Sözlük Kullanarak Türkçe El yazısı Tanıma

Murat Şekerci[1] **Rembiye Kandemir**[2]

[1,2] Bilgisayar Mühendisliği Bölümü
Mühendislik-Mimarlık Fakültesi
Trakya Üniversitesi, 2250, Edirne

[1]e-posta: muratsekerci@hotmail.com [2] e-posta: rembiyeg@trakya.edu.tr

ÖZET

Bu çalışmada, Türkçe el yazısı için bir tanıma sistemi geliştirilmiştir. Tanıma sistemi büyük harf ve küçük harfleri ayrı ayrı tanıyacak şekilde tasarlanmıştır. 172 kişiden, 29 adet büyük harf ve küçük harf olmak üzere toplam 9976 karakter örneği alınarak veri tabanı oluşturulmuştur.
172 kişilik örnekler topluluğundan 100 kişilik örnek eğitim amaçlı 72 kişiden alınan karakter örnekleri ise karakter tanıma algoritmasının performansını test etmek için kullanılmıştır. Karakter tanıma aşamasında, K-en yakın komşuluk sınıflama yöntemi kullanılmıştır. Doğrulama aşamasında sınırlı sayıda kelime içeren bir sözlük kullanılarak anlamsız harf gruplarının seçilmesi engellenenmiş ve yanlış tanınan kelimelerin düzeltilmesi sağlanmıştır. Geliştirilen sistem, el yazısı metin örnekleri ile test edilmiştir.

Anahtar sözcükler*: El yazısı tanıma, K-en yakın komşuluk, sözlük*

1. GİRİŞ

El yazısı tanıma konusunda çok sayıda çalışma yapılmıştır[1]. Türkçe el yazısı tanıma konusunda ise yapılan çalışma sayısı sınırlıdır. Bunlardan bazıları; Berrin Yanıkoğlu ve arkadaşlarının önerdiği Türkçe el yazısı tanıma sistemi[2], Ayhan Erdem ve arkadaşlarının önerdiği yapay sinir ağları ile el yazısı karakter tanıma sistemi[3], Esra Vural ve arkadaşlarının önerdiği çevrimiçi yazı tanıma sistemleri verilebilir[4]. Makine baskısı Türkçe yazı tanıma alanında yapılmış çalışmalar da mevcuttur [5-6].

Bu çalışmada, Türkçe el yazısı için bir tanıma sistemi geliştirilmiştir. Bu sistem, yazı görüntüsü üzerinde işlem yapmakta ve görüntü üzerindeki gürültülerin temizlenmesi, satırların, kelimelerin, karakterlerin ayrıştırılması, karakterlerin tanınması ve son işleme aşamalarından oluşmaktadır. Karakter tanıma aşamasında korelasyon yöntemi kullanılarak, K-en yakın komşuluk sınıflama yöntemi ile tanıma oranı arttırılmıştır. Son işleme aşamasında sınırlı sayıda kelime içeren bir sözlük kullanılarak ve anlamsız harf gruplarının seçilmesi engellenerek yanlış tanınan kelimeler düzeltilmiştir.

2. ÖN İŞLEM

2.1.Karakterlerin standart hale getirilmesi

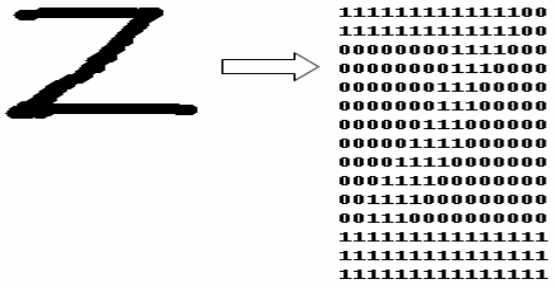
Veri toplama işlemi iki aşamada gerçekleştirilmiştir.

1. Aşama: 172 kişiden büyük ve küçük harf olmak üzere toplam 9976 harf toplanılmış ve bu harflerden 5800 adedi eğitim amaçlı, 4176 adedi test amaçlı kullanılmıştır.
2. Aşama: Yazı tanıma işleminden kullanılmak üzere farklı kişilerden el yazısı metin örnekleri toplanmıştır.

Her iki veri toplama işlemindeki örnekler görüntü işleme yöntemleri kullanılarak gri seviye tonlamasına dönüştürülerek görüntü üzerindeki gürültüler temizlenmiştir.

Farklı kişilerden alınan örnek harfler aynı standartta yazılmadığından bu karakterleri standart hale dönüştürmek için bazı işlemler uygulanmıştır.

İkili koda dönüştürme işleminde karakter görüntüsü, birbirine eşit 15 yatay ve 15 dikey dilime bölünerek 225 alan meydana getirilmiştir. Her bir alan sırasıyla taranarak içinde siyah piksel bulunduran alan 1, hiç siyah piksel bulundurmayan alan ise 0 ile ifade edilmiştir. Böylece, Şekil 1 de görüldüğü gibi bütün el yazısı karakter görüntüleri 0 ve 1'lerden oluşan 225 karakter uzunluğunda standart bir şekle dönüştürülmüştür.



Şekil 1. Karakter görüntüsünün ikili koda dönüştürülmesi

Elde edilen bu karakter görüntüleri veri tabanında kullanılmıştır.

2.2. Veri tabanının oluşturulması

Tasarlanmış olan karakter tanıma sisteminde kullanılan veritabanı, text dosyası tabanlıdır. Küçük harf ve büyük harf olmak üzere iki ayrı text dosyası mevcuttur. Bu dosyalarda bir karakter iki satırla ifade edilmektedir; birinci satıra karakterin ascii kodu, ikinci satıra karakterin ikili kod karşılığı kullanılmıştır.
Örneğin, "a" ve "b" için text dosyasındaki kayıtları
97
01011101110101000101000101101 0
98
01001001001001010110010110101 0
şeklindedir.

3. KULLANILAN YÖNTEMLER

3.1. Benzeme katsayısının hesaplanması

Yapılan çalışmada, karşılaştırılan ikili kodlardaki eşleşen bit sayısından yaralanılmıştır.

$X = \{ x_1,\ldots,x_i,\ldots,x_n \}$ ve $X' = \{ x_1',\ldots,x_i',\ldots,x_n' \}$ ikili görüntü değerleri olmak üzere (1) denklemi ile her bir görüntü için benzeme katsayısı hesaplanmaktadır.

$$\sum_{i=1}^{i=n} b_i = \left\{ \begin{array}{ll} 0, & eğer\ x_i \neq x_i' \\ 1, & eğer\ x_i = x_i' \end{array} \right\} \quad (1)$$

Örneğin, ikinci bölümde anlatılan "a" ve "b" karakter görüntülerinin ikili kodları karşılaştırılırsa, eşleşen bitlerde benzeme değeri 1 arttığından dolayı, "a" ve "b" nin aralarındaki benzeme değeri 18 olarak hesaplanır.

Karakter tanıma işlemine tabi tutulan aday karakterin ikili kodu, veri tabanında bulunan bütün kayıtlarla karşılaştırılır ve en çok benzeme değerini aldığı karakter belirlenir ve bu sınıfa dahil olduğu kabul edilir.

Bu şekilde yapılan karakter tanıma işlemlerinde yanlış sonuçlar elde edilebilir. En çok benzeme değerini veren sınıf doğru sınıf olmayabilir. Bu nedenle karakter tanıma işlemi K-en yakın komşuluk sınıflama yöntemi ile kuvvetlendirilmiştir.

3.2. K-en yakın komşuluk

K-en yakın komşuluk, sınıflama yöntemlerinden birisidir. Buradaki K, 1'den büyük ve genelde tek sayi olarak seçilen bir tam sayıdır. K-en yakın komşuluk yönteminde, sadece bir tane en büyük benzeme değerine değil, K tane en büyük benzeme değerine bakılarak sonuca ulaşılır.

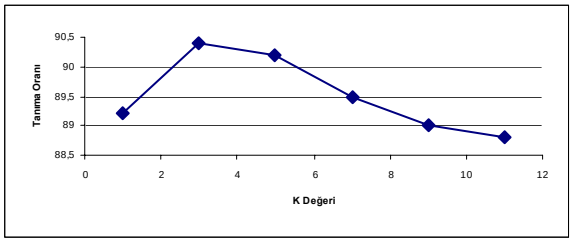
Sıra	Veritabanındaki Örnek	Benzeme Değeri
1	e	198
2	c	189
3	c	187
4	c	185
5	e	169

Tablo 1. Örnek tablo

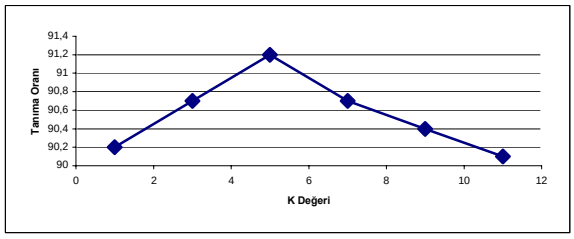
Tablo 1. de, test örneği olan "c" nin, veritabanında bulunan "e" lerden birisine çok benzediği, ama bununla birlikte veritabanında bulunan üç adet "c" ye de oldukça benzediği söylenebilir. Eğer böyle bir durumda sadece benzeme katsayısı kullanılırsa tanıma algoritması, 'c' test örneğini "e" olarak kabul eder.

K-en yakın komşuluk yöntemi kullanıldığında ise test verisinin "c" olduğu kabul edilir. Yani K-en yakın komşuluk yönteminde K adet en benzer sonuç içinde, en çok sayıda eleman bulunduran sınıf seçilmektedir [8].

Bu çalışmada, en uygun K'yı bulmak için 72 kişiden alınan veriler ile testler yapılmıştır. Yapılan testler sonucunda küçük harfler için K değerinin 3, büyük harfler için ise K değerinin 5 olduğu durumlarda en iyi sonuçların elde edildi gözlenmiştir. Test sonuçları grafik olarak Şekil 2' de verilmektedir.



a) Küçük Harfler



b) Büyük Harfler

Şekil 2. K değerlerine göre tanıma oranları

En uygun K değerleri ile yapılan sınıflandırma işlemlerindeki tanıma oranları ayrıntılı bir şekilde Tablo 2 ve Tablo 3'te gösterilmektedir.

Karakter	Adet	Doğru	Yanlış	Tanıma Oranı
a	46	37	9	%80,4
b	66	62	4	%93,9
c	62	62	0	%100
ç	34	32	2	%94,1
d	63	62	1	%98,4
e	59	51	8	%86,4
f	51	45	6	%88,2
g	59	47	12	%79,6
ğ	54	39	15	%72,2
h	53	52	1	%98,1
ı	60	53	7	%88,3
i	55	47	8	%85,4
j	36	35	1	%97,2
k	28	24	4	%85,7
l	40	36	4	%90
m	61	59	2	%96,7
n	65	65	0	%100
o	69	68	1	%98,5
ö	60	52	8	%86,6
p	62	61	1	%98,3
r	65	63	2	%96,9
s	56	49	7	%87,5
ş	41	37	4	%90,2
t	53	52	1	%98,1
u	59	57	2	%96,6
ü	57	39	18	%68,4
v	63	63	0	%100
y	44	36	8	%81,8
z	54	39	15	%72,2
Toplam	1575	1424	151	%90,4

Tablo 2. Küçük harflerdeki tanıma oranları. K=3

Karakter	Adet	Doğru	Yanlış	Tanıma Oranı
A	57	52	5	%91,2
B	57	46	11	%80,7
C	62	62	0	%100
Ç	39	39	0	%100
D	33	28	5	%84,8
E	60	54	6	%90
F	56	48	8	%85,7
G	60	54	6	%90
Ğ	60	53	7	%88,3
H	67	61	6	%91,0
I	66	65	1	%98,4
İ	49	35	14	%71,4
J	44	42	2	%95,4
K	46	40	6	%86,9
L	57	57	0	%100
M	58	55	3	%94,8
N	44	39	5	%88,6
O	68	65	3	%95,5
Ö	58	52	6	%89,6
P	57	55	2	%96,4
R	43	41	2	%95,3
S	58	55	3	%94,8
Ş	46	35	11	%76,0
T	57	57	0	%100
U	57	56	1	%98,2
Ü	53	45	8	%84,9
V	68	68	0	%100
Y	60	50	10	%83,3
Z	63	54	9	%85,7
Toplam	1603	1463	140	%91,2

Tablo 3. Büyük harflerdeki tanıma oranları. K=5

5. SATIRLARIN VE KELİMELERİN AYRIŞTIRILMASI

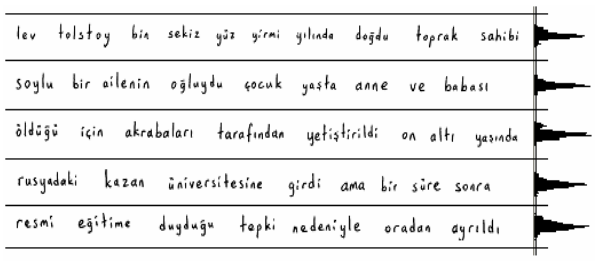
İkili seviyeye indirgenen ve gürültüleri temizlenen görüntünün yatay doğrultudaki histogram grafiği çıkartılır. Dikey doğrultuda bir eşik seviyesi belirlenir (3). Bu eşik seviyesinin altında kalan vadilerin orta noktasından yatay doğrultuda kesim yapılır. Bu kesimler satırları birbirinden ayıran kesimlerdir. Bazen, histogramda oluşan çok küçük vadilerden dolayı, fazla sayıda satır hesaplanabilir veya kelimeler kesime uğrayabilir. Bu problemi çözmek için Min Boşluk Değeri belirlenir (4). Bu değer, minimum vadi genişliğidir. Bu değerin altındaki boşluklar işleme alınmamaktadır.

Satır ayrıştırma işlemi için;
Eşik Değeri = Görüntü Genişliği / 900 (3)
Min Boşluk Değeri = Görüntü Yüksekliği/1000 (4)

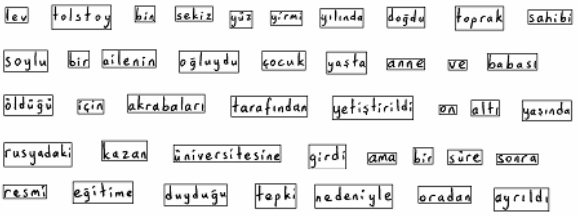
Şekil 3'de bir metindeki satırların ayrıştırılması sonucu gösterilmektedir. Satırlar birbirinden ayrıştırıldıktan sonra her bir satır üzerinde kelime ayrıştırma işlemi yapılır (5). İşlem sonunda görüntüdeki bütün kelimeler saptanır.

Kelime ayrıştırma işlemi için;
Eşik Değeri =Görüntü Genişliği / 90 (5)

Kelimeler belirlendikten sonra, her bir kelime çerçevelenir. Şekil 4'de belirlenmiş kelimeler gösterilmektedir.



Şekil 3. Satırların bulunması



Şekil 4. Kelimelerin çerçevelenmesi

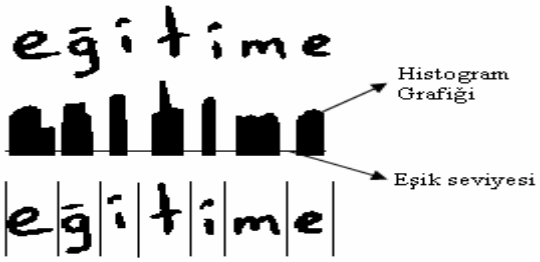
6. KARAKTERLERİN AYRIŞTIRILMASI

Kelime görüntüsünün dikey doğrultudaki histogram grafiği çıkartılır [9]. Yatay doğrultuda bir eşik seviyesi belirlenir. Bu aşamada eşik seviyesi 5 seçilmiştir. Bu seviyenin altında kalan vadilerin orta noktalarından dikey doğrultuda kesim yapılır. Bu kesimler karakterleri birbirinden ayıran kesimlerdir.

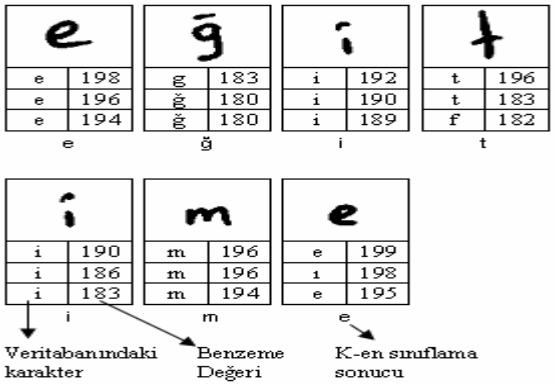
7. KELİMENİN TANINMASI VE SON İŞLEME

Karakterler ayrıştırılma işleminden sonra, her bir karakter çerçevelenir ve tanıma algoritmasına sunulur. Sonuçlar alınır ve kelime sonucu üretilir. Üretilen kelime, son işleme aşamasında, sözlük kullanımı ile doğrulanır. Kelimede yanlış tanınan karakterler varsa bunlar sözlük yardımı ile düzeltilir. Örneğin Şekil 7'da gösterilen kelime görüntüsü, "eğitima" olarak tanınırsa sözlük kullanımı ile sistem tarafından "eğitime" olarak düzeltilmektedir.

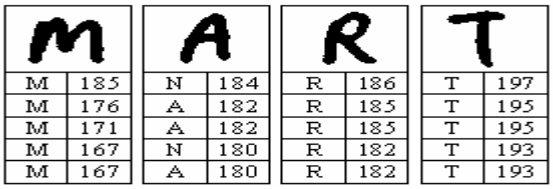
Şekil 8'de görüldüğü gibi kelime görüntüsünden ayrıştırılan her bir karakter, tanıma işlemine tabi tutulur. Küçük harfler için K değeri 3 seçildiğinden dolayı, veritabanında, bu karaktere en çok benzeyen ilk üç karakter listelenir. Büyük karakterlerde ise K değeri 5 seçildiğinden dolayı, veritabanında, bu karaktere en çok benzeyen ilk beş karakter listelenir. Listeler oluşturulduktan sonra listedeki en çok sayıya sahip sınıf seçilerek karakter sonucu üretilir.



Şekil 7. Karakterlerin ayrıştırılması



a)



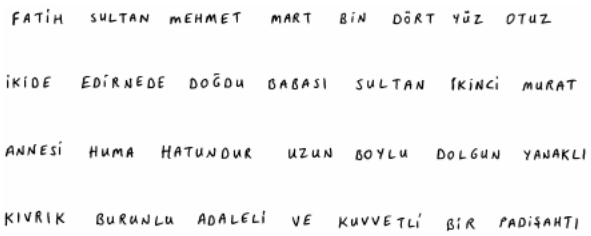
b)

Şekil 8. Karakterlerin K-en yöntemi ile tanınması.
a) küçük harfler, b) büyük harfler

Şekil 9, Şekil 10 ve Şekil 11 de sisteme verilen el yazısı metinler ve sistem tarafından üretilen sonuçlar gösterilmektedir.

lev tolstoy bin sekiz yüz yirmi yılında doğdu toprak sahibi soylu bir ailenin oğluydu çocuk yaşta anne ve babası öldüğü için akrabaları tarafından yetiştirildi on altı yaşında rusyadaki kazan üniversitesine girdi ama bir süre sonra resmi eğitime duyduğu tepki nedeniyle oradan ayrıldı

Şekil 9. Küçük harf yazılan metin için tanıma sisteminin ürettiği sonuç



Şekil 10. Büyük harf yazılan metin görüntüsü

FATİH SULTAN MEHMET MART BİN DÖRT YÜZ OTUZ
İKİDE EDİRNEDE DOĞDU BABASI SULTAN İKİNCİ MURAT
ANNESİ HUMA HATUNDUR UZUN BOYLU DOLGUN YANAKLI
KIVRIK BURUNLU ADALELİ VE KUVVETLİ BİR PADİŞAHTI

Şekil 11. Büyük harf yazılan metin için tanıma sisteminin ürettiği sonuç

8. SONUÇ

Bu çalışmada, önerilen el yazısı tanıma sistemine küçük harflerle yazılmış 152 kelime ve büyük harflerle yazılmış 62 kelime içeren metinler ayrı ayrı sunularak test edilmiştir. Tanıma sistemi bütün kelimeleri doğru tanıyarak %100 başarı sağlamıştır. Önerilen sistem ayrık yazılan kelimeler üzerinde işlem yaptığı için, bağlı veya birleşik yazılan kelimeler bu çalışmanın dışında tutulmuştur. Kullanılan sözlükteki kelime sayısı sınırlı tutulmuştur. Bu sözlükteki kelime sayısı artırılarak son işleme aşaması geliştirilebilir. Bununla beraber sözlükteki kelime sayısı arttıkça sistemin kullanacağı zaman da artacaktır. Önerilen sistem, ayrık yazılan Türkçe el yazılarının dijital ortamlara aktarılması gerektiği durumlarda kullanılabilir.

10. KAYNAKLAR

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

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

[3] O. Ayhan ERDEM ve Emre UZUN "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

[4] Esra Vural, Hakan Erdoğan, Kemal Oflazer Berrin Yanıkoğlu, "Türkçe İçin Tablet PC Ortamında Çevrimiçi Yazı Tanıma Sistemi", IEEE, 2004

[5] M. Oğuzhan Külekçi, "Turkish machine print character recognition by artificial neural networks" , 2002

[6] Sait Ulaş Korkmaz, G. Kırçiçeği, Y. Akıncı, Volkan Atalay, "A Character Recognizer for Turkish Language ", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)

[7] Nabiyev, V. V., Yapay Zeka, Seçkin Yayınevi, Ankara, 2003

[8] P. J. Grother, G.T. Candela ve J.L. Blue, "Fast Implementations of Nearest Neighbor Classifiers", Pattern Recognition, 30 (1997), 459--465.

[9] Zhixin Shi ve Venu Gov1ndaraju "Segmentation and recognition of connected handwritten numeral strings", Pattern Recognition, Vol. 30, No. 9, pp. 1501 1504. 1997