10	50	%92
harf	20	50	%97
harf	30	50	%94

Tablo 2: Birinci Veritabanı İçin Durum Sayılarının Değişiminin Harf Modelindeki Etkisi

Model	Durum Sayısı	Kelime Sayısı	Başarı
kelime	30	50	%89
kelime	50	50	%94
kelime	70	50	%96

Tablo 3: Birinci Veritabanı için Durum Sayılarının Değişiminin Kelime Modelindeki Etkisi

Prototip geliştirme aşamasında yapılan yukardaki deneylerden sonra 1000-kelimelik veri tabanında daha gerçekçi sonuçlar elde edilmiştir. Bin kelimeden oluşan 2. veri tabanı, veri toplanması amacıyla , yüzer kelimeden oluşan on ayrı sözcük kümesine ayrılmıştı ve 30 farklı kişi kümelerin birindeki kelimeleri veri olarak yazılmıştı.

Bu veri tabanı ile yapılan ilk deneyde toplam verinin 2 kümesi test (toplam 200 kelime), 8 kümesi (toplam 800 kelime) eğitim için ayrılmıştır. Bu şekilde, Gizli markov Modelini eğitmekte kullanılan sözcükler ve yazarları ile testte kullanılan sözcükler ve yazarları birbirinden tamamen farklı seçilmiş olmaktadır. Eğitim ve test kelimelerinin örtüşmemesi kelime modeline uygun bir yapı olmadığı için harf modeli denenmiştir.

Bu deneyde eğitim ve test setlerinin seçiminin olabildiğince rastgele yapılması için 5 farklı dağılım ile deney tamamlanmıştır (örn. 1 ve 2 nolu kümeler teste, gerisi eğitime ayrılmıştır). Tablo 4'te bu farklı dağılımlar için alınan sonuçlar sunulmaktadır, bu sonuçların benzerliği bu deneyin sonuçlarının belli eğitim kümelerine çok bağlı olmadığını, gürbüz olduğunu göstermektedir. Dolayısıyla, harf modelleri ile 1000-kelimelik veri tabanının yazar-bağımsız olarak sınanmasındaki başarı oranı %91.1'dir. Yazar-bağımsız, yani eğitim ve test setindeki yazarların tümüyle farklı olmaları durumu, en zor durumdur.

Test Kümesi	Başarı
1,2	%92.6
3,4	%92.5
5,6	%89.6
7,8	%89.8
9,10	%91.2
Ortalama:	%91.1

Tablo 4: İkinci Veritabanı için, Harf Modelleri ile, Farklı Test Kümelerindeki Başarı Oranları.

Ikinci deneyde test ve eğitim amacıyla kullanılan kelimelerin ortak, yazarlar farklı seçilmiştir. Eğitim sırasında tanınacak test aşamasında tanınacak kelimelerin örneklerinin görülmüş olması başarıyı artırır. Bu amaçla 1000 kelimelik veri kümesi 10 grupx100 kelimex3kişi şeklinde olduğu için, her gruptan iki kişi eğitim, bir kişi test için ayrılmıştır. Böylece her gruptaki bütün kelimeler hem eğitim hem test kümesine eklenir.

Bu deneydeki başarı oranı Tablo 5'de gösterildiği gibi %90.4 olarak bulunmuştur. Bu başarı yukarda bahsedilen deneyden (Tablo 4) daha düşük bulunmasına rağmen, aradaki ufak fark eğitim verilerinin toplam veriye oranındaki düşüşle (%80'den %66'ya) açıklanabilir.

Model	Yazar Sayısı	Kelime Sayısı	Başarı
harf	30	1000	%90.4

Tablo 5: İkinci Veritabanı İçin Eğitim Kelimelerinin Test Kelimeleriyle Ortak Seçilmesindeki Başarı Oranları

Son deneyde farklı öznitelik seçimlerindeki başarı oranları incelenmiştir. Çeşitli öznitelik kombinasyonları denenmiştir. Öznitelikler sırasıyla x koordinatının birinci türevi (dx), y koordinatının birinci türevi (dy), x koordinatının ikinci türevi (ddx), y koordinatının ikinci türevi (ddy), ve basıncın yüzde değişimi (basınç) olarak gösterilmiştir. Tablo 6'da görüldüğü gibi en iyi öznitelik kombinasyonu tüm niteliklerin seçilmesiyle oluşur; ancak aradaki fark çok küçüktür.

Öznitelikler	Yazar Sayısı	Kelime Sayısı	Başarı
dx,dy	30	1000	%89.3
dx,dy,ddx,ddy	30	1000	%88.7
dx,dy,ddx,ddy,basınç	30	1000	%90.4

Tablo 6: İkinci Veritabanı İçin Farklı Özniteliklerin Seçilmesindeki Başarı Oranları

5. Sonuçlar ve Sonsöz

Bu deneylerde oldukça başarılı sonuçlar elde edilmiştir. En belirgin hatalar kelimelerin olduklarından daha kısa tahmin

Yazılan Kelime	Tanınan Kelime
bilgisayarımdan	bilgisayarıma
odalarının	odalarda
isteyenler	istiyorum
yaptım	yapım

Tablo 7: Örnek hatalar, sistemin benzer kelimeleri karıştırdığını göstermektedir.

edilmesidir. Her harf farklı sayıda durumla gösterilirse bu sorun ortadan kalkabilir. Bir başka çeşit hata grubu ise yazımda *durum* olan kelimeyi *olurum* olarak tanımadır. Bu iki kelime birbirinden ayırd edilemez şekilde yazılabildiklerinden dolayı bu ayrımın yapılabilmesi anlam çözümleme yapılmadan mümkün değildir.

Ayrıca Türkçe'deki ç,ş,ğ,ü gibi harflerde bulunabilecek geçikmiş vuruşlardan kaynaklanabilecek hatalar mevcuttur. Bu sorun genelde çevrimiçi yazı tanımada noktalı harflerde de görülmektedir, fakat Türkçe için böyle harflerin çokluğu yüzünden sorun daha büyümektedir. İleriki safhalarda geciken vuruşların ayrı bir sembolle gösterilip modellenmesi planlanmaktadır. Kelime modelinde ise gecikmiş vuruşlu bir kelimenin tüm olası yazımları değerlendirilebilir ve buna göre seçim yapılabilir.

Son yıllarda gelişmekte olan Tablet PC ürünlerine uygun Türkçe tanıma modülü henüz geliştirilmemiş bir uygulama alanıdır [4]. Şu an için 1000 kelimeden oluşan bu prototip sistemin, ileride dil modelleme modülleriyle entegre edilerek geniş dağarcıklı bir ürün haline getirilmesi planlanmaktadır.

Konuşma tanıma sistemlerindekine benzer bir şekilde, sistemin başarısı kelime sayısı ile genelde ters orantılı olduğundan kelime sayısı arttıkça başarının düşmesi beklenir ve normaldir. Ancak 1000-kelimelik veri kümesi e-posta yazışmalarından derlendiği için aynı kök kelimenin pekçok halini (örn. bize, bizler vb.) zaten içermektedir. Bu yüzden kelime sayısı arttığı zaman performanstaki düşmenin çok belirgin olması beklenmemektedir.

İlerideki çalışmalarda sistem birden fazla Gauss karışım modelleriyle ve daha az durum kullanılarak test edilecektir.

6. Kaynakça

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

[2] Yanıkoğlu B., Sandon P. A., Segmentation of off-line cursive handwriting using linear programming. Pattern Recognition, 31(12):1825-1833, 1998. 6

[3] Young S., et al. The HTK Book v3.0. Cambridge University, 1999

[4] http://msdn.microsoft.com/msdnmag/issues/03/10/TabletPC/default.aspx

[5] http://www.microsoft.com/windowsxp/tabletpc/ multilanguagecd.asp

[6] Oflazer K., Two-level description of Turkish morphology, Literary and Linguistic Computing 9(2), 1994.

[7] Hu J., Lim S. G., Brown M. K., "Writer independent online handwriting recognition using an HMM approach" Pattern Recognition, January 2000.