

**İNSAN HESAPLAMASI YAKLAŞIMI İLE TÜRKÇE
ARŞİVLERİN SAYISALLAŞTIRILMASI**

İBRAHİM GÜMÜŞ

**YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ**

**TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

HAZİRAN 2012

ANKARA

Fen Bilimleri Enstitü onayı

Prof. Dr. Ünver KAYNAK

Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

Doç. Dr. Erdoğan Dođdu

Anabilim Dalı Başkanı

İbrahim GÜMÜŞ tarafından hazırlanan İNSAN HESAPLAMA VE TÜRKÇE ARŞİV SAYISALLAŞTIRMA adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. Osman ABUL

Tez Danışmanı

Tez Jüri Üyeleri

Başkan :Prof. Dr. Tahir KHANİYEV

Üye : Yrd. Doç. Dr. Esra KADIOĞLU URTİŞ

Üye : Yrd. Doç. Dr. Osman ABUL

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

.....
İbrahim GÜMÜŞ

Üniversitesi : TOBB Ekonomi ve Teknoloji Üniversitesi
Enstitüsü : Fen Bilimleri
Anabilim Dalı : Bilgisayar Mühendisliği
Tez Danışmanı : Yrd. Doç. Dr. Osman ABUL
Tez Türü ve Tarihi : Yüksek Lisans – Haziran 2012

İbrahim GÜMÜŞ

İNSAN HESAPLAMASI YAKLAŞIMI İLE TÜRKÇE ARŞİVLERİN SAYISALLAŞTIRILMASI

ÖZET

İnsanların mevcutta bilgisayarlardan daha iyi yaptığı hesaplama sorunlarının çözümü için insan hesaplaması yöntemi uygun bir yaklaşımdır. Bu konuda yaşanan sorun ve zorlukları azaltmaya yardımcı olabilmek için insanın beyin gücünü, yüksek algı kapasitesini kullanan yeni yaklaşımlar geliştirilmekte ve denenmektedir. Arşiv sayısallaştırma da bu hesaplama sorunlarından biridir. Tez çalışmasında reCAPTCHA' ya benzeyen arşiv sayısallaştırma sistemi olan trCAPTCHA tanıtılmaktadır. trCAPTCHA, esasen eski Türk arşivlerinin sayısallaştırılmasını hedeflemektedir. trCAPTCHA' yı reCAPTCHA'dan ayıran en önemli özellik reCAPTCHA'nın kullandığı genel sözlüğün trCAPTCHA tarafından kullanılmamasıdır. Bunun yerine taranan her sözcükle, optik karakter okuyucu sonuçlarından üretilen her alternatif metinle trCAPTCHA kendi yerel sözlüğünü oluşturmaktadır.

Anahtar Kelimeler: İnsan hesaplama, reCAPTCHA, Arşiv Sayısallaştırma, trCAPTCHA

University : TOBB University Of Economics and Technology
Institute : Institute of Natural and Applied Sciences
Science Programme : Computer Engineering
Supervisor : Asst. Prof. Dr. Osman ABUL
Degree Awarded and Date : M.Sc. – June 2012

İbrahim GÜMÜŞ

**DIGITIZING TURKISH ARCHIVES USING HUMAN COMPUTATION
APPROACH**

ABSTRACT

Human Computation is an appropriate approach for solving computational tasks at which people are far better than computers. To ease the difficulties for solving such tasks new techniques and methods have been developed and tested. Archive digitization is such a task. In this thesis, we present our reCAPTCHA like archive digitization system, called trCAPTCHA, mainly targeting old Turkish archives and hence Turkish speaking audience. trCAPTCHA differs from reCAPTCHA that it uses no global dictionary but constructs a local dictionary for each scanned word from alternative texts generated through OCR readings.

Keywords: Human Computation, reCAPTCHA, Archive Digitization, trCAPTCHA

TEŐEKKÜR

Çalıőmalarım boyunca deęerli yardım ve katkılarıyla beni yönlendiren tez danışmanım Yrd. Doç. Dr. Osman Abul'a, yine çalıőmam boyunca manevi desteklerinden dolayı TOBB Ekonomi ve Teknoloji Üniversitesi Bilgi İşlem Birimi müdürü Hüseyin Çotuk başta olmak üzere çalıőma arkadaşlarım Ahmet Ömercioęlu, Yahya Şirin ve Mehmet Ali Karakaya'ya ve hayatım boyunca bana vermiş oldukları manevi destekten dolayı aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

ÖZET.....	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ	x
KISALTMALAR	xii
1. GİRİŞ	1
1.1. Yapay Zeka	3
2. İNSAN HESAPLAMASI.....	6
2.1. İnsan Hesaplaması.....	6
2.1.1. Açık Kontrol	7
2.1.2. İnsan Hesaplaması Yaklaşımı Algoritmaları	10
2.1.3. Görev Tasarımı ve Teşvikler.....	11
2.1.4. İnsan Hesaplaması Yaklaşımına Ait Sistem İncelemeleri	14
3. ARŞİV SAYISALLAŞTIRMA VE reCAPTCHA	27
3.1. Arşiv Sayısallaştırma	27
3.1.1. Arşiv Sayısallaştırma Konusunda Dünyada Durum	29
3.1.2. Arşiv Sayısallaştırma Konusunda Türkiye’ de Durum	30
3.1.3. Arşiv Sayısallaştırmada Türkiye’de Yaşanan Sorunlar	31
3.2. reCAPTCHA Sisteminin İncelenmesi	31
3.2.1. reCAPTCHA	31
3.2.2. reCAPTCHA Çalışma Sistemi.....	33
4. trCAPTCHA: TÜRKÇE ARŞİV SAYISALLAŞTIRMA SİSTEMİ.....	35
4.1. trCAPTCHA Sistem Tasarımı.....	36
4.1.1. Çekirdek Modülü	37

4.1.2.	Web Modülü	39
4.2.	Kullanılan Yazılım ve Teknolojiler	46
4.2.1.	OCR	46
4.2.2.	ORACLE.....	55
4.2.3.	VAADIN	57
4.2.4.	HIBERNATE	58
4.2.5.	SPRING	59
5.	DENEYSEL PERFORMANS ANALİZİ	61
5.1.	Kullanılan Veri Kümelerinin Oluşturulması.....	61
5.2.	Performans Unsurları	63
5.2.1.	Resmin Bozulması	63
5.2.2.	Resim Boyutları	64
5.2.3.	Sonuçların Sınıflandırılarak Seçilmesi.....	65
5.3.	Gerçek Ortamda trCAPTCHA	68
6.	SONUÇ	72
6.1.	Gelecek Çalışmalar	73
7.	KAYNAKLAR	75
8.	ÖZGEÇMİŞ	78

ÇİZELGELERİN LİSTESİ

Çizelge 4.1 OCR Programlarının veri setine göre sonuçları.....	55
Çizelge 4.2 Veri Tabanı Şeması.....	56
Çizelge 5.1 Şekil 5.1'nin OCR Tarafından Üretilen Sonuçları.....	62
Çizelge 5.2 Şekil 5.2'nin OCR Tarafından Üretilen Sonuçları.....	63
Çizelge 5.3 Veri Kümelerine Ait Sonuçlar	65

ŞEKİLLERİN LİSTESİ

Şekil 1.1 Kullanıcılara Gösterilen Bir Resim.....	3
Şekil 2.1 Açık Kontrolün Üç Alanı.....	8
Şekil 2.2 ESP Oyunundan Bir Görünüm	15
Şekil 2.3 Peekaboom Oyunundan Bir Görünüm.....	16
Şekil 2.4 Verbosity Oyunundan Bir Görünüm.....	18
Şekil 2.5 TagATune Oyunundan Bir Görünüm	20
Şekil 2.6 Mekanik Turk ismi verilen makinenin bir resmi	21
Şekil 2.7 AMT’ de Görev Oluşturma Ekran Görüntüsü	22
Şekil 2.8 AMT’nin Kullanıcılara Sunduğu İşlevsellikler	23
Şekil 2.9 CAPTCHA Örnekleri	25
Şekil 3.1 OCR Programları Tarafından Zor Çözümlenecek Bir El Yazısı Metni.....	32
Şekil 4.1 Uniform Bilgi Sistemi ve trCAPTCHA Giriş Ekranı Yapısı.....	35
Şekil 4.2 trCAPTCHA Akış Şeması	36
Şekil 4.3 Çekirdek Modülü Yapısı.....	37
Şekil 4.4 OCR Çıktı Yapısı.....	38
Şekil 4.5 OCR Programları Tarafından Üretilen XML dosyasının İçeriği	39
Şekil 4.6 OCR Sonuçlarının Yüklenmesi İçin Kullanılan Ekran.....	41
Şekil 4.7 OCR Sonuçlarının Sisteme Aktarılmadan Önce Gösterilen Ekran.....	41
Şekil 4.8 Sistemdeki Veri Setlerinin Sorgulandığı Ekran.....	42
Şekil 4.9 trCAPTCHA Kelimeleri Ekranı.....	43
Şekil 4.10 trCAPTCHA Resim Sonuçları Ekranı	44
Şekil 4.11 trCAPTCHA Dataset İstatistikleri Ekranı.....	45
Şekil 4.12 Yazı Tipi Courier, Yazı Rengi Siyah Olan Metin Resmi	53
Şekil 4.13 Yazı Tipi Courier, Yazı Rengi Gri Olan Metin Resmi	53
Şekil 4.14 Yazı Tipi Justy, Yazı Rengi Siyah Olan Metin Resmi	53

Şekil 4.15 Yazı Tipi Justy, Yazı Rengi Gri Olan Metin Resmi	53
Şekil 4.16 Yazı Tipi Times, Yazı Rengi Siyah Olan Metin Resmi.....	54
Şekil 4.17 Yazı Tipi Times, Yazı Rengi Gri Olan Metin Resmi	54
Şekil 4.18 Yazı Tipi Verdana, Yazı Rengi Siyah Olan Metin Resmi	54
Şekil 4.19 Yazı Tipi Verdana, Yazı Rengi Gri Olan Metin Resmi.....	54
Şekil 4.20 Vaadin Genel Mimarisi	57
Şekil 4.21 Hibernate Detaylı Mimarisi	59
Şekil 4.22 SPRING Mimarisi.....	60
Şekil 5.1 Veri Kümesi 1’den Seçilen Rastgele Bir Resim	62
Şekil 5.2 Veri Kümesi 2’den Seçilen Rastgele Bir Resim.....	63
Şekil 5.3 Standart Kelime Resmi	65
Şekil 5.4 Boyutları Değiştirilmiş Kelime Resmi	65
Şekil 5.5 Veri Kümesi 1 İçin Tahmin Oranları Dağılımı.....	67
Şekil 5.6 Veri Kümesi 2 İçin Tahmin Oranları Dağılımı.....	68
Şekil 5.7 Kullanıcı Etkisi Olmadan Doygunluğa Ulaşmış Kelime Sayısı	69
Şekil 5.8 Kullanıcı Etkisi Olmadan Veri Kümelerinin Doygunluk Durumu	70
Şekil 5.9 Kullanıcı Etkisi İle Doygunluğa Ulaşmış Kelime Sayısı.....	70
Şekil 5.10 Kullanıcı Etkisi İle Veri Kümelerinin Doygunluk Durumu	71

KISALTMALAR

Kısaltmalar Açıklama

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
APSA	The Adaptive Puzzle Selection Algorithm
ARMS	Archives and Records Management Section
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CRUD	Create, Read, Update Delete
CRT	Cathode Ray Tube
DAWG	Directed Acyclic Word Graph
DPI	Dot Per Inch
ESP	ESP Games
FPSA	The Fresh-first Puzzle Selection Algorithm
GPL	GNU Public License
GWAP	The Game With A Purpose
HQL	Hibernate Query Language
HTML	Hyper Textup Markup Language
OCR	Optical Character Recognition
OPSA	The Optimal Puzzle Selection Algorithm
ORM	Object/Relational Mapping
RIA	Rich Internet Applications
RPSA	The Random Puzzle Selection Algorithm
IP	Internet Protocol
RDBMS	Relational Database Management Systems
SQL	Structured Query Language
TOBB ETÜ	Türkiye Odalar ve Borsalar Birliği Ekonomi ve Teknoloji Üniversitesi
TÜİK	Türkiye İstatistik Kurumu
XML	Extensible Markup Language
JDBC	Java Database Connectivity
AMT	Amazon Mechanical Turk

1. GİRİŞ

Son 30 yılda bilgisayarlar akademik antikalardan endüstriyel makinelere dönüşmüş ve sonuç olarak modern yaşamın her alanına nüfuz etmiş, vazgeçilmez bir hal almıştır. Hafife alınan birçok gündelik işi yapmak, günümüz bilgisayarlarının geniş hesaplama yetenekleri olmasaydı imkânsız bir hal alırdı. Yüksek karmaşıklıkta kontrol sistemleri tarafından sürekli ve titizlikle yapılan düzenlemeler ile pilotsuz uçan hava araçları bulunmaktadır. Tüm dünyada ölçüm istasyonlarının giderek daha güvenilir hava tahminleri oluşturmaları için bilgisayarlar binlerce veriyi analiz edip birleştirmektedir. Bugün birçok tren, sürücüsü (insan) olmadan yoluna devam edebilmektedir. Bu listeyi farklı örnekler ile daha da genişletmek mümkündür. Genel olarak belirtmek gerekirse bilgisayarlar hızlı, daha doğru ve çok çeşitli sorunlarda insanların sezgiselliğine göre daha az hata eğilimi göstermektedir. Bununla birlikte zor ve farklı sorunların çözülmesi beraberinde henüz çözülmemiş daha zor ve daha karmaşık sorunları getirmektedir. Konuşulan cümleyi anlamak, bir kedi ile bir köpeği birbirinden ayırmak veya bir resmin içeriğini tanımlamak gibi ortalama beş yaşında bir çocuğun rahatlıkla yapabileceği işlemler şu anki ileri düzeydeki bilgisayar sistemleri tarafından oldukça zordur. Bu şaşırtıcı eksiklikleri gidermek için bulunan yaklaşımlardan bir tanesi, bilgisayarların, insanların düşündüğü şekilde düşünmelerini sağlayan yeni yöntemlerin araştırılıp denenmesidir. Yapay sinir ağları, makine öğrenme yaklaşımları bunlara birer örnektir. Benzer şekilde bu sorunları bilgisayarların çözebileceği hale getirecek daha iyi algoritmalar geliştirilmektedir. Bahsedilen yaklaşımlara ek olarak 2006 yılında Luis von Ahn tarafından geliştirilen “İnsan Hesaplaması (Human Computation)” yöntemiştir. Bu yöntem, insanların algı, kavramsal zeka gibi henüz bilgisayarların tam olarak sahip olamadığı yetenekleri bilgisayarların çözemedikleri sorunlarda kullanmak ve bilgisayarların insanlar gibi düşünmesine alt yapı sağlayacak olan verilerin sağlanmasını amaçlamaktadır.

Bu yaklaşım sayesinde insanların beyinlerini adeta birer bilgisayar işlemcisiymiş gibi kullanarak büyük ölçekli hesaplamaların küçük parçalarını çözmek

hedeflenmektedir. Bu yaklaşımın bir başka özelliği ise kullanıcıların farkında olarak veya olmayarak değişik yöntemler ile değerli çıktılar üretmelerini sağlamaktır.

İnsan hesaplama sistemlerinin bir örneği olan ve tez çalışmasında detaylarından bahsedilen ESP Oyunu (ESP Game); birçok insanın haftada 40 saatten fazla oynadıkları online bir oyundur. İnsanlar bu oyunu oynayarak, internet üzerinde bulunan görüntüler için anlamlı, doğru anahtar kelime etiketleri sağlamaktadırlar. Bu etiketler ile birlikte insanlar internet üzerindeki resimlerin arama doğruluğunu arttırmaktadır. Eğlenerek, farkında olmadan, amaçları sadece oyundaki puanlarını arttırmak olan insanlar bir yandan da internet üzerindeki sayısız resmi etiketlemektedirler. Bundan dolayı oyun olarak geliştirilen ESP tarzında insan hesaplama sistemlerine “Maksatlı Oyunlar (GWAP – Games With A Purpose)” denmektedir. Tez dokümanında “İnsan Hesaplama” isimli bölümde ESP Oyunu ve benzeri diğer maksatlı oyunlar incelenmiştir.

Gelişimsel açıdan bakıldığı zaman maksatlı oyunlardan sonra CAPTCHA sistemi gelmektedir. Bu sistem ile insanın algı ve seçiciliğinin bilgisayardan gösterdiği farklılık kullanılarak, bir güvenlik tedbiri olarak CAPTCHA'nın hangi alanlarda nasıl karşımıza çıktığına da değinilecektir. CAPTCHA, maksatlı oyunların ve diğer insan hesaplama sistemlerinin aksine insanlardan aldığı geri beslemelerin herhangi bir faydasının olmadığı CAPTCHA'yı kullananların insan olduklarını kanıtlamaya yaramaktadır.

CAPTCHA gibi insanlardan aldığı verilerin faydasız olduğu bir sistem, insan hesaplama yönteminin doğası ile uyuşmamakta ve gerçek bir problemi çözmemektedir. Bundan dolayı CAPTCHA sisteminin bir adım daha gelişmiş hali olan ve arşiv sayısallaştırma fikrinin temelini oluşturan reCAPTCHA sistemi geliştirilmiştir. Bu sistemin temeli optik karakter okuyucu programlarının, arşivlerin içeriklerinin tanımlanmasındaki yetersizliğe dayanmaktadır. Böylelikle OCR programları tarafından tanımlanamayan bölümlerin sonuçlarının insanlardan toplanmasıdır. Bu yaklaşımla insanlardan elde edilen veriler değerli kılınmaktadır. reCAPTCHA OCR' ların ürettiği sonuçları İngilizce sözlükten geçirerek sözlükte bulunmayan kelimelerin insanlar tarafından tanımlanması arzulanmaktadır.

Tez kapsamında reCAPTCHA gibi arşivlerin sayısallaştırma amacı taşıyan uygulamalara paralel bir uygulama, trCAPTCHA, geliştirilmiştir.

eyvallah etmezdi.

Şekil 1.1 Kullanıcılara Gösterilen Bir Resim

Kullanılan ve tezde de adı geçen bu uygulamalar kendi buldukları kültürün, o kültüre ait kullanılan dilin özelliklerini taşır. Örneğin, Şekil 1.1’de gösterildiği gibi bir resim kullanıcılara sunulurken, resimdeki kelimelerin kullanıcı tarafından tanınması ürettiği sonuçların doğruluğu açısından önemlidir. Türkçe dil ve kültürüne yakın olan kullanıcılar tarafından kolaylıkla çözümlenebilirken, farklı dil ve kültüre sahip kullanıcılar tarafından “evval lah etmezdi.”, “ewallab etmezdi.” gibi sonuçlar üretilebilmektedir. Uygulamayı kullanan kişinin ait olduğu kültür ve dil ile ne kadar yakından ilişkili bir resim seçerseniz, ilgili resim ile ilgili o kadar doğru ve çok sayıda bilgi elde edilebilir. Bu tezde anlatılan uygulama ile resimler için Türkçe içerik sağlayan bir arşiv sayısallaştırma sistemi geliştirilmesi amaçlanmaktadır. Uygulamanın çok sayıda kullanıcı tarafından kullanılması ile birlikte zor ve zahmetli olan arşiv sayısallaştırma süreci daha kısa sürede gerçekleştirilebilir.

1.1. Yapay Zeka

İnsan zekasına özgü algılayış, anlayış, biliş, düşünüş, karar verme gibi kapasitelerinin ve bunlara bağlı süreçlerin doğal olmayan yollarla herhangi bir şekilde üretme çabası yapay zekayı ortaya çıkarmaktadır. Canlı varlıklarda ortaya çıktığında zeka olarak belirtilen yetenek ve oluşlar analiz edilerek zekanın temel ve gerçek anlamda ne olduğu ortaya konulabilmektedir. Böylelikle bu yetenek ve oluşların, gerçekleştikleri çevresel etkenler ve diğer değişkenler göz önüne alındığında yapay olarak gerçekleştirilebilmesi çalışmaları yapay zeka olarak ortaya çıkmaktadır.

Yapay zeka kavramını ilk kez 1956 yılında kullanan ve terimin mucidi kabul edilen John McCarthy, yapay zekayı "Makineleri zeki yapan mühendislik ve bilim dalı." olarak tanımlamıştır [39]. Bu tanımdan yola çıkarak canlı varlıklardaki zekanın daha çok makineler üzerinde, onlara bu yetenekleri kazandırabilmek için uygulanmaya çalışıldığını ifade etmek mümkündür.

Yapay zekânın ortaya çıkışı bilimde birçok şeyi tetikleyen sebep olan savaş olmuştur yine. İkinci Dünya Savaşı'nda Alan Turing, Nazi Almanyası' nın Enigma makinesinin şifre algoritmasını çözmeye çalışan ünlü matematikçilerdendi. İngiltere, Bletchley Park'ta şifre çözme amacı ile başlatılan çalışmalar, Turing'in prensiplerini oluşturduğu bilgisayar prototipleri olan Heath Robinson, bombe ve Colossus bilgisayarları, Boolean cebirine dayanan veri işleme mantığı ile Makine Zekası kavramının oluşmasına sebep olmuştu. Böylelikle yapay bir nesne olan makineye zeka özellikleri kazandırma çabası yapay zekanın ortaya çıkmasına zemin sağladı.

Daha sonraları yapay zeka konusundaki ilk çalışma McCulloch ve Pitts tarafından yapılmıştır. Bu araştırmacılar, yapay sinir hücrelerini kullanan hesaplama modeli, önermeler mantığı, fizyoloji ve Turing'in hesaplama kuramına dayanan bir model olmuştu. Herhangi bir hesaplanabilir fonksiyonun sinir hücrelerinden oluşan ağlarla hesaplanabileceğini ve mantıksal işlemlerin gerçekleştirilebileceğini gösterdiler. 1950'li yıllarda Shannon ve Turing bilgisayarlar için satranç programları yazıyorlardı. İlk yapay sinir ağı temelli bilgisayar SNARC, MIT'de Minsky ve Edmonds tarafından 1951'de yapıldı. Daha sonra Newell ve Simon, "insan gibi düşünme" yaklaşımına göre üretilmiş ilk program olan General Problem Solver (Genel sorun çözücü) 'ı geliştirmişlerdir.

Yapay zeka üzerine araştırmalar yapan bilim adamlarının bir kısmı insan gibi düşünen sistemler kurmak isterken, diğer yandan sadece belli sorunların çözümüne yönelik rasyonel sistemler kurmak istediler. Bundan sonraki yıllarda ise mantığa dayalı çalışmalar çoğunlukta olmuştur ve başarımlarının izlenebilmesi için yapay sorunlar ve yapay dünyalar kullanılmıştır, simülasyon popülerlik kazanmıştır.

Daha sonraları bu başarımları ortaya koymak için kullanılan yapay sorunlar gerçek hayatı, insan yaşantısının içinde geçtiği ortamı temsil etmekten uzak oldukları için eleştirildiler. Bu eleştiriden yola çıkarak yapay zekanın yalnızca bu alanlarda başarılı olabileceği ve gerçek yaşamdaki sorunların çözümüne uygulanamayacağı öne sürülmüştür.

Her sorunu çözecek genel amaçlı program yerine belirli bir alandaki bilgi ve veriyle donatılmış programlar kullanma fikri yapay zeka alanına yeni bir soluk kazandırdı. Kısa sürede uzman sistemler adı verilen bir yöntem gelişti. Fakat burada çok sık rastlanan durum, bir otomobilin tamiri için önerilerde bulunan uzman sistem programının otomobilin ne işe yaradığından haberi olmamasıydı. Alan Turing; Turing testi olarak adlandırılan ve bir bilgisayarın veya başka bir sistemin insanlarla aynı zihinsel yetiye sahip olup olmadığını ölçen bir test geliştirmişti. Genel anlamda bu test bir uzmanın, makinenin performansı ile bir insaninkini ayırt edip edemeyeceğini ölçer. Eğer ayırt edemezse, makine insanlar kadar zihinsel yetiye sahip demektir. Bu testte bir insan ve bir bilgisayar, deneyi yapan kişiden gizlenir. Deneyi yapan hangisiyle haberleştiğini bilmeden bunların ikisiyle de haberleşir. Deneyi yapan kişinin sorduğu sorular ve deneklerin verdiği cevaplar bir ekranda yazılı olarak verilir. Amaç, deneyi yapanın uygun sorgulama ile deneklerden hangisinin insan, hangisinin bilgisayar olduğunu bulmasıdır. Eğer deneyi yapan kişi güvenilir bir şekilde bunu söyleyemez ise, o zaman bilgisayar Turing testini geçer ve insanlar kadar kavrama yeteneğinin olduğu varsayılır.

Tez çalışması kapsamında geliştirilen Türkçe olan ve reCAPTCHA benzeri bir sistem ile insan hesaplaması tekniğini kullanan çoklu örnekler birlikte tanıtılacaktır.

2. İNSAN HESAPLAMASI

Bilgisayarlar birçok görevi mükemmel bir şekilde yerine getirmektedir. Milyarlarca matematiksel işlemi saniyeler içinde yapabilir ve büyük miktarlardaki veriyi işleyebilir, sıralayabilir, filtreleyebilir ve önceden tanımlanmış kuralları uygulayabilir. Böylece, insanlar tarafından yapılması uzun zaman gerektiren çok sayıda işlem günümüz bilgisayarları tarafından çok daha hızlı bir şekilde yapılabilmektedir. Bilgisayarlar son 50 yılda pek çok açıdan önemli ölçüde gelişmiş olmalarına rağmen, çoğu insan için temel olan kavramsal zeka ya da algısal yeteneklere sahip değildirler. Bu bölümde bilgisayarların henüz sahip olmadığı bu yetenekleri öğretmek ve insanlar ile işbirliği sağlama yöntemi olan insan hesaplama incelenecektir.

2.1. İnsan Hesaplaması

İnternet üzerindeki resimlerin etiketlenmesi, taranmış bir kitabın yazıya aktarılması, Wikipedia'nın İngilizceden diğer dillere çevrilmesi... vb. sorunların çözümlerini, bilgisayarların kendi başlarına elde etmesi mümkün değildir. 60 yıldan fazla bir süre sonra bile bilgisayar bilimi üzerinde yapılan araştırmalarda da, bilgisayarların kötü performanslar ile gerçekleştirdiği sorunlar hala olacaktır. Algı ile ilgili konular hala bir bilgisayar tarafından yerine getirilmesi zor bir görevdir. Örneğin, uygulamaların resimleri kategorize etmek için kullandığı teknikler pek çok yönden yetersiz kalmaktadır. Çünkü bu uygulamalar çoğunlukla resmin içerisinde bulunduğu veya bulabildiği nesnelere göre tanımlamakta ve kategorize etmektedir. Bu yöntem yetersiz kalmaktadır, çünkü resimlerin genelini kavramsal olarak değerlendiremediği için de yanıltıcı, işlemesi zor sonuçlar ortaya çıkabilmektedir [9]. Farklı örnek ile değerlendirilecek olursa bir şarkının içerisinde hangi enstrümanların olduğunu, şarkının türünü ve uyandırdığı ruh halini belirlemek oldukça zordur [20,21,22]. Birçok görevin bilgisayarlar için zor olmasının bir başka nedeni de, dil engelidir. Herhangi bir metin, yazılı veya sözlü olsun, formal bir dil özelliğine sahip değildir. Ancak bunun yerine gazete makalesi gibi doğal dilden meydana gelen bir yazıyı anlamak, özetleme yapmak bir bilgisayar için yine de çok zor bir görevdir.

Sonuç olarak, çok karmaşık birçok hesaplama problemi vardır, örneğin tüm bu bilinen algoritmalar daha büyük problemlere ölçeklenememektedir. Bu iyi bilinen NP-complete problemlerini (gezgin satıcı problemleri, paketleme ve zamanlama sorunları vb. gibi) içermektedir. Biyoloji alanında da, protein katlanması ve genetik dizilimini belirleme dâhil olmak üzere birçok önemli sorunların hesaplama karmaşıklığının fazla olması, algoritmik yaklaşımların başarılı olmasında önemli bir engeldir.

Bilgisayarların tatmin edici çözümler bulamadığı büyük ölçekli görevleri insanların yardımıyla çözmek mümkün müdür? Bu sorudan yola çıkarak 2006 yılında Luis von Ahn tarafından ortaya "insan hesaplama (human computation)" kavramı atılmıştır. Bu kavram; insan beceri ve yeteneklerinden yararlanarak, bilgisayarların henüz çözmediği veya çözmekte yetersiz kaldığı büyük ölçekli hesaplamaları çözmek ve bilgisayarlara bu insan yeteneklerini öğretmeyi amaçlamaktadır.

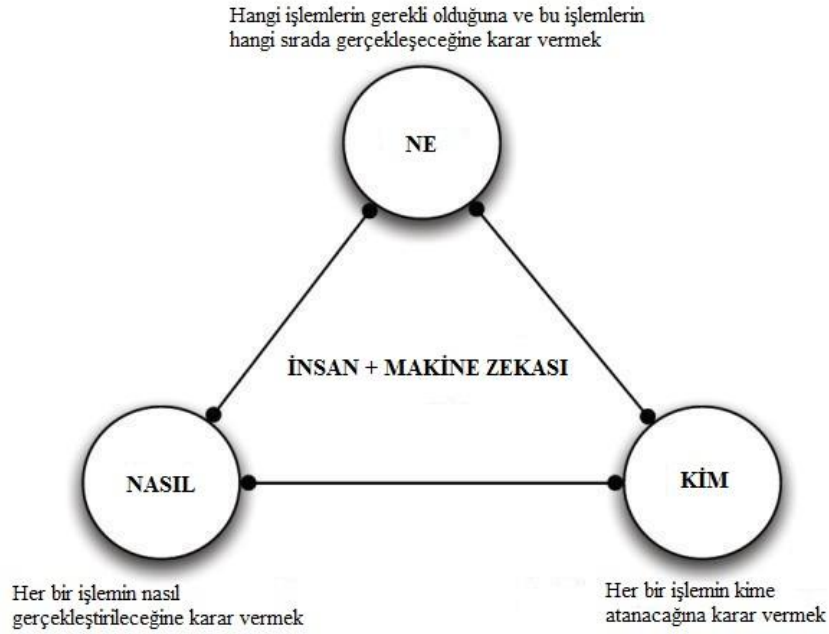
İnsan hesaplama terimi iki temel fikri kapsamaktadır. Birinci fikir, insanların zaten ilgilendikleri oyun oynamak, internet sitelerine (elektronik posta hesapları, sosyal ağlar, ... vb.) girmek/kaydolmak gibi işlemleri gerçekleştirirken öte yandan da anlamlı görevler yaptırılmayı hedeflemektedir. İkinci fikir ise insanların hesaplama işlemlerini nasıl gerçekleştireceklerini kontrol eden sistemler geliştirerek eldeki problemlerin doğru ve etkili bir şekilde çözülmesini hedeflemektedir.

2.1.1. Açık Kontrol

İnsan hesaplama yöntemi, kitlesel dış kaynak kullanma (*Crowdsourcing*), sosyal hesaplama (*Social computing*) ve kolektif zeka (*Collective Intelligence*) gibi birçok kavram ile ilişkilidir. Fakat insan hesaplama yönteminin ilişkili olduğu bu kavramların hiçbiri açık kontrol (*explicit control*) fikrini vurgulamamaktadır. Nitekim bu kavramlar sistemin kontrol edemediği veya kasten etmediği hesaplama sonucunun büyük bir kısmının bir grubun arasındaki doğal dinamikler (örn. koordinasyon ve rekabet) tarafından tanımlandığını varsaymaktadır. Bu kavramların açık bir ayrıştırma veya görevlerin atama yöntemleri olmadığı gibi insanların doğru

söylediklerini garanti eden açık bir şekilde tasarlanmış kontrol mekanizmaları bulunmamaktadır.

İnsan hesaplama arařtırmaları, insan davranıřı alıřmaları üzerine odaklanmak yerine, algoritmalar üzerine odaklanmaktadır. Bu algoritmalar ya neyin, nasıl ve kim tarafından iřleneceđini açıka belirtmektedir ya da iyi tanımlanmıř problemleri özmek için insanların abalarını organize etmektedir. İnsan hesaplamanın üç temel yönü vardır. Bunlar Őekil 2.1’de belirtilen “Ne”, “Kim”, ”Nasıl” yönleridir. Hem insanların hem de makinelerin bu açık kontrolden sorumlu olduđu unutulmamalıdır.



Őekil 2.1 Açık Kontrolün Ü Alanı

“Ne” Yönü

Bir hesaplama sorununa özüm üretmek amacıyla, problemin nasıl özüleceđini tam olarak özetleyen bir algoritmaya sahip olunması gerekmektedir. Bir dizi operasyon ve kontrol yapılarının kombinasyonlarına sahip olan algoritmalar, operasyonların nasıl düzenlenip alıřtırılacađını tanımlamaktadır. Geleneksel anlamdaki algoritmalara benzer bir Őekilde bazı insan hesaplama algoritmaları diđerlerine göre daha verimlidir. İnsan hesaplamanın “ne” yönü ile ilgili bazı arařtırma soruları ařađıdaki gibidir;

- Hangi görevler insan müdahalesine gerek kalmadan makineler tarafından gerçekleştirilebilir? Daha verimli ve doğru hesaplamalar yapabilmek için insanların ve makinelerin hesaplama yetenekleri yükseltilebilir mi [1] ?
- Karmaşık görevler yönetilebilir hesaplama birimlerine nasıl ayrıştırılır? Ayrıştırılan birimler insanların düşüncelerini işlemek için nasıl sıralanır?
- Birçok insan tarafından gelen temel gerçeğin olmadığı karmaşık çıktılar nasıl toplanır?

“Kim” Yönü

Bazı görevlerin uzman olmayan kişiler tarafından gerçekleştirilmesi yeterliyken, diğer görevler yoğun bilgi ve özel uzmanlık gerektirmektedir. İnsan hesaplamasının “kim” yönü ile ilgili bazı araştırma soruları aşağıdaki gibidir;

- Görevleri yönlendirmek için verimli algoritmalar ve ara yüzler (örneğin arama veya görüntüleme ekranları gibi) nelerdir?
- İnsanların zaman ile değişebilecek uzmanlıkları nasıl modellenilebilir?

“Nasıl” Yönü

Son olarak “nasıl” yönü sistemin insan katılımını nasıl motive edeceğine ve en iyi yeteneklerine göre hesaplama görevlerini nasıl yerine getirecekleri sorusuyla ilgilidir. İnsan hesaplamasının “nasıl” yönü ile ilgili bazı araştırma soruları aşağıdaki gibidir;

- İnsanların kişisel ihtiyaçlarına cevap veren bir ortam oluşturarak, sistem ile uzun süreli etkileşimi nasıl sağlanır?
- İnsanları doğru çıktılar üretmeye teşvik eden oyun mekanizmaları nasıl tasarlanır?
- İnsanların birbirleriyle nasıl ilişki kurduklarını tanımlamak için yeni pazarlar, organizasyonel yapılar ya da etkileşim modelleri nelerdir?

2.1.2. İnsan Hesaplaması Yaklaşımı Algoritmaları

İnsan hesaplama algoritmasını tanımlamak için sayısal algoritmalarda alışık olunan dil kullanılabilir. Bir algoritmanın iyi bilinen beş özelliği şunlardır;

- Girdi (*Input*)
- Çıktı (*Output*)
- Sonluluk (*Finiteness*)
- Yararlılık (*Effectiveness*)
- Belirlilik (*Definiteness*)

Sayısal algoritmalar gibi bu beş özellik aynı zamanda insan hesaplama algoritmaları için de uygulanır. Açıkça her insan hesaplama algoritması bir resim kümesi gibi bazı *girdilere* sahiptir. İnsan emeği ve hesaplama işlemlerinin kombinasyonu ile bir takım *çıktılar* üretir. İnsan hesaplama algoritmasının açık işlem kontrolü *sonluluğu* garanti etmelidir. Örneğin sonlu sayıdaki adımlardan veya bir süre sonra bir cevabın var olduğu gibi.

Yararlılık ifadesi, bir dizi özel talimatlara bağlı olarak belli bir miktar zamanda tamamlanabilen yeterince küçük ve kolay işlemlerin her biri anlma gelir. *Belirlilik* ifadesi genel olarak algoritmanın her bir adımı açıkça belirlenmiş ve kesin olarak tanımlanmıştır, algoritmanın tekrarlanması aynı sonucu tekrar üretmesi gibidir.

Açıkçası döngü içerisindeki insanlar söz konusu olduğunda bu özellik nadiren garanti edilir. Fakat görevleri daha küçük alt görevlere bölerek, mümkün olduğu kadar az belirsizlik elde etmeye çalışılır.

Kontrol Yapıları (Control Structures)

İnsan hesaplama algoritmalarının başlıca yeni unsuru, işlemlerin bilgisayarlar yerine insanlar tarafından gerçekleştirilecek olmasıdır. Ancak işlemler dizisini sıralamak için geleneksel algoritma tasarımlarında kullanılan kontrol yapılarının aynısı kullanılabilir. Bunlardan en önemlileri şunlardır;

- Dizi veya İterasyon (*Sequence or Iteration*): İşlem listesinin hangi sırada çalıştırılacağını belirtir.
- Seçim (*Selection or Choice*): Koşulları (*if-then-else*) ve koşul sağlandığı takdirde yürütülecek özel işlemi belirtir.
- Yineleme (*Repetition or Looping*): Sonlandırma koşulu sağlanıncaya kadar tekrar tekrar yürütülecek işlemi belirtir.
- Paralel (*Parallel*): Eş zamanlı yürütülecek işlemler kümesini belirtir.

2.1.3. Görev Tasarımı ve Teşvikler

Bu kısımda bir insan hesaplama görevinin tasarımı ve bu tasarımın bir sistemin performansına olan etkilerinden bahsedilecektir. Özellikle görev tasarımı çıktıların kalitesini, cevaba ulaşmak için geçen süreyi ve insanların çaba göstermeleri için bir teşviğin olup olmadığını etkilemektedir. İnsanlardan çabalarını ortaya çıkaran sistemlerin nasıl tasarlanacağına dair bilgiler verilecektir.

2.1.3.1. Görev Tasarımı

Bir insan hesaplama görevi tasarlarırken eş üretim sistemlerinin tasarımında kullanılan tanıdık kararlardan bazıları kullanılır. İnsan hesaplama görevi tasarımı sırasında dikkate alınması gereken beş önemli tasarım kararı aşağıdaki gibidir.

- Bilgi: Bilginin görüntülenmesi, sırası insanların performansları üzerinde etki etmektedir ve yanlılıkları soru cümlelerinde önemli rol oynamaktadır. Bundan dolayı insan hesaplama görevleri tasarlanırken gösterilecek bilgiler, sorulacak sorular insanları yönlendirmemek açısından titizlikle belirlenmelidir. Karmaşık görevlerde insanlara kısmi çözümler sunularak yardım edilebilir.
- Ayrıntı: Görev tasarımı için en uygun ayrıntı nedir sorusuna özel bir kapsam için deneysel olarak cevaplandırılabilir. Ancak tipik bir durumda küçük görevler büyük görevlere göre daha iyi sonuçlara yönelmektedir. Bir görevi ayrıntılarına

göre alt görevlere bölerek insanlara sunulması ile elde edilecek sonuçlar, görevin tamamıyla sunulmasından elde edilecek sonuçlardan daha etkilidir.

- **Bağımsızlık:** Her bir alt görevin birbirinden bağımsız olması karmaşık bir tasarım sorunudur. Diğer bir taraftan diğer insanların görevlerinden bağımsız olan görevleri yerine getirmek daha kolaydır. İnsan hesaplama görevleri olarak yapısal, kontrollü ve bağımsız görevler oluşturmak önemlidir.
- **Kalite Kontrol:** Yedekleme toplama, filtreleme tekniklerinin doğrulaması ile bağlanabilmektedir. Ek olarak sosyal normlara ve yaptırımlara itiraz edebilir, yasal kontratlar kullanılabilir, daha sofistike görev izleme gerçekleştirebilir, topluluk duygusu oluşturabilir yada gelecekte iyi insanları daha fazla iş ile ödüllendiren bir itibar sistemi kurulabilir.

2.1.3.2. Teşvikler

Geliştirilen insan hesaplama sisteminin en önemli performans unsurlarından bir tanesi olan teşvikler insanların gelip gelmeyeceklerini ve görevi yapıp yapmayacaklarını, hangi insanların görevi yapacaklarını, görevi ne kadar iyi yapacaklarını konu eden başlıklar ile tanımlanabilmektedir.

Farklı insan hesaplama sistemlerinde önemli rol oynayan teşvikler şunlardır; para, yüksek skorların listesi, eğitim, eğlence, katkıda bulunma bilinci (Galaxy Zoo, [SETI@Home](#) gibi projeler), gönüllülük, dünyaya yardım etmek istemektir.

- **Para:** Para ödülleri kullanmanın avantajı, rastgele bir ayrıntı seçilebilir ve ödemenin hangi seviyede doğru olduğunu test etmek kolaydır. Para ödülleri birçok insan için ilgi çekici olabileceği gibi amacı keyif almak olan insanlar üzerinde ise ters etki yapabilmektedir.
- **Erişebilirlik/Güvenlik:** CAPTCHA, reCAPTCHA ve tez kapsamında geliştirilen trCAPTCHA sistemlerinin teşvik modelini oluşturur. İnsan hesaplamayı erişim ve güvenlik kapsamında kullanmanın avantajı, insanlar belirli internet sitelerine ulaşabilmek için bu işi ücretsiz ve zorunlu olarak gerçekleştirmek zorunda

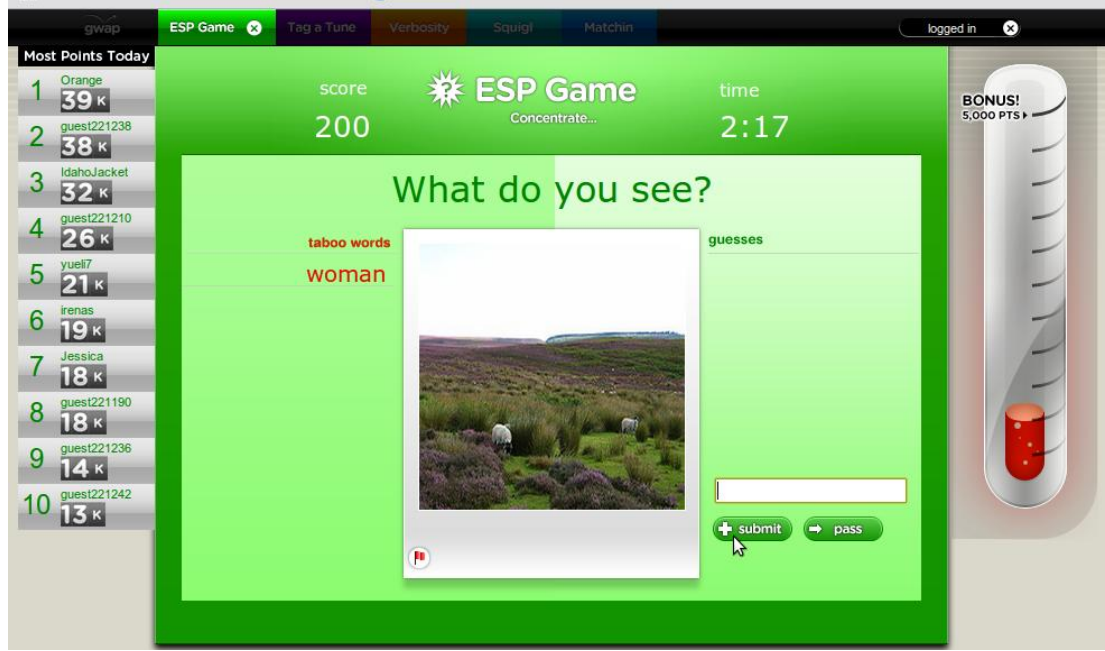
olmalarıdır. Tamamlanan işlerin fazlalığı internet sitesinin popülerliği ile doğru orantılıdır. Yeni fikrin adapte edilmesi oldukça zordur çünkü şuan ki fikirler (örnek olarak reCAPTCHA gibi) zaten çok iyi çalışmaktadır. Ek olarak bu işlerin başarı oranları % 90 veya üzerinde olması gerekmektedir aksi takdirde bu iş güvenlik/erişim işi olarak kullanılamaz.

- Oyun/Eğlence: Oyunların kullanılmasının avantajı birçok insanın doğasından kaynaklanan zevkli oyun oynayarak işlerin ücretsiz olarak gerçekleştirilmesidir. Eğlenceli bir oyun muazzam miktarda iş üretmektedir. Örneğin ilerleyen kısımlarda detaylı incelenecek olan ESP oyununu insanlar haftada 40 saatten fazla oynamaktadır. Fakat bu tarz eğlenceli oyunların tasarımı bilimden daha çok sanat gerektirmektedir. Büyük miktardaki oyuncuların sayılarını zaman içerisinde muhafaza etmek, oyuncuların oyuna karşı ilgilerinin giderek azalacağından dolayı zordur. Bu yaklaşım rastgele olarak ölçeklendirilemez. Para ödüllendirmesindeki yöntemin aksine daha çok insana ihtiyaç duyulduğunda oyunu daha eğlenceli hale getirmek kolay değildir.
- Gönüllülük/Katılım Duygusu: İnsan hesaplama sistemlerinin pozitif yönlerinden bir tanesi ise gönüllülük esasına (örnek olarak www.galaxyzoo.org) dayanmasıdır. İşin kalitesi belirsiz olabilir. Bazı kapsamlarda insanlar belki para ödeyerek elde edilen işlerden daha yüksek kalitede işler sağlarken, bu bazı kapsamlarda ise düşük kalitede çıktılar üreten uzman olmayan insanları çekebilir. Bu teşvik yönteminin en önemli problemi, kapsamın veya görevin insanların inandıkları bir şey olması zorunluluğudur. Kısaca insanlar, yaptıkları işlerin toplumu geliştirdiklerine inanmaktadır.
- Öğrenme/Merak: İnsanlara öğrenme deneyimi sağlayarak ödüllendiren insan hesaplama sistemleri çok nadir fakat son derece ilginçtir. Bu tarz bir insan hesaplama sistemi olan www.duolingo.com projesi hala beta sürümündedir. Bu proje büyük miktardaki metinleri çevirmeyi amaçlamaktadır. Amaç uzman çevirmenler kullanmadan çeviriler yapmaktır. Bunun yerine, işin büyük bölümü sadece yeni bir dil öğrenmeye başlayan kişiler tarafından yapılmaktadır.

2.1.4. İnsan Hesaplaması Yaklaşımına Ait Sistem İncelemeleri

2.1.4.1. ESP Oyunu (ESP Game)

İnternet üzerinde milyonlarca resim bulunmakta ve bu resimler için uygun metinsel açıklamaları sağlayacak yöntemler mevcut değildir. Bilgisayarlı görme (computer vision) henüz resimlerin içeriklerini tanımlayan geniş çaplı, kullanışlı bir uygulama oluşturamamıştır. Resimlerin doğru tanımlanması resim arama motorları, erişilebilirlik amacıyla, görme engelli bireylerin internet sitelerinde gezinmelerini sağlayan uygulamalar gibi birçok uygulama tarafından ihtiyaç duyulmaktadır. Şu anki resim kategorize eden uygulamalar birçok açıdan yetersiz kalmaktadır. Çünkü bu uygulamalar internet üzerindeki resimlerin içeriklerinin, buldukları internet sayfasındaki metinler ile ilişkisi olduğunu varsaymaktadır. Bu yanlış varsayımdan ötürü Luis von Ahn ve Laura Dabbish 2004 yılında insan hesaplamasının ilk uygulaması olan ESP oyununu oluşturmuşlardır. ESP kelimesi Extra Sensory Perception (Ekstra Duyusal Algılama) kelimelerinin baş harflerinden oluşan bir kısaltmadır. Bu oyun sayesinde son derece zahmetli ve sıkıcı olan resim etiketleme işlemini yarışma ortamı sunarak keyifli hale getirmişlerdir [9,10,11,12]. Bu gibi oyunlar sayesinde internetteki görüntülerin çoğunu etiketleme gibi bir hedef belirlemişlerdir [10].



Şekil 2.2 ESP Oyunundan Bir Görünüm
[<http://www.gwap.com/gwap/gamesPreview/espgame/> adresinden alınmıştır.]

Resim etiketleme oyunları genel olarak bilgisayarlı görme teknikleri araştırmaları için etiketlenmiş resimlerin bulunduğu, öğrenme algoritmalarını geliştirmek için ihtiyaç duyulan büyük veritabanlarının oluşturulmasına, aynı zamanda internet’de görüntü arama ve uygunsuz içeriğin filtrenmesi için bir yöntem oluşturmaktadır. Bu yaparken sadece belirli bir zümrenin bilgilerinden değil, sistemi kullanan toplumun her kesimden kullanıcıların ortak noktalarını esas almaktadır.

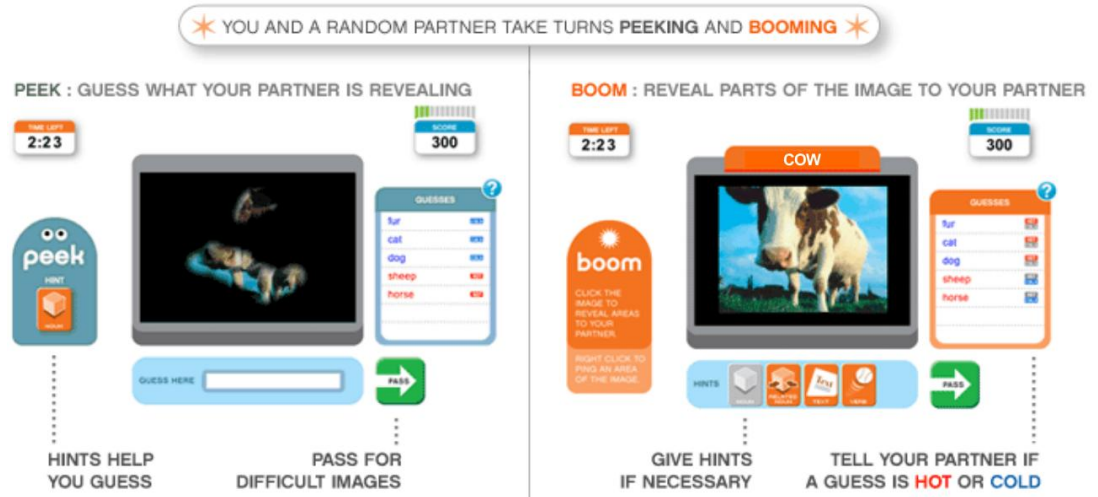
ESP Oyun Kuralları

ESP oyunu iki ortak tarafından oynanmaktadır ve aynı anda bir seferde birden fazla çift çevirim içi olarak bu oyunu oynayabilmektedir. Çiftler çevrimiçi olan kişiler arasından rastgele seçilmektedir. Çiftler karşısındaki beraber yarıştıkları kullanıcının kim olduğunu bilmez ve iletişime geçemezler. Tek ortak oldukları şey gördükleri resimdir [9,10,11,12]. Oyuncular iki dakika elli saniye içerisinde mümkün olduğunca çok resimle ilgili tahminler oluşturarak ilerliyorlar. Herhangi bir resim için ortak tahminde buldukları anda ESP oyunu oyunculara etiketlemeleri için bir başka resmi gösteriyor. Resimleri tahmin sırasında ortak kanılardan oluşturulmuş bir

yasaklı kelime listesi (taboo words) bulunmaktadır. Bu sayede resimler için farklı tahminler elde etmek amaçlanmıştır. Bir kullanıcı grubuna gösterilen resimler farklı kullanıcı gruplarına da gösterildiği için resimler hakkında maksimum etiketler oluşturuluyor.

2.1.4.2. Peekaboom Oyunu

İnsanlar az bir çaba harcayarak resimlerde hangi nesnelere olduğu, nerede olduklarına, arka planda nelerin olduğu v.b. özellikleri anlama ve analiz etme yeteneğine sahiptirler. Buna rağmen bilgisayar sistemleri resimdeki nesnelere yerlerini bulma gibi temel görevleri gerçekleştirilmede sorunlar yaşamaktadırlar. Önerilen ve test edilen bilgisayar ile görüntüleme teknikleri için kullanılan algoritmalar verimli sonuçlar vermemektedir [13]. Bu problemlerden dolayı 2006 yılında Luis Von Ahn tarafından resimler hakkında daha detaylı açıklamalar elde etmek amacıyla Peekaboom oyunu geliştirilmiştir [14]. ESP oyunundan farkı; esp oyunu resimler hakkında bilgiler verirken, Peekaboom ise bir resimde çizilmiş nesnelere bölgelerini vermektedir.



Şekil 2.3 Peekaboom Oyunundan Bir Görünüm
[http://peekaboom.org adresinden alınmıştır.]

Şekil 2.3’ de gösterilen ekran görüntüsünde sol taraftaki ekran Peek rolüne sahip oyuncunun ekranı, sağ taraftaki ekran ise Boom rolüne sahip kullanıcının ekranını göstermektedir.

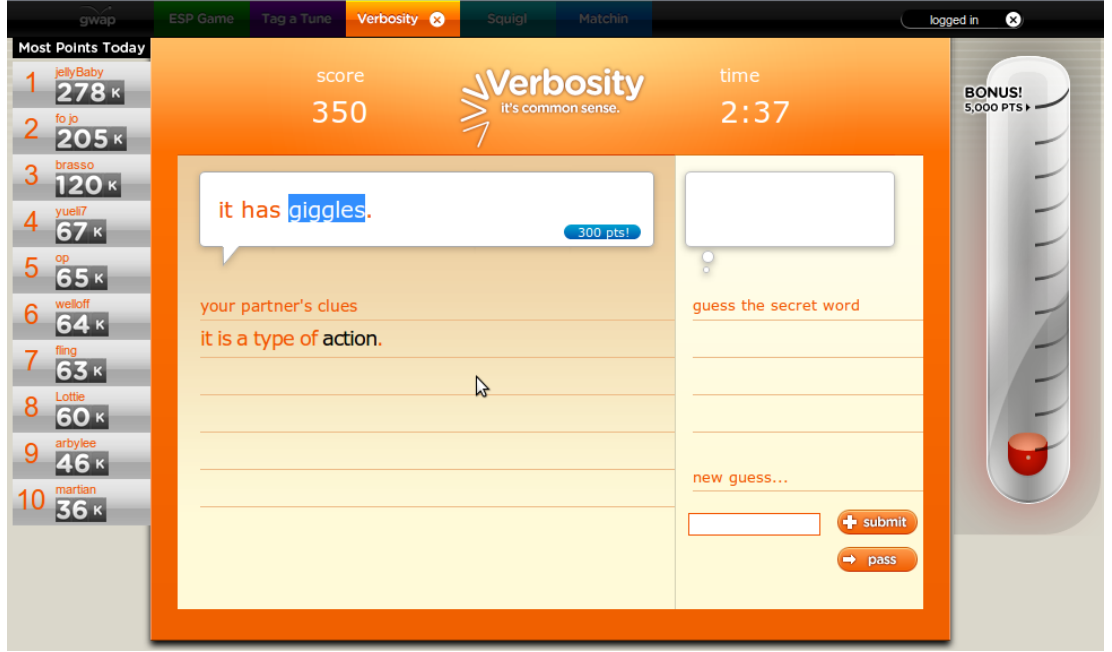
Peekaboom Oyun Kuralları

Diğer maksatlı oyunlardaki (GWAP) gibi Peekaboom oyunu da her oyun oturumu için rastgele seçilmiş iki oyuncu ile oynanmaktadır. ESP oyununda, oyun anında aynı işlemleri gerçekleştiren tek bir role sahip oyuncular olmasına rağmen, Peekaboom oyunundaki oyuncular farklı iki role sahiptirler. Birinci oyuncuya rastgele bir resim ve bu resme ait bir tanımlama gösterilmektedir. Bu tanımlamalar ESP oyunundan üretilmiş tanımlamalar da olabilmektedir. Birinci oyuncu gösterilen resmin parçalarını ikinci oyuncuya gösterir. Birinci oyuncuya Boom adı verilir. İkinci oyuncu ise sadece resmin birinci oyuncu tarafından gösterilen alanlarını görebilmektedir. İkinci oyuncuya da Peek adı verilir. Peek’in görevi kendisine gösterilen resim alanlarına göre Boom’un gördüğü kelimeyi tahmin etmektir. Eğer ikinci oyuncu tahmin edilmesi istenen kelimeyi bulursa bu başarımı bir puan ile ödüllendirip, yeni resimler gösterilir ve aynı adımlar tekrar edilir. Peek ve Boom birbirleriyle haberleşemedikleri için Boom rolüne sahip oyuncu gördüğü kelimeyi resim üzerinde düzgün bir şekilde işaretlemeye gayret etmelidir.

2.1.4.3. Verbosity

Birçok farklı resim etiketleme ile ilgili oyunlar olmasından başka dünya hakkında basit gerçekleri toplamak için Luis Von Ahn ve arkadaşları tarafından Verbosity adlı bir oyun geliştirilmiştir [19]. “Kedi bir hayvandır.” cümlesi hemen hemen her insan için ortak, gerçek bir bilgi iken bilgisayarlar için herhangi bir şey ifade etmemektedir. Geçen 20 yıl boyunca herkes için ortak, gerçek bilgilerin bir veritabanında toplanması ile ilgi yoğun çabalar gösterilmiştir [15, 16, 17]. Böyle gerçek bilgilerin olduğu geniş bir veritabanını otomatik muhakeme yeteneği ile birleştirildiğinde, bilgisayar sistemlerinin yeteneklerini daha da arttırabilir. Bu gerçekleri elle girmek can sıkıcı olduğu gibi aynı zamanda hata yapma payı

yüksektir. Bu problemten kaynaklanarak bu tarz verileri bir oyun yardımıyla toplamak gerçekleştirilmesi mümkün bir alternatiftir.



Şekil 2.4 Verbosity Oyunundan Bir Görünüm
[<http://www.gwap.com/gwap/gamesPreview/verbosity/> adresinden alınmıştır.]

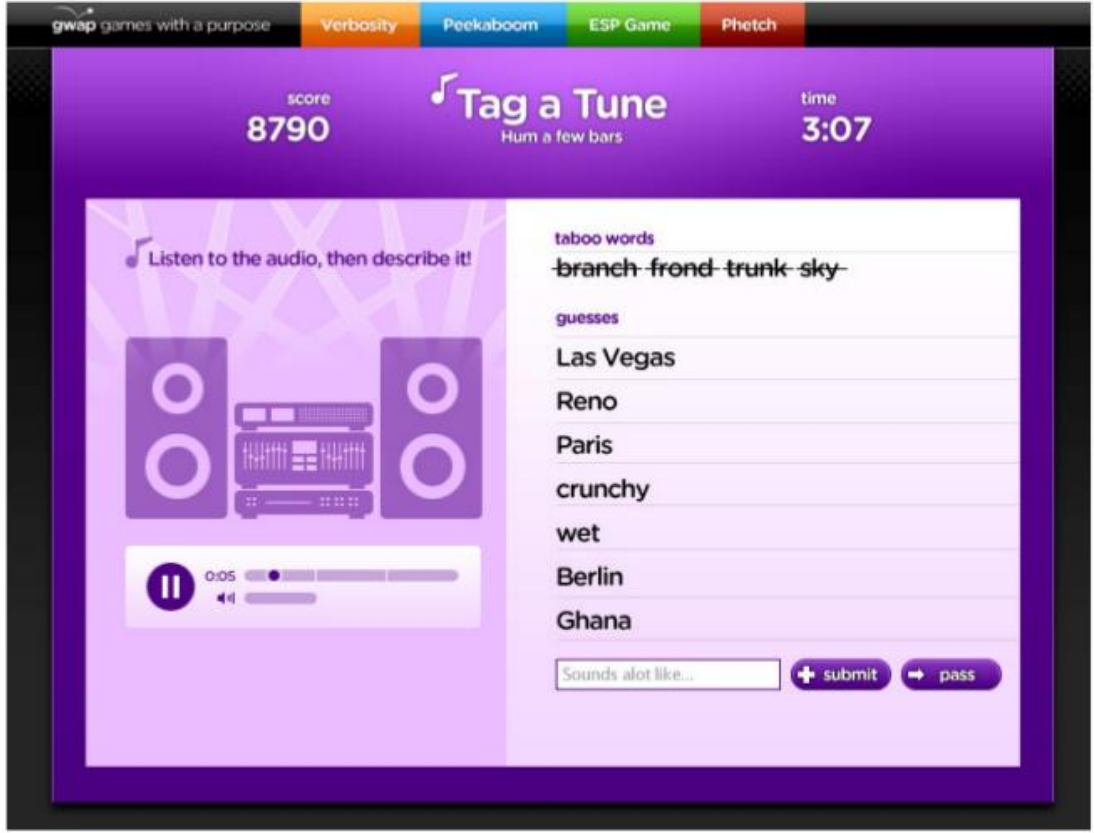
Verbosity Oyun Kuralları

Verbosity Taboo[18] adı verilen oyundan esinlenmiştir. Verbosity ve ESP arasındaki fark oyuncuların Verbosity’de farklı roller üstlenmesidir. Rastgele seçilen iki oyuncu tarafından çevrimiçi oynanan Verbosity’de oyuncuların biri anlatıcı (Narrator) diğeri ise tahmin edici (Guesser) olur. Anlatıcı gizli bir kelime alır ve diğeri oyuncunun kelimeyi tahmin etmesi için ipuçları gönderir. İpuçları daha önce belirlenmiş şablonlardan oluşur ve anlatıcı 7 şablondan birini istediği sırada seçerek ve şablondaki boş yere istediği veya alakalı kelimeyi yazarak diğeri oyuncuya gönderir. Örneğin, kelime “Dizüstü Bilgisayar”, anlatıcı size diyebilir ki : "Bir klavyesi var." [19] Diğeri oyuncu gelen cümleye göre gizli kelimeyi tahmin etmeye çalışır. Anlatıcı tüm tahminleri görebilir ve yaklaşıp yaklaşmadığı konusunda bilgi verebilir [19]. Böylelikle bir oyunda tek bir kelime ile alakalı cümleler toplanmış olur. Böylelikle de nesnelere özellikleri ve onun parçaları, etmenleri vb kullanıcılar

tarafından tespit edilmiş olur. Örneğin, anlatıcı “Bir klavyesi var.” dediğinde dizüstü bilgisayarın bir klavyesi olduğunu öğrenmiş oluruz [19].

2.1.4.4. TagATune Oyunu

İnsanlar bazı şeyleri düzenlemek amacıyla ve sonraki zamanlarda bunlara erişimi kolaylaştırmak amacıyla etiketler. İnternet üzerindeki multimedya nesnelерinin çoğalmasıyla birlikte internetin içeriğini düzenlemek için ortak etiketleme yaygın bir strateji olarak ortaya çıkmıştır. Bu stratejiden yola çıkarak Luis Von Ahn ve Edith Law ses dosyalarını etiketlemek için TagATune adında maksatlı bir oyun geliştirmişlerdir [8]. Prototip olarak geliştirdikleri bu oyunda ses dosyalarına etiket oluşturmak için ESP oyunun mekanizmasını kullanmışlardır. Bu yaklaşımı kullanmak beraberinde birçok problem getirmiştir. İki oyuncunun ses dosyasının etiketi üzerinde anlaşması resmin üzerinde anlaşmasından daha zordur. Sonuç olarak prototip üzerinde oynanan oyunların % 36’sında oyuncular pas demişler ve maksatlı oyunların amaçlarından bir tanesi olan hoş vakit geçirme sağlanamamıştır. Bu problemi çözmek için ESP oyunu mekanizmasını kullanan prototipi, girilen bilgide anlaşma (Input Agreement) adında yeni bir mekanizma ile tekrar geliştirmişlerdir. Bu yeni mekanizma ile geliştirilmiş oyunu oynayan kullanıcıların % 0.5’ i pas demişler ve oynanan oyunların % 80’inde oyuncular aynı cevabı vermişlerdir.



Şekil 2.5 TagATune Oyunundan Bir Görünüm
[<http://www.gwap.com/gwap/gamesPreview/tagatune/> adresinden alınmıştır.]

TagATune Oyun Kuralları

Diğer maksatlı oyunlar gibi TagATune da iki kişinin ortaklaşa oynadığı bir oyundur. Her turda, her bir oyuncuya kısa bir müzik çalınır. Müzik her iki kullanıcı için aynı olabilir fakat bu her zaman geçerli değildir. Kullanıcılar dinledikleri müzikleri isteğine bağlı tanımlar. Kullanıcıların yaptıkları tanımlamalar birbirlerine gösterilmektedir. Tanımlamalar yapıldıktan sonra oyuncular dinletilen müziğin aynı veya farklı olduklarını seçebilirler. Oyuncular doğru seçim yapıp, bu seçim üzerinde anlaşılırsa, puan ile ödüllendirilirler. Oyuncular aynı fikirde olduklarında ayrı ayrı müziklere girdikleri tanımlar doğru olarak dikkate alınır. Oyuncular, etiketler üzerinde anlaşmayı başarmaktan daha kolay olan ikili seçimde (aynı, farklı) anlaşmaya varmak zorundadır. Bu önemli olan keyifli bir oyun olma fikrini desteklemektedir.

2.1.4.5. Mechanical Turk

Bu bölümde incelenen insan hesaplama sistemlerinin (ESP, VERBOSITY... vb.) ortak özelliği, işin yapılması için motivasyon unsuru olarak herhangi bir parasal ücrete ihtiyaç duymamalarıdır. Nitekim bazı alanlarda kullanıcılara para ödemedi iş yaptırmak mümkün değildir. Amazon Mechanical Turk (AMT) gibi binlerce kullanıcının küçük ölçekli insan hesaplama görevlerini (örneğin: bir resim etiketleme, bir cümlenin çevrilmesi... vb) tamamladığı ve karşılığında para kazandıkları kitlesel dış kaynak kullanan (*crowd sourcing*) platformlar bulunmaktadır.



Şekil 2.6 Mekanik Turk ismi verilen makinenin bir resmi

Amazon Mechanical Turk 2005 yılının sonlarına doğru yayınlanmıştır [35]. Mechanical Turk ismi 18. yüzyılda Macar asıllı Wolfgang von Kempelen tarafından geliştirilen dünyanın ilk satranç oynayan makinesinin isminden gelmektedir. Wolfgang von Kempelen'in geliştirdiği makine Şekil 2.6'da gösterilmiştir.

AMT istemcilerin (*requester*), çalışanlara (*workers*) görevler gönderdiği ve çalışanların gönderilen bu görevleri tamamladıktan sonra karşılığında ücret aldıkları bir insan hesaplama platformu sunmaktadır. İstemcilerin dağıttıkları bu görevlere Human Intelligence Tasks'in baş harflerinden oluşan HITS denilmektedir. Kullanıcı motivasyonu olarak parayı kullanan AMT kısa sürede giderek popüler hale gelmiştir. 2010 yılında AMT'de yaklaşık 400.000 çalışan bulunmaktaydı [34].

Design HIT Templates

Welcome! You have not created any templates yet. To get started, use one of the sample templates below. Visit the [Resource Center](#) or read the [User Guide](#) for help.

Sample HIT Templates		
HIT Template Name	HIT Title	
Basic Open-ended Question	Answer a Simple Question See an example	Start with this template
Blank Template	Default Title See an example	Start with this template
Data Collection	Find the Website Address for Restaurants See an example	Start with this template
Data Correction	Provide the correct spelling of search terms See an example	Start with this template
Data Extraction	Get Product Name from Image See an example	Start with this template
Image Filtering	Flag offensive content images (WARNING: This HIT may contain offensive content. Worker discretion is advised.) See an example	Start with this template
Image Tagging	Tag an image See an example	Start with this template

Şekil 2.7 AMT' de Görev Oluşturma Ekran Görüntüsü
[<https://www.mturk.com/mturk> adresinden alınmıştır.]

Mechanical Turk üzerinden dağıtılan tipik görevler, sınıflandırma (örn: görüntü, müzik, belgeler... vb), çeviri ve orijinal içerik oluşturma (yorumlar, hikayeler, blog iletileri... vb) içerir [36]. Psikologlar, sosyologlar ve ekonomistler daha önce laboratuvar ortamlarında yapılan deneylerini Mechanical Turk üzerinden dağıtmaya başlamışlardır. Böylelikle daha büyük, küresel ve heterojen bir kitle havuzuna sahip olmaktadır [37].

The screenshot shows the Amazon Mechanical Turk (AMT) interface. At the top, there is a navigation bar with the Amazon Mechanical Turk logo and the text 'Artificial Intelligence'. There are buttons for 'Your Account', 'HITs', and 'Qualifications'. A notification indicates '96,650 HITs available now'. Below the navigation bar, there is a search bar and a filter for 'HITs'. The main content area displays a list of HITs under the heading 'All HITs'. The list shows 1-10 of 2085 results. The first four HITs are visible, each with a title, requester, expiration date, reward, and time allotted. The HITs are: 'Copy Text from Business Card', 'Flag offensive content images (WARNING: This HIT may contain offensive content. Worker discretion is advised.)', 'WHERE ARE YOU? A 5-second HIT from techlist. Instant automatic approval.', and 'Just Click a Button (Instant Qualification)'. Each HIT has a 'View a HIT in this group' link.

Requester	HIT Expiration Date	Reward	Time Allotted	HITs Available
Oscar Smith	Dec 3, 2010 (1 hour 59 minutes)	\$0.02	10 minutes	2
ProBoards, Inc.	Dec 17, 2010 (1 week 6 days)	\$0.01	7 minutes	1
techlist	Dec 3, 2010 (2 hours 45 minutes)	\$0.01	5 minutes	169
Paul H	Dec 3, 2010 (13 minutes 57 seconds)	\$0.01	2 minutes	1

Şekil 2.8 AMT'nin Kullanıcılara Sunduğu İşlevsellikler
[https://www.mturk.com/mturk adresinden alınmıştır.]

AMT' deki görevler bir API üzerinden programsal olarak oluşturulabilmesinin yanında Şekil 2.7'de gösterildiği gibi bir dizi şablon yardımıyla el ile de yapılabilmektedir. AMT sistemine çalışan olarak katılan kullanıcılara Şekil 2.8' deki gibi çeşitli işlevsellikler sunulmaktadır. Böylelikle çalışanlar (kullanıcılar) kendilerine uygun (parasal, içerik, bitiş süresi... vb) görevleri görebilir ve bu görevleri kabul edebilirler.

2.1.4.6. CAPTCHA

Birçok durumda yalnızca insanların sisteme erişimine izin vermek istenmektedir. Aksi durumda internet forumları veya elektronik posta hesapları gibi iletişim araçlarında istenmeyen reklamların dağıtılması şeklinde yapılan istismar örnekleri bulunmaktadır. Bir başka örnek ise otomatik oylamanın mümkün olduğu, kolayca manipüle edilebilen internet anketlerdir. Yapılması istenilen işlemlerin bilgisayarlar tarafından otomatik olarak değil de bizzat insanlar tarafından yapılmasını sağlamanın gerekli olduğu kanısına varılmıştır. Makine hesaplamaları ve insan düşüncüsü arasındaki farkların avantajını kullanmak için çeşitli yaklaşımlar geliştirilmiştir [7].

“İnsanlar geçebilir fakat bilgisayarlar geçemez.” prensibi ile çalışan CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart),

2000 yılında Manuel Blum ile birlikte öncü çalışmalar yapan Luis Von Ahn tarafından geliştirilen ve tamamen otomatik Turing Testi olarak tanımlanan bir uygulamadır. Turing testi bilgisayarın, insanlardan ayırt edilememeleri durumunda başarılı sayıldıkları bir test olarak kabul edilir. CAPTCHA ise bu testin tam tersi olarak düşünülebilir. Bilgisayarların çözemeyeceği ama insanların çözebileceği bir testtir. Bazı durumlarda bilgisayarlar kendilerinin çözemeyeceği veya çözmesi zor olacağı problemler üretebilirler. Örneğin bir bilgisayar için verilen bir yazıyı resme çevirmek ve resmi deforme etmek önemsiz bir iş iken bu işin tam tersinin gerçekleştirilmesi oldukça zordur. Bir sistemin potansiyel kullanıcılarına (bu kullanıcılar insan veya bilgisayarlar olabilir) bu tarz üretilmiş bir problem gösterilir ve kullanıcıdan üretim işleminin tersine çevrilmesi istenir. Sistem sonucu, bu problemi oluşturmak için kullandığından dolayı bilmektedir. Bundan dolayı sistem kullanıcı tarafından girilen girdinin doğruluğunu kolayca karşılaştırabilir. Sistemin insan kullanıcıları tarafından bu problem kolayca çözülebilir.

Veriler ve kullanıcılara gösterilecek resimleri üretmede kullanılan algoritmalar herkes tarafından ulaşılabilir. Bu durum böyle olmasaydı, hatalı tasarımlar, kolayca kırılabilir sorunlara yol açabilirdi. Bilgisayarların insan zekâsından daha iyi olduğunu savunan görüşler olsa da bu uygulama karşısında bu görüşler çürümektedir. Çünkü harfleri tanıma yeteneğine sahip olan bilgisayarlar deforme edilmiş bu harf kümelerini çoğu zaman tanımamaktadır. Oysa insanlar çoğu zaman hatasız olarak tanımakta ve bu testi geçebilmektedir [3,4,5]. Bu yöntem insan zekâsı ve insan algısının bilgisayardan üstün olabileceğini de ortaya çıkarmış olmaktadır. Bu testi doğru yanıtlayanın insan olduğu kabul edilmektedir. Şekil 2.6'deki CAPTCHA örneklerine baktığımızda yazı içeriğinin anlaşılmasını güçleştirmek için tahrip ve üzerinde bazı işlemler gerçekleştirildiği görünmektedir.



Şekil 2.9 CAPTCHA Örnekleri

CAPTCHA sadece resim olarak düşünülemez. Zira farklı yöntemler ile bilgisayarların testi geçmesi zorlaştırılabilir. Şunu unutmamak gerekir ki her uygulamayı çürütmek için yeni uygulamalar da geliştirilmektedir. CAPTCHA bilgisayarlara yapılan “spam amaçlı botların” önlenmesi için geliştirilmiş olsa da öyle karşı uygulamalar geliştirilmiştir ki zaman içinde CAPTCHA dahi yetersiz kalmıştır. Bu nedenle CAPTCHA için de sürekli geliştirilmesi ve teknolojiye uyarlanması gereken bir uygulamadır demek yerinde olacaktır. Örneğin CAPTCHA sadece resim olarak kabul edilirse görme engelli kullanıcılar için bazı durumlarda giriş yapmak olanaksız hale gelecektir. Bilgisayar ile insanı ayırt etmeye çalışırken görme engelli kullanıcılarında erişimini engelleyeceğinden, farklı bir uygulama geliştirilmiş ve bu noktada ses kullanarak görme engelli insanların da bu testi geçmesi sağlanmıştır. CAPTCHA'nın kullanıldığı uygulamalardan bir kısmı şöyledir;

- İnternet Anketleri: Kasım 1999’da slashdot.com “Hangi üniversite bilgisayar bilimlerinde en iyi?” sorulu bir internet anketi yayınlamıştır. Diğer internet anketlerde olduğu gibi aynı IP adresi üzerinden birden fazla oy kullanılmasını engellemek amacıyla oy kullananların IP adresleri kayıt altına alınmıştır. Carnegie Mellon üniversitesindeki öğrenciler üniversitelerini binlerce kez

oylayan bir program geliřtirmişler ve Carnegie Mellon üniversitesinin sonucu hızla büyümeye başlamıştır. Diğer günlerde MIT üniversitesinden öğrenciler kendi oylama programlarını yazmışlar ve bu anket, oylama botları arasındaki bir yarışmaya dönüşmüştür. Anket bittiğinde MIT 21.156 oy, Carnegie Mellon üniversitesi 21.032 oyla ve diğer üniversiteler 1.000 oyla bitirmişlerdir. Bu tarz sonuçlara sahip internet anketlerine nasıl güvenilebilir? Bundan dolayı bu anketlerin sadece insanlar tarafından yanıtlanması istenmektedir.

- Bedava Elektronik Posta Servisleri: Yahoo, Microsoft, Google gibi bedava elektronik posta hizmetleri sağlayan firmalar önceden özel bir tip saldırıya maruz kaldılar. Bu saldırı binlerce elektronik posta hesabı alan botlardı. Yeni bir elektronik posta hesabını alırken bu hesabı açmak isteyen insan olduğunu kanıtlamak gerekli olduğundan dolayı CAPTCHA'lar kullanılmıştır.
- Solucanlar ve Spam: CAPTCHA'lar elektronik postalarda solucan ve spam karşıtı makul bir çözüm sunuyor. Çözüm diğer bilgisayar başında insan olduğunu bildiği zaman elektronik postaları kabul ediyor. Diğer bilgisayarın başında insan olup olmadığı da CAPTCHA'lar vasıtasıyla biliniyor. www.spamarrest.com gibi birkaç firma bu fikri zaten ticarileştirmiştir.
- Sözlük Saldırıların Önlenmesi: Pinkas ve Sander [6] CAPTCHA'ları şifreli sistemlerde sözlük saldırılarını önlemek amacıyla kullanılabileceğini önermişlerdir. Önerdikleri sistem basitçe şu şekilde; bilgisayarların rastgele şifreler deneyerek sisteme girmeye çalışmasını önlemek ve sisteme girmek isteyen insan olduğunu kanıtlamak.

3. ARŞİV SAYISALLAŞTIRMA VE reCAPTCHA

3.1. Arşiv Sayısallaştırma

Tüzel ya da gerçek kişilerin birbirleriyle olan ilişkileri sonucu ortaya çıkan yazılı belgelerin, çeşitli görsel ve işitsel kaynakların, belirli nesnelere bir amaç doğrultusunda sistemli olarak ya da belirli süreçler doğrultusunda saklanması işlemi genel olarak arşivlemeyi tanımlamaktadır. Arşivin ortaya çıkışını da devletlerarası ilişkiler sağlamıştır. Devlet ile vatandaşları arasındaki ilişkiler ve devletin kendi kurumları arasındaki ilişkiler ise arşivin gelişmesini sağlamıştır.

Papirüs, deri parçaları, vb. malzemeler üzerinde ortaya çıkan belgeler ve buna bağlı saklama, arşivleme tarihi gelişimini kağıt üzerinde ve kopyalama yöntemleriyle birlikte daha gelişmiş ortamlarda ve daha gelişmiş yöntemlerle sürdürmüştür. Belgeler üzerindeki ham veri ve belgelerin kendilerinin herhangi bir sınıflama, sıralamaya maruz kalmamış niteliksiz hali de zamanla yeterli gelmemiştir. Belgelere gizlilik dereceleri, tarih, dosya plan kodları gibi tanımlayıcı, sıralayıcı nitelikler eklenmiştir. Belgelerin sonraki zamanlar için durumlara ve olaylara delil olmalarının yanında üzerlerindeki verilerin de enformasyona dönüştürülebilmesi arşive bir başka boyut kazandırmıştır.

Belgelerin iletimi, kopyalanması, saklanması gibi işlerde kağıt ortamının yetersiz kalması, devletlerin ve insanların gelişen beklentileri, elektronik çağın başlangıcıyla birlikte arşivlemede yeni yaklaşımları, yeni yöntemleri ortaya çıkarmıştır.

Verilerin ve belgelerin saklanabileceği ortamlar çeşitli askeri ihtiyaçlar doğrultusunda sürekli gelişim göstermiş ve bunun sonucunda farklı yöntemler ortaya çıkmıştır.

En temel tanımıyla belgelerin elektronik, sayısal ortamda saklanmasıdır. Fakat belgelerin kâğıt ve benzeri ortamlarda basılı olarak bulunması sebebiyle dijital arşivi kâğıt ve benzeri ortamlarda basılı olarak bulunan belgelerin elektronik ortamlarda

saklanması süreci olarak tanımlayabilmekte mümkündür. Asıl amaç belgelerin saklanması olarak görünse de arşivlemek sadece belge saklamak değildir. Ham verilerin yüksek oranda bulunabilirliği ve işlenebilirliği de amaçlanmaktadır. Dolayısıyla dijital arşivi belgeleri saklamak ve bulunabilirliğini, bilgiye dönüştürebilirliğini sağlamak şeklinde tanımlayabiliriz.

Dijital arşivlerde belgelerin özellik ve nitelik açısından da donatılabilirliği sağlanmaktadır. Etiketleme, tasnif etme, sınıflandırma, karşılaştırılabilirlik özellikleri ile donatılmış belgeler bilgi hiyerarşisi açısından belgeyi anlamlı bilgiye dönüştürebilmeyi de mümkün kılmaktadır. Ham veri arşive alınmış belgenin ilk hali olarak düşünülebilir. İçerisindeki veriler arşive eklendiğinde sınıfı ya da tarihi hakkında da bir veri mevcuttur fakat belgenin içeriği işlenmediği ya da başka bir kavramla ilişkilendirilmediği için ham veri şeklindedir. Tanımda da belirtildiği üzere arşivin sadece saklamak işlemi olmadığı düşünüldüğünde belgedeki veriyi enformasyona yaklaştırmada ilk adım arşivlemek olarak düşünülebilir. Fakat bu da tek başına yeterli değildir. Klasik arşivleme yöntemleri düşünüldüğünde bu işlem çok zahmetlidir ayrıca pek çok açıdan kısıtlılıklar mevcuttur. Kavramsal olarak ele alındığında dijital arşive yönelimin, ortaya çıkışının bu şekilde sebepleri bulunmaktadır.

Bilgi teknolojilerinin tetiklemesiyle üretilen bilgi miktarının geçmiş yüzyıllarla karşılaştırılmayacak ölçüde artış göstermesine rağmen karmaşıklaşan bilgi yığınlarına erişimin geçmişten daha güç olması ilginç bir ironiyi oluşturmaktadır [23, 24]. Giderek çeşitlenen bilgi kaynakları ve belgesel yapılar sorunları da aynı ölçüde karmaşıklaşmaktadır [24].

Birleşmiş Milletler Arşivler ve Belge Yönetimi Birimi (ARMS) tarafından 2006 yılında yayımlanan Belge Dijitalleştirme Rehberi'ne göre dijitalleştirme, kâğıt belge, fotoğraf, grafik malzemeler gibi fiziksel/analog materyallerin elektronik ortama ya da elektronik ortamda depolanan imajlara dönüştürmesi işlemi [26] olarak tanımlanmaktadır. Dijitalleştirme uygulamalarının temel olarak üç nedenden ötürü yapıldığı dile getirilmektedir:

- Kâğıt belge ve depolama maliyetinin azaltılması: Seçilmiş dijital belgelerin dijital ortamda depolanması kâğıt ve depolama alanı maliyetinde azaltma yarattığı gibi hayati belgelerin (vital records) korunması açısından önemlidir.
- Kurumsal İçerik Yönetimi (Enterprise Content Management) Çözümlerinin Uygulanması: Belgelerin dijitalleştirilmesi ve elektronik belgelerin kullanımının artması, kurumsal süreçlerde farklı bilgi kaynaklarının paylaşımını kolaylaştıracağı için kurumsal içerik yönetimi faaliyetleri etkin biçimde gerçekleştirilebilmektedir.
- Arşivsel Koruma: Dijitalleştirme orijinal kopyaların kullanımını azaltacağı için arşiv belgelerinin uzun süre korunmasında önemli avantajlar sağlayacağı gibi çoklu kullanım olanakları da yaratabilmektedir [26].

Dijitalleştirme kurumsal bilgi ve belge yönetimi programlarının bir parçası olarak uygulanmak durumundadır. Kurumların hâlihazırda dosyalama, bilgi güvenliği ve belge saklama vb. planlarının olması ilgili alanda dijitalleştirme uygulamalarını kolaylaştıracaktır. Bilgi ve belge kaynaklarının kanıt niteliğinin sürdürülebilmesi için elektronik kopyaların özgün (authentic), bütün (complete) ve erişilebilir (accessible) olması gerekir. Öte yandan hassas ya da gizlilik değeri olan belgelerin dijitalleştirilmesi dikkatli gerçekleştirilmelidir. Bu tür belgeler için uygun metadata, güvenlik ve erişim unsurları mutlaka tanımlanmalıdır. Bu tür belgelere sadece gerekli yetkilere sahip personelin tanımlanmış sınırlamalarla erişimi sağlanmalıdır. Gizlilik değeri taşıyan belgelerin dijitalleştirilmesi ve tanımlanması işlemleri diğer belgelerden ayrı tutulmalıdır.

Arşiv dijitalleştirme ile alakalı tanım ve bilgilerden sonra arşiv dijitalleştirmenin dünyadaki ve Türkiye'deki durumlarını incelemek gerekmektedir.

3.1.1. Arşiv Sayısallaştırma Konusunda Dünyada Durum

- 2011 yılında dijital dünyada var olan bilginin büyüklüğü 2006 yılına oranla 10 kat daha büyük olmuştur.

- Dünyanın en zengin kütüphanesinde (Kongre Kütüphanesi) 170 milyon belge var. İnternet ortamında 550 milyar belge var. Dünya üzerinde her bir kişiye 90 belge düşmektedir.
- Her yıl Kongre Kütüphanesi'ni 37 000 kez dolduracak kadar bilgi üretilmektedir. Bu bilginin %92'si manyetik ortama kayıtlıdır.
- Dünyada her yıl 2 Exabyte (100 katrilyon byte) bilgi üretiliyor (20 milyar adet The Economist dergisi).
- ABD'de yılda 80 milyar fotoğraf, 2 milyar röntgen filmi çekiliyor. Günde 610 milyar elektronik posta gönderiliyor. Dünya'da her yıl üretilen bilgi için 1,5 milyar gigabyte'lık saklama ortamı gerekiyor.

3.1.2. Arşiv Sayısallaştırma Konusunda Türkiye' de Durum

TÜİK'in 2010 yılı araştırmasına göre [25];

- Hanelerin %34'ünde masaüstü ve %17'sinde dizüstü bilgisayar var. Toplam %51. 2 evden birinde bilgisayar var.
- İnternet abonesi 2003 yılında 19 000, 2006'da 2,8 milyon ve 2010 Haziran'ında 7,7 milyon. İnternette yıllık büyüme oranı %25.
- Hanelerin %42'si İnternete ulaşabiliyor. Bu oran kentlerde %49. İnterneti dergi gazete okuma amacıyla kullanım oranı %59. Ailelerin yarısı İnterneti radyo dinleme ve TV izleme amacıyla kullanıyor.
- Cep telefonu üzerinden hizmet alanların sayısı şimdilik 1 milyon.
- 2010 yılı ortasında cep telefonu abone sayısı 61,5 milyon. 3G abone sayısı 11,4 milyon.
- Kuruluşlarda bilgisayar kullanım ve İnternet erişimine sahiplik oranları 2009 yılı itibariyle %90,7 ve %88,8. İnternet erişimine sahip girişimlerin internet sayfasına sahiplik oranı 2009 yılı Ocak ayında %58,7'dir.
- Milli Kütüphane koleksiyonunda bulunan 26 700 cilt yazma eserden yaklaşık 25.200 cildinin dijital ortama aktarılması tamamlanmıştır. Bu yazmalara ait dijital ortama aktarılan sayfa sayısı da yaklaşık 3 525 000 poza ulaşmıştır. 1100 adet sesli kaset kitaptan 387 adeti dijital ortama aktarılmıştır.

- Devlet Arşivleri Genel Müdürlüğü Cumhuriyet Arşivi'nde 9 386 457 dijital materyal bulunmaktadır.

3.1.3. Arşiv Sayısallaştırmada Türkiye'de Yaşanan Sorunlar

Arşivler sayısallaştırılırken bir takım sorunlar ile karşı karşıya gelinmektedir. Türkiye'de arşivlerin sayısallaştırılmasında karşılaşılan sorunların başlıcaları şunlardır [27];

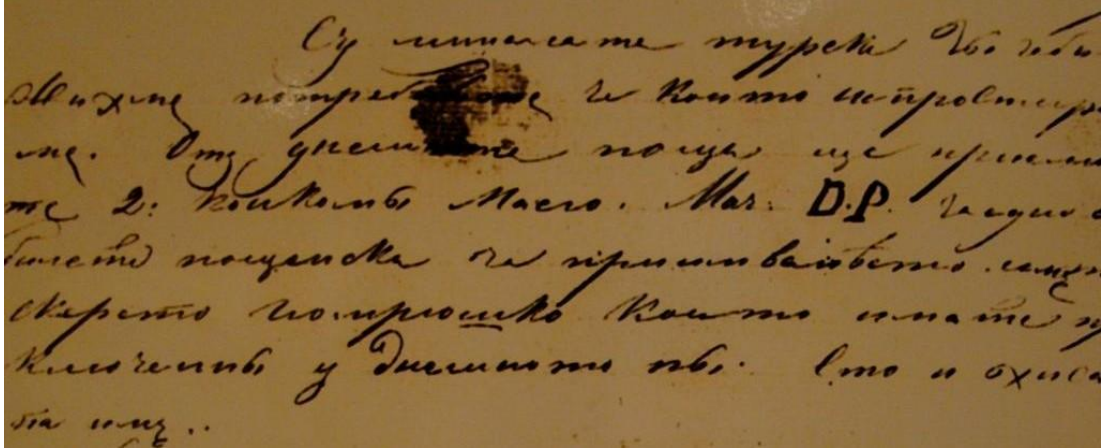
- Kurumlardaki “muhafazakâr/korumacı” tutum.
- Farkındalık eksikliği.
- “Mevzuat (yasal yapı)” ile ilgili sorunlar: 5846 sayılı Fikir ve Sanat Eserleri Kanunu.
- Telif Hakları sorunu ve korkusu.
- İşbirliği istekliliği ve örneklerinde eksiklik.(Toplu katalog vb.)
- Her kurumun kendi başına bir şeyler yapmak istemesi/ yapıyor olması. (Aynı materyalin dijitalleştirilmesi vb.)
- Bu konuya liderlik yapacak, eşgüdüm sağlayacak bir kurumun olmaması.
- Standartların eksikliği. (Üst veri, teknik şartname vb.)
- Proje temelli çabaların azlığı.
- Başarılı örneklerin yeterince paylaşılamaması.
- AB (Avrupa Birliği) perspektifinin yeterince güçlü olmaması.
- Ulusal bir politikanın belirlenmemiş olması.

3.2. reCAPTCHA Sisteminin İncelenmesi

3.2.1. reCAPTCHA

CAPTHCA'lar internet üzerindeki hizmetleri kötüye kullanan programları önleyen yaygın güvenlik önlemlerindedir. Bu yaklaşıma göre dünya üzerindeki tüm

kullanıcılar tarafından günde yaklaşık 100 milyon CAPTCHA çözülmektedir [3recapthca]. CAPTCHA'lar internet hizmetlerinin kötüye kullanılmasını önlemek için geniş ölçekte verimli olmasına rağmen, her bir insan bunları çözmek için zihinsel çaba harcamaktadır. İnsanların düşünsel kaynaklarının rastgele üretilmiş testleri çözerek boşa harcandığı fark edilmiştir. Bilgisayar çağından önce yazılmış fiziksel kitaplar, eski yazılar bilgilerin korunması ve dünya çapında erişiminin sağlanması amacıyla toptan dijitalleştirilmektedir. Sayfalar resim olarak tarandıktan sonra optik karakter okuyucu programları (OCR) tarafından metin dosyalarına dönüştürülür. Bu metinsel dönüşüm kitapların indekslenmesi, aranabilmesi ve farklı formatlarda saklanabilmesi açısından yararlı olmaktadır. Dijitalleştirme işlemindeki engellerden biri OCR programlarının taranan resimleri deşifre ederken mükemmellikten uzak olmasıdır.



Şekil 3.1 OCR Programları Tarafından Zor Çözülünecek Bir El Yazısı Metni

Şekil 3.1'e benzer eski yazılarda OCR programları kelimelerin % 20 sini ancak tanıyabilmektedir. Bununla beraber insanlar ise bu tarz işlemde daha doğru sonuçlar vermektedir. Örneğin "anahtar ve doğrulama" yöntemiyle resimleri metne aktaran iki insan tarafından kaydedilen metinler, kelime seviyesinde % 99' dan daha fazla doğru sonuç üretilmiştir [38, 2]. Ne yazık ki işleri resimleri yazıya aktarmak olan insanları kullanmak oldukça pahalıdır. Bu kişiler sadece çok önemli belgeleri aktarmakta kullanılmaktadır. ReCAPTCHA olarak adlandırılan bu yöntem şu an için 200.000 den fazla internet sitesinde kullanılmakta ve insanların rastgele üretilen karakter

resimleri çözmek yerine OCR programlarından geçememiş kelimeleri çözmelerini amaçlamaktadır. Kullanıcılar tarafından girilen sonuçlar dijitalleştirme işlemini iyileştirir. CAPTCHA gibi kullanıcıları ayırabilmek için (insan veya bilgisayar) sistemin kullanıcıların girdilerini kontrol edebilmesi gerekir. Bunun için reCAPTCHA kullanıcıya iki tane kelime verir. Bu kelimelerden bir tanesinin sonucu sistem tarafından bilinirken diğeri ise kullanıcıdan istenir. Eğer kullanıcı kontrol için verilen kelimeyi düzgün olarak girerse, diğeri kelimeyi bilinmeyen kelimeye sonuç olarak kabul eder.

3.2.2. reCAPTCHA Çalışma Sistemi

Taranmış resim iki farklı OCR program tarafından analiz edilir. OCR programlarının verdiği sonuçlar standart karakter dizisi algoritmaları tarafından eşleştirilir ve her bir sonuç bir İngilizce sözlükten araştırılır. Her iki OCR programı tarafından farklı deşifre edilen veya İngilizce sözlükte yer almayan herhangi bir kelime şüpheli olarak işaretlenir. Analizlere göre şüpheli kelimelerin % 96'si en az bir OCR programı tarafından hatalı deşifre edilmiştir. Bunun tersine şüpheli olarak işaretlenmeyen kelimelerin % 99.75 her iki OCR programı tarafından doğru deşifre edilmiştir. Bir şüpheli ve bir tane kontrol amaçlı kelime resim olarak kullanıcıya her seferinde farklı sıralama ile gösterilir. Dijitalleştirme sürecinde insanların hatalarını azaltmak için reCAPTCHA her bir şüpheli kelimeyi birden fazla kullanıcıya farklı şekillerde (farklı deformasyonlar ile) gönderir. Kullanıcı kontrol kelimesini doğru girerse bu kontrol kelimesinin bağlı olduğu bilinmeyen kelimeye ait cevabı ise olası sonuç olarak ele alınır. Birden fazla kullanıcı herhangi bir bilinmeyen kelime için ortak cevap verirse o bilinmeyen kelime kontrol kelimesi olarak gösterilmeye başlanır. ReCAPTCHA'nın bilinmeyen kelime olarak birden fazla kullanıcıya gönderdiği kelimelere girilen sonuçlar arasında fark olduğunda nihai sonuç olarak en yüksek oya sahip olanı seçer. Oylama mekanizması şu şekildedir; kullanıcıların verdiği her bir sonuç bir oy olarak sayılır ve OCR programının tahmini ise 1.5 oy olarak dikkate alınır. Dijitalleştirme süreci için bir tahminin doğru yazıma sahip bir kelime olarak seçilebilmesi için en az 2.5 oy almalıdır. Standart bir OCR programındaki doğruluk

oranı % 83.5'lere (24080 kelimededen 3976 hatalı) sahipken, reCAPTCHA sisteminde ise bu oran % 99.1'lere (24080 kelimededen 216 hatalı) ulaşmaktadır.

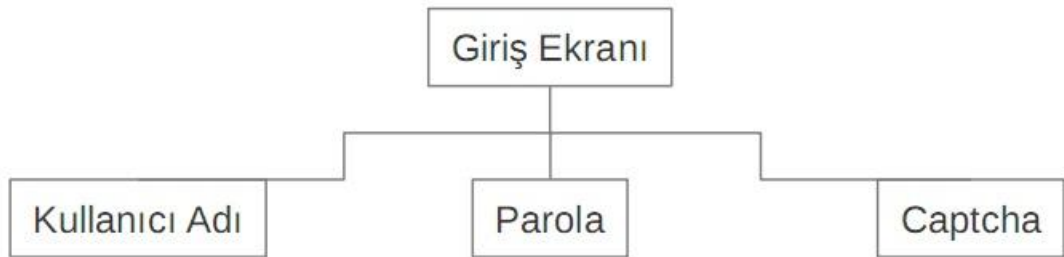
Bu sonuçlar boşa harcanan insan işlem gücünün bilgisayarların henüz çözemedikleri sorunları çözmekte sağladıkları faydanın açık bir kanıtıdır.

4. trCAPTCHA: TÜRKÇE ARŞİV SAYISALLAŞTIRMA SİSTEMİ

Luis von Ahn tarafından 2006 yılında ortaya atılan İnsan Hesaplama yöntemi tez dokümanında da detayları belirtildiği gibi başarıyı yakalamış bir yöntemdir. Bu başarıyı sağlamasında etken olan sistemlerden bir tanesi ve tez çalışması kapsamında geliştirilen uygulama olan trCAPTCHA' ya ilham kaynağı olan reCAPTCHA' dır. trCAPTCHA da reCAPTCHA gibi "İnsan Hesaplama" bölümünün alt başlıklarından bir tanesi olan teşviklerde bahsedildiği gibi güvenlik teşvikine dayanmaktadır. Ara yüz olarak da reCAPTCHA' ya benzer şekilde geliştirilmiştir. Bu benzerlik Şekil 4.1' de gösterildiği gibi trCAPTCHA'nın entegre olduğu Uniform Bilgi Sistemine girmek için gerekli olan kullanıcı adı ve şifrenin yanında resim havuzundan seçilmiş iki kelime resmi içermektedir.

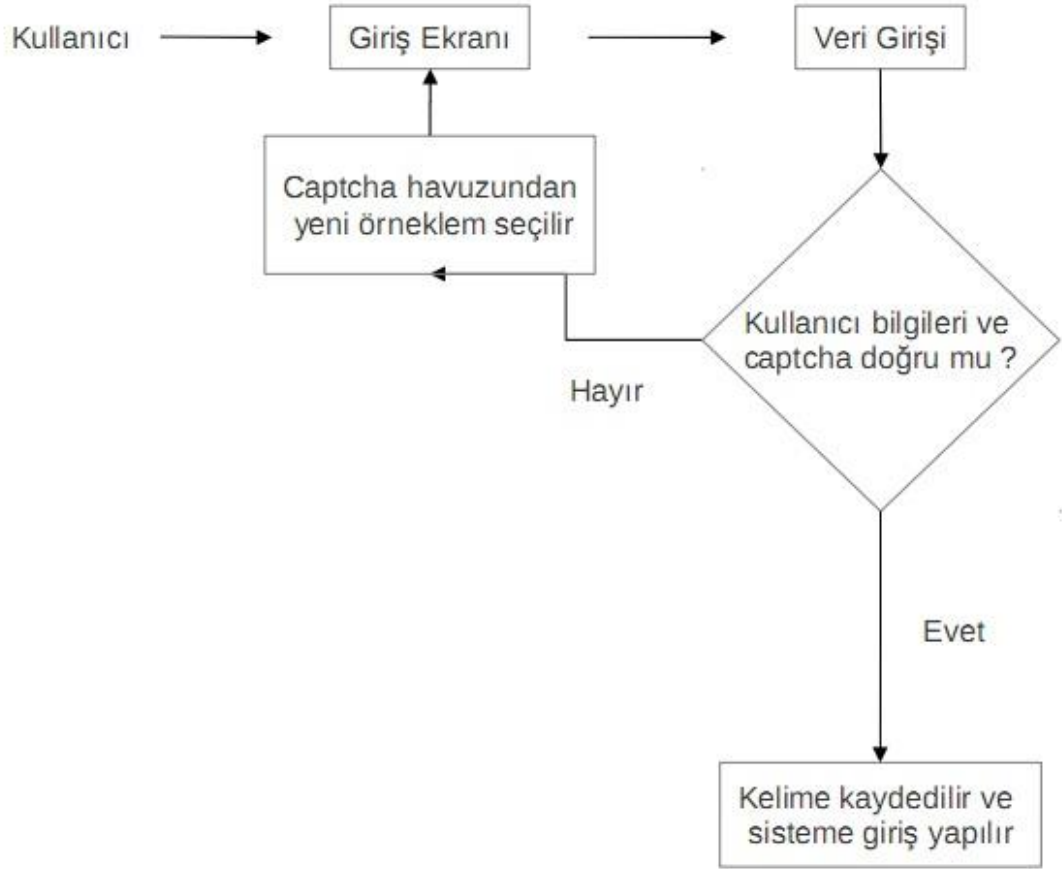
trCAPTCHA uygulamasının geliştirme aşamasında yapılan analizlerin sonuçlarında reCAPTCHA' nın eksiklikleri fark edilmiştir. Bu eksiklikler ve trCAPTCHA' nın bu eksiklikleri gidermek için kullandığı yöntemler ilerleyen kısımlarda daha detaylı olarak ele alınacaktır. Bu eksikliklerden bazıları şunlardır;

- reCAPTCHA'nın sadece İngilizce dilinde destek vermesi
- reCAPTCHA'nın metin içeriklerini sözlüğe göre kıyaslaması ve metin içeriğini bilmeden varsayım yapması.
- Sözlük ile yapılan varsayımdan kaynaklanan, kelimelerin doygunluğa süresini etkilemesi.



Şekil 4.1 Uniform Bilgi Sistemi ve trCAPTCHA Giriş Ekranı Yapısı

trCAPTCHA'nın genel işleyişi Şekil 4.2'de gösterilmektedir. Bu işleyişe göre istemde kayıtlı kullanıcı adı ve parola mevcutsa ve bu iki alan birbirleriyle uyuyorsa kullanıcı captchayı da doğru girmişse captcha'daki kelimeler kaydedilir ve sisteme giriş sağlanmış olur. Eğer aksi durum söz konusuysa yani kullanıcı adı ve parola uyuşmuyorsa kullanıcı tekrar giriş ekranına yönlendirilir.



Şekil 4.2 trCAPTCHA Akış Şeması

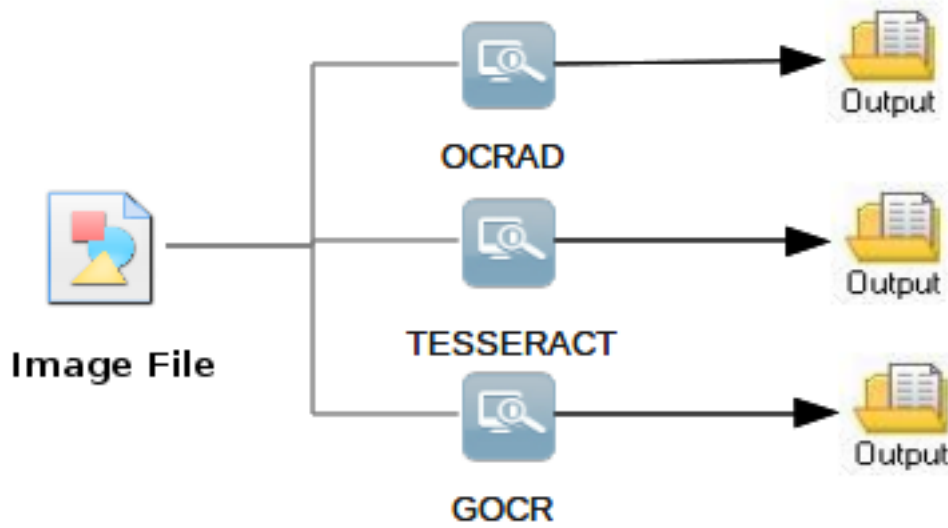
4.1. trCAPTCHA Sistem Tasarımı

Türkiyedeki arşiv dijitalleştirme ile ilgili yaşanan problemlerden ötürü türkçe tabanlı olan reCAPTCHA benzeri bir insan hesaplama sistemi tez kapsamında geliştirilmiştir.

Geliştirilen uygulama iki ana modülden oluşmaktadır. Bunlardan ilki optik karakter okuyucu uygulamalarının ve mantık kısmının işlediği çekirdek modülü, ikincisi ise kullanıcı etkileşiminin ve istemci sunucu arasındaki iletişimin sağlandığı web modülüdür.

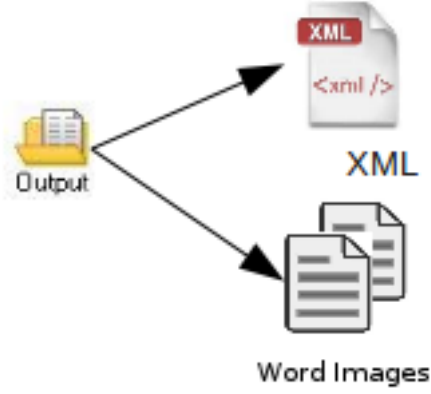
4.1.1. Çekirdek Modülü

Uygulamanın web modülünde gösterilecek verileri oluşturan temel modüldür. Bu modülde, taranmış olan resim Ocrad, Tesseract, Gocr olmak üzere üç farklı optik karakter okuyucu programından geçirilmektedir. Tez çalışması kapsamında bu programlara Optik Karakter Okuyucu Bölümünde anlatılan eklentiler yapılmıştır. Şekil 4.3’de gösterildiği gibi optik karakter okuyucu uygulamaları resmi işledikten sonra kendilerine özel çıktıları üretmektedirler.



Şekil 4.3 Çekirdek Modülü Yapısı

Şekil 4.4’de gösterildiği gibi OCR programları tarafından üretilen bu çıktıların her biri XML dosyaları ve kelime resimlerinden oluşmaktadır. Sonuçları XML dosyasında toplamamanın amacı uygulamaya sonradan optik karakter okuyucu programlarının kolaylıkla eklenebilmesini kolaylaştırmaktır.



Şekil 4.4 OCR Çıktı Yapısı

Optik karakter okuyucu programlarının ürettiği XML dosyasının yapısı Şekil 4.5’ de gösterilmektedir. XML dosyası “app” etiketi ile başlamakta, üretildiği OCR uygulamasının adını ve işlem yaptığı dokümanın adını içermektedir. “app” etiketi en az bir tane “line” etiketi barındırmaktadır. Bu etikette ise kaynaktan okunan kelimelerin kaynaktaki hangi satıra denk geldiğini ifade etmektedir. “line” etiketi de en az bir tane “word” etiketi içermek zorundadır. “word” etiketinde de satırın kaçınıcı kelimesi olduğu, bu kelime için tahmin edilen sonuç ve tahmin oranını içermektedir. Ayrıca OCR programı kelimeyi tam olarak tanımlayamadığı durumlarda alternatif kelimeleri belirtmek için “word” etiketine benzer “aword” etiketi içermektedir.

```
<?xml version="1.0" encoding="UTF-8"?>
<app name="tesseract" document="ds1_page001.tiff">
<line id="0">
<word id="0" value="Günümüzde" rate="81">
</word>
<word id="1" value="kol" rate="85">
<award value="kol" rate="65"/>
<award value="xol" rate="47"/>
</word>
<word id="2" value="gezen" rate="85">
</word>
<word id="3" value="bir" rate="96">
</word>
<word id="4" value="yanlış" rate="82">
</word>
<word id="5" value="var:" rate="88">
<award value="var;" rate="58"/>
</word>
...
</line>
...
</app>
```

Şekil 4.5 OCR Programları Tarafından Üretilen XML dosyasının İçeriği

4.1.2. Web Modülü

Geliştirilen uygulama reCAPTCHA türevi olan bir uygulama olduğundan daha fazla insanın Türkçe arşivi sayısallaştırması arzulandığından dolayı çekirdek modülden üretilen sonuçlar bir web ara yüzünde kullanıcıya gösterilmektedir. Web modülünün dört temel görevi bulunmaktadır. Bunlar;

- Optik karakter okuyucu programlarından gelen sonuçların yüklenmesi.
- Sonuçların toplu olarak görüntülenmesi ve üzerinde operasyonlar yapılması.
- Sayısallaştırma sonuçlarının görüntülenmesi
- Kullanıcılardan tahminlerin alınması

Herhangi bir optik karakter okuyucu uygulaması bir kelime için kelime doyunluk eşik değerinden (KDED) fazla bir tahmin üretirse o kelime doyunluğa ulaşmıştır denir. Doyunluğa ulaşan kelimeler optik karakter okuyucu programlar tarafından tanımlanamayan kelimeleri nitelendirmekte kullanılır.

Tez uygulamasının entegre olduğu Uniform Bilgi Sisteminin kullanıcı giriş ekranından faydalanmaktadır. Giriş ekranında kullanıcı adı ve kullanıcı şifresi haricinde bir resim gösterilmektedir. Bu resim iki tane kelime resminden oluşmaktadır. Bu kelime resimlerinden bir tanesi yukarıda bahsedilen doyunluğa ulaşmış bir kelime resmini içermektedir. Bunun amacı sistemi aldatmaya yönelik bir girişimde bulunmamasını ve doğru sonuç ürettiğini kontrol etmektir.

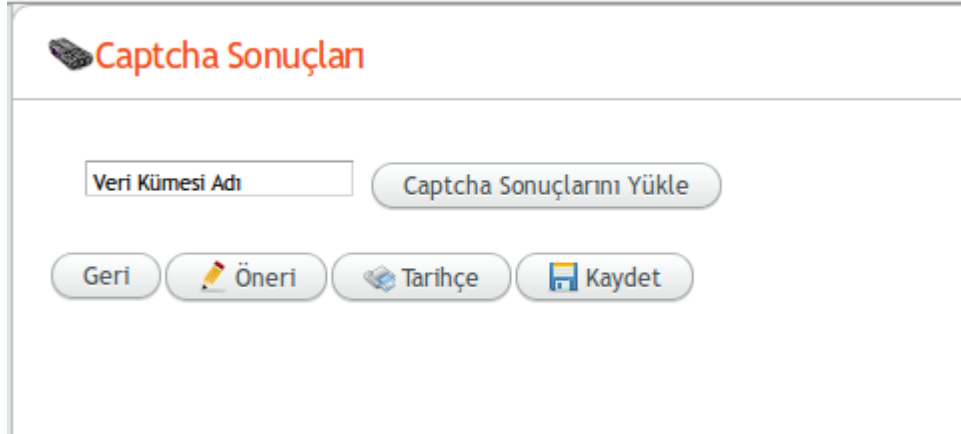
Optik karakter okuyucu programlarından üretilen kelime sonuçları herhangi bir sözlük ile karşılaştırılmamaktadır. Tez uygulaması için her ne kadar Türkçe tabanlı bir arşiv sayısallaştırma sistemi olarak tanımlansa da bu yaklaşım ile uygulamayı mümkün olduğu kadar dil bağımsız hale getiriyor. Ve optik karakter okuyucu programların bir kelime için ürettiği alternatif kelimeler ve kelime keskinlik oranları önem kazanmaktadır.

4.1.2.1. trCAPTCHA Ekranları

Tez kapsamında geliştirilen uygulamaya ait verilerin yönetilebilmesi amacıyla uygulamanın entegre olduğu Uniform Bilgi Sistemi olan ekranlar aşağıda belirtildiği gibidir.

‘trCAPTCHA Upload’ Ekranı

Optik karakter okuyucu programları tarafından oluşturulan verileri Uniform Bilgi Sistemine aktarmak için kullanılmaktadır. Şekil 4.3’de gösterilen formata sahip sıkıştırılmış dosya Şekil 4.6’de gösterilen “Captcha Sonuçlarını Yükle” düğmesi ile yüklenir. Yükleme işlemi sorunsuz bir şekilde tamamlandıktan sonra sistem, yüklenecek olan verilerin Şekil 4.7’ da gösterildiği gibi küçük bir istatistiğini gösterip yüklenecek olan veri seti isimlendirilerek kaydedilmektedir.



Şekil 4.6 OCR Sonuçlarının Yüklenmesi İçin Kullanılan Ekran



Şekil 4.7 OCR Sonuçlarının Sisteme Aktarılmadan Önce Gösterilen Ekran

Yüklenen veri setleri üzerinde çeşitli manipülasyonlar yapmak ve verilerin veya veri setlerinin anlık çözülme istatistiklerine bakmak amacıyla Şekil 4.8’de gösterilen “Captcha Resim Sorgulama” ekranı kullanılmakta ve sistemde mevcut olan verilerin resim adına ve veri setine göre sorgulama yapılabilmektedir.



Resim Adı	Dataset	Yükleme Tarihi
img001_1.jpg	DATASET-1	20.03.2012
img003_1.jpeg	DATASET-1	20.03.2012
img001.jpeg	DATASET-1	20.03.2012
img002_1.jpeg	DATASET-1	20.03.2012
img001_2.jpeg	DATASET-1	20.03.2012
img002_3.jpeg	DATASET-1	20.03.2012
img001_3.jpg	DATASET-1	20.03.2012
img005_1.jpeg	DATASET-1	20.03.2012

24 kayıt

Şekil 4.8 Sistemdeki Veri Setlerinin Sorgulandığı Ekran

Bu ekran içerisinde dört tane alt ekran barındırmaktadır. Bunlar;

‘trCAPTCHA Kelimeleri’ Ekranı

“trCAPTCHA Kelimeleri” ekranı ile Şekil 4.9’de gösterildiği gibi doyumluğa ulaşmış ve ulaşmamış kelimeler ve eğer varsa bu kelimelere ait alternatif kelimeler gösterilmektedir. Bu ekran sayesinde kullanıcılara gösterilmesi istenmeyen kelimeler düzenlenebilir, kelimelerin içerikleri ve tahmin oranları düzenlenebilir ve kelimeler manuel olarak doyumluğa ulaştırılabilir.

trCaptcha Kelimeleri

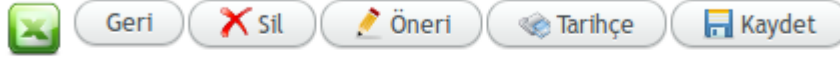


Satır Numarası 0
Kelime Numarası 3
Uygulama Adı gocr
Sonuç
Oran
Tamam
Göster
Kelime Resmi

oluyor

Alternatif Kelimeler

Tahmin	Sayaç
oluyop	94
o1uyor	66
o1uyop	60



OCR Adı	Satır No	Kelime No	Sonuç	Oran	Tamam ?
tesseract	0	0	Fakat	112	Evet
tesseract	0	1	Amiel	99	Evet
gocr	0	2	_sll	38	Hayır
gocr	0	3	oluyor	101	Evet
gocr	0	4	da	102	Evet
gocr	0	5	bu	101	Evet
tesseract	0	6	ıztırap	108	Evet
gocr	0	7	verici	100	Evet

Şekil 4.9 trCAPTCHA Kelimeleri Ekranı

‘trCAPTCHA Resim Sonuçları’ Ekranı

“trCAPTCHA Resim Sonuçları” ekranı Şekil 4.8’de gösterildiği gibi sorgulama ekranından seçilen resmin anlık istatistiğini hem görsel hem de sayısal olarak sunmaktadır. Şekil 4.10’da gösterildiği gibi ekranın “Resim Dosyası” başlıklı kısmında sonuçları gösterilen resim bulunmaktadır. “trCAPTCHA Sonuçları” başlıklı kısmında ise, ilgili resmin metinsel hali, doygunluğa ulaşmış kelimeler yeşil renkte ve doygunluğa ulaşmamış kelimeler ise kırmızı renkte olacak şekilde gösterilmektedir. “trCAPTCHA Sonuçları” başlıklı kısmının en altında ise ilgili resme ait sayısal istatistikler (Toplam Kelime Sayısı, Doygun Kelime Sayısı ve Belgenin Tamamlanma Oranı) bulunmaktadır.

trCaptcha Resim Sonuçları : img001_1.jpg Belgesi Sonuçları

UF

trCaptcha Sonuçları

Fakat Amiel ...sll oluyor da bu ızdırap verici vecd haline gelebiliyor? Filozof kendisinde «sevkita-biyyi, hamleyi ve böylece cür’et ve itimadı öldüren aşırı derecedeki teemmülün gelişmesiyle» yavaş yavaş bu hale düşüyor. Tahlil, kendini unutmamanın, bırakmanın girdiği bütün duyguları yok eder. Halbuki tahlile karşı duyulan temayül sıkılğalıktan ıstikak eder ve onunla karışır. Sıkılğan insan, tabii olarak kendi kendini müşahedeye doğru gider ve bilmukabele kendi kendini müşahede edenler de tabii olarak sıkılğan olur. İnsan duyguları üzerinde düşünmeğe kalkınca, onlardan korkar; bu duygulawa kendini kaptırmağa cesaret edemediği gibi bunları meydana çıkarmağa da cesaret edemez; bunları boğmağa veya hiç olmazsa zaptetmeğe çalışır; ger_Xten bu duyguların hi_ durdurmağa ve hafifletmeğe, nihayet mahundaolup bitenler- de bundan ibarettir. hiyetlerini değıştir-ğeme muvaffak olur. Amiel’in ru-

Toplam Kelime Sayısı : 122
Toplam Doygun Kelime Sayısı : 101
Bekleyen Kelime Sayısı : 21
Belgenin Tamamlanma Oranı : % 82

Geri Öneri Tarihçe

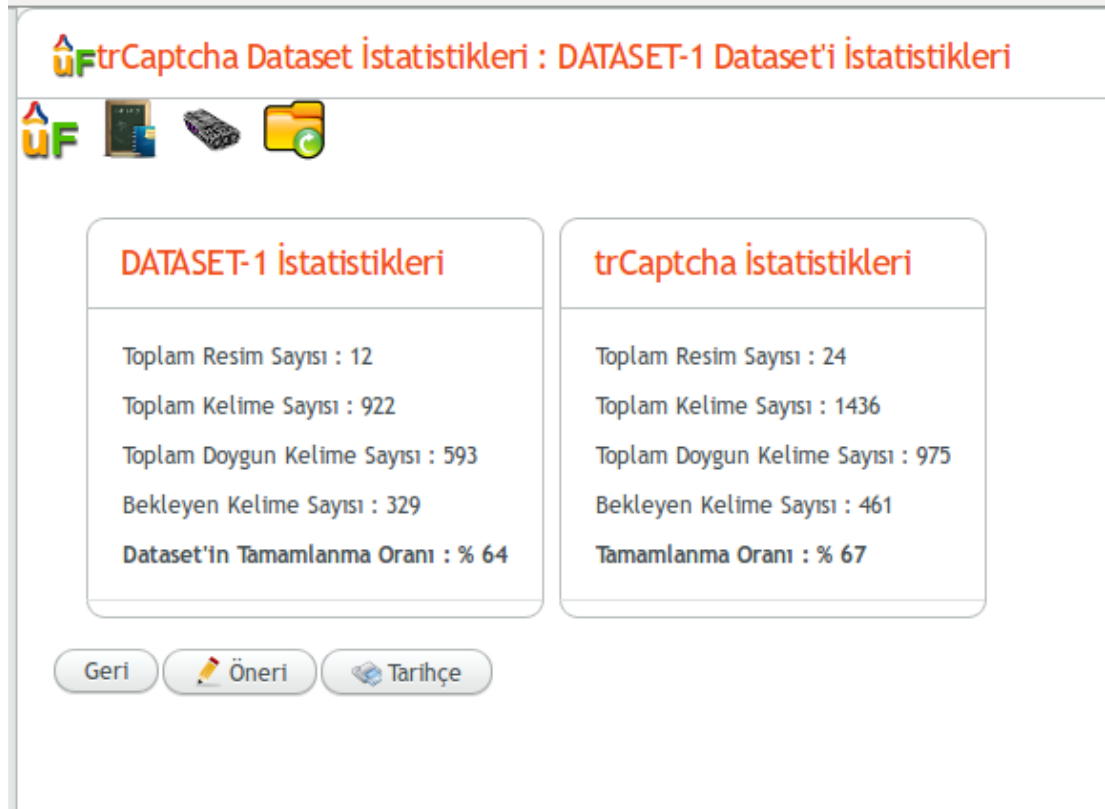
Resim Dosyası

Fakat Amiel nasıl oluyor da bu ızdırap verici vecd haline gelebiliyor? Filozof kendisinde «sevkita-biyyi, hamleyi ve böylece cür’et ve itimadı öldüren aşırı derecedeki teemmülün gelişmesiyle» yavaş yavaş bu hale düşüyor. Tahlil, kendini unutmamanın, bırakmanın girdiği bütün duyguları yok eder. Halbuki tahlile karşı duyulan temayül sıkılğalıktan ıstikak eder ve onunla karışır. Sıkılğan insan, tabii olarak kendi kendini müşahedeye doğru gider ve bilmukabele kendi kendini müşahede edenler de tabii olarak sıkılğan olur. İnsan duyguları üzerinde düşünmeğe kalkınca, onlardan korkar; bu duygulara kendini kaptırmağa cesaret edemediği gibi bunları meydana çıkarmağa da cesaret edemez; bunları boğmağa veya hiç olmazsa zaptetmeğe çalışır; gerçekten bu duyguların hızını durdurmağa ve hafifletmeğe, nihayet mahiyetlerini değıştirgeme muvaffak olur. Amiel’in ruhunda olup bitenler de bundan ibarettir.

Şekil 4.10 trCAPTCHA Resim Sonuçları Ekranı

‘trCAPTCHA Dataset İstatistikleri’ Ekranı

“trCAPTCHA Dataset İstatistikleri” ekranı ile Şekil 4.7’de gösterildiği gibi sorgulama ekranında seçilen resmin ait olduğu veri setine ait istatistikleri ve tüm captcha istatistikleri gösterilmektedir.



Şekil 4.11 trCAPTCHA Dataset İstatistikleri Ekranı

Şekil 4.11’da gösterilen ekranda istatistiksel bilgi taşıyan alanlar şunlardır;

- Toplam Resim Sayısı: Veri seti içerisinde bulunan belge resimlerinin sayısı.
- Toplam Kelime Sayısı: Veri seti içerisinde bulunan belgelerdeki kelime sayılarının toplamı.
- Toplam Doygun Kelime Sayısı: Veri seti içerisinde bulunan belgelerin doygunluğa ulaşmış kelime sayılarının toplamı.
- Bekleyen Kelime Sayısı: Veri seti içerisinde bulunan belgelerin doygunluğa ulaşmamış kelime sayısının toplamıdır.
- Dataset Tamamlanma Oranı: Veri setinin doygunluğa ulaşma oranıdır. Bu oran şu formül ile hesaplanmaktadır.

$$\text{Dataset Tamamlanma Oranı} = (\text{Doygun Kelime Sayısı} / \text{Kelime Sayısı}) * 100$$

4.2. Kullanılan Yazılım ve Teknolojiler

4.2.1. OCR

OCR, görüntüler üzerindeki el yazısı, baskı ya da makineler tarafından uygulanmış karakterleri mekanik ya da elektronik olarak çevrilmesi işini yapmaktadırlar. Grafik dosyalarının üzerindeki karakterler ve bu karakterlerin oluşturduğu metinler bu şekilde düzenlenebilir hale getirilmektedir genellikle. Karakterlerin tanınması için optik kaynaktan gelen karakter algısının çeşitli değişkenler göz önünde bulundurulması belirli yöntemlerle değerlendirilmesi gerekmektedir. Bu durum OCR' ı örüntü tanıma, yapay zekâ, bilgisayarla görme gibi alanların konusu haline getirmektedir.

Optik karakter tanıma çalışmaları ilk olarak İkinci Dünya Savaşı yıllarında başladı fakat 1950' lere kadar ticari bir girişim olmadı bu alanda. Savaş sonrası ekonomileri daha çok temel gereksinimlere yatırım yapıyordu. Fakat bu alandaki gelişmelere sebep olan şey gelişen ticari ve finansal aktiviteler oldu. Amerikan Bankacılık kuruluşu ve Finansal Hizmetler Endüstrisi banka çeklerinin otomatik olarak işlenebilmesi için ortak bir dil kullanılması çalışmasını yaptılar. Fakat daha sonra seçimleri manyetik mürekkep tanıma yönünde oldu.

OCR uygulamalarında ilk başlarda görmeyi sağlayacak kısım için ışıklar ve aynalar sabit kalıplardan görüntüyü iletecek şekilde kullanıldı. Hareketli olan diskin üzerine yansıtılan görüntü siyah ve beyaz veri olarak yansıtıldı ve parçalara bölündü. Daha sonra ise elektronik bit'lere çevrildi. Uygulamanın mantıksal kısmında ise sınırlı sayıdaki karakter seti ve siyah beyaz ayrımı ile tanıma sağlanmaya çalışıldı. Tüm bu işlemler sınırlı şartlar altında sağlanabiliyordu. Kağıt ya da tanınacak metnin olduğu yüzey belirli bir hızda, belirli doğrultuda ve belirli bir açıda daha önceden tespit edilmiş noktalara hareket ettirildiğinde başarı elde edilebiliyordu. Yani işlemin gerçekleşmesi için çok fazla sabit değişkenle ve çevresel faktörle mücadele etmek gerekiyordu.

Daha sonraki OCR uygulamalarının aletleri 1960'ların sonralarında ortaya çıktı. Televizyonlarda kullanılan CRT yani katot ışın tüpü, ufak bir ışık kaynağı ve fotoğraf çoklayıcı “curve following” (eğri takibi) adı verilen yöntem için kullanılıyordu. Bu yöntem ile verinin yeri, yazı tipindeki ve yazı kaynağındaki tasarımı, bunların yerleşimi fark etmeksizin yazı kaynağını görme sağlanmış oldu. Bu teknik belirli şartlar sağlandığında el yazısının otomatik tanınmasını sağlayan konsept olarak öne çıktı.

Üçüncü nesil tanıma aygıtları ise 1970'lerin başında foto-diyot dizileri sayesinde ortaya çıktı. Bu ufak alıcılar belirli bir düzende dizildiklerinde dokümanın baskı hızında görüntüyü yansıtabiliyorlardı. Bu görsel yaklaşımsa kızıl ötesi yaklaşımları ortaya çıkardı.

Tez çalışması sırasında üç adet optik karakter okuyucu programı incelenip kullanılmıştır. Bu optik karakter okuyucu programları aşağıda detaylı bir şekilde incelenmiştir.

4.2.1.1. GOCR

GOCR projesi Joerg Schulenburg tarafından 2000 yılında GPL altında geliştirilmeye başlanmıştır [29]. GOCR, C programlama dili kullanılarak geliştirilmiştir. Taranmış resimleri (yazı içeren) text formatına çeviren GOCR, ASCII/UTF-Text, htm/xml, Tex formatlarında çıktılar sağlamaktadır. Başlıca özellikleri şunlardır [29]:

- Taşınabilirlik.
- Font büyüklükleri aynı olacak. Farklı font büyüklüklerinde sonuçlar kötüleşiyor, karakter tanıma güçleşiyor. İyi sonuçlar 20-60 piksel arasındaki fontlarda sağlanıyor.
- Renkli resimler ile çalışmıyor. (Sadece gri tonlarındaki ve siyah/beyaz resimlerle çalışıyor.)
- Eğitilemez. (İstenildiğinde verdiği sonuçları kaydedip kontrolü arttırabilir. Fakat bu sonuçların düzgün olması için sonuçları % 100 bilmesi gerekiyor.)

- Otomatik algılamaya sahip (İstenildiğinde kapatılabilir.)
- Bir Boyutlu Barkod (1D Barcode) okuyabilir.
- Her karakter kendi piksel örüntüsüne göre tanınır.
- Yavaş.
- Temiz (gürültü içermeyen) ve beyaz zemin üzerine siyah yazılı ve yüksek çözünürlükteki resimlere ihtiyaç duymaktadır.
- Döndürülmemiş (çevrilmemiş) resimler.
- Basit karakter yerleşimi ve resmin sadece yazı içermesi.
- Sadece latin fontları çevirebilir. Bazı eklentilerle Almanca, İspanyolca dillerine özgü karakterleri de tanımlayabilir.
- Resim tipi olarak PNM, PGM, PBM formatlarını kabul eder. Fakat .jpg, .jpeg, .bmp, .tiff, .png, .ps, .eps formatlarını da kendi içerisinde .pgm uzantılı dosyalara çevirip kabul eder.

GOOCR'da Yapılan Değişiklikler

Tez çalışması sırasında bir takım ihtiyaçları gidermek için GOOCR projesine eklemelerde bulunulmuştur [29]. GOOCR projesinin standart halinde, sonuçlar karakter bazlı verilmektedir. Sonuçları kelime bazlı üretecek şekilde dönüştürüldü. Karakter tanıma sırasında GOOCR'ın emin olamadığı karakterlerde, bu karakterlere ait alternatifler verilmekteydi. Bu alternatif karakterler alternatif kelimelere dönüştürüldü. Kelimeleri resim halinde, orijinal resimden çıkartma işlemi eklendi. Karakter bazlı keskinlik oranları kelime bazlı olacak şekilde değiştirildi. Bu değişiklikler neticesinde üretilen sonuçları XML formatında saklayabilme özelliği eklendi.

4.2.1.2. TESSERACT

Tesseract projesi 1985 de HP tarafından HP-UX makineleri için geliştirilmeye başlanmıştır. 2005 yılında Apache lisansı altında açık kaynak kodlu proje haline getirildi [28]. Tesseract C++ programlama dilini kullanarak geliştirilmiştir. Taranmış resimleri text formatına dönüştürmektedir. Başlıca özellikleri şunlardır [28]:

- Taşınabilir.
- Sayfa yerleşim analizi bulunmaktadır. (Kısıtlı bir şekilde)
- Sonuç çıktılarını txt, hocr/html formatlarında verebilmektedir.
- Arayüzü bulunmamaktadır. Çalıştırma konsoldan yapılmaktadır.
- Tek kolonlu, yazı içeren resimleri yazı formatına çevirir.
- Kendi içerisinde basit bir sözlüğü vardır. Sözlük kelimelerini DAWG halinde tutmaktadır.
- Eğitilebilir. Sonuçları düzenleme özelliğini sağlayan yan programları bulunmaktadır.
- Ondan fazla (10+) dil desteğine sahip. Fakat otomatik dil algılama mevcut değildir. Çalıştırılırken taranmış resmin hangi dilde olduğunu açıkça belirtmek gerekmektedir.
- Kelime bazlı sonuçlar verebiliyor.
- Taranmış resmi TIFF formatında kabul ediyor.
- Kütüphane desteği var. Farklı programlama dillerinde bindingleri mevcut.

TESSERACT Çalışma Prensipleri

TESSERACT optik karakter okuyucu programı çalışma esnasında geçirdiği evreler şunlardır;

- Resim girdisini alır.

- Resim için eşik değerini (threshold value) hesaplar ve bu değere göre resmin iyileştirilmesini yapar.
- Sayfa yerleşim analizini yapar.
- Yazı içeren bölgelerdeki bileşenleri oluşturur.
- Karakter bölgelerini belirler.
- Satırları ve kelimelerin yerleşimini yaparak bulur.
- Bulduğu satırların kelimelerini iki aşamalı tanımadan (Recognize Word Pass 1 ve Recognize Word Pass 2) geçirir.
- Tanıma işleminden sonra çıktı formatını hazırlar ve çıktıyı oluşturur.

TESSERACT'da Yapılan Değişiklikler

Tez çalışması sırasında bir takım ihtiyaçları gidermek için TESSERACT projesine eklemelerde bulunulmuştur [28]. TESSERACT projesinin standart halinde, sonuçlar kelime bazlı verilmektedir. Fakat emin olamadığı kelimelerin alternatiflerini vermesi sağlandı. Kelimeleri resim halinde, orijinal resimden çıkartma işleminde mevcut olan problemler giderildi. Alternatif kelimelerin keskinlik oranları elde edildi. Bu değişiklikler neticesinde üretilen sonuçları XML formatında saklayabilme özelliği eklendi.

4.2.1.3. OCRAD

OCRAD projesi Antonio Diaz tarafından 2004 yılında GPL altında geliştirilmeye başlanmıştır [30]. OCRAD C++ programlama dili kullanılarak geliştirilmektedir. Taranmış resimleri text formatına dönüştürmektedir. OCRAD hala devam eden bir araştırma projesidir. Başlıca özellikleri şunlardır [30]:

- Sadece gri tonlarındaki ve siyah/beyaz resimlerle çalışmaktadır.
- Eğitilemez.
- Yerleşim analizi bulunmaktadır. (Taranmış resimlerdeki blokları, kolonları ayırmaya yarar.)
- Öznitelik çıkarımına göre tanınıyor (Karakterleri şekillerine göre tanıyor.).

- Hızlı. Hızlı olmasının nedeni her bir karakterin şeklini karakter şekillerinin bulunduğu bir veri tabanından karşılaştırmamasıdır.
- Temiz (gürültü içermeyen) ve beyaz zemin üzerine siyah yazılı ve yüksek çözünürlükteki resimlere ihtiyaç duymaktadır.
- Döndürülmemiş (çevrilmemiş) resimler.
- Basit Karakter yerleşimi ve resmin sadece yazı içermesi.
- Resim tipi olarak PNM, PGM, PBM formatlarını kabul eder.
- Kütüphane desteği bulunmaktadır.
- Araştırma projesi olduğundan dolayı karakter tanıma için kullanılan algoritma sürümden sürüme değişiklik gösterir.

OCRAD Çalışma Prensipleri

OCRAD optik karakter okuyucu programı çalışma esnasında geçirdiği evreler şunlardır;

- Resim girdisini alır.
- Resim için eşik değerini (threshold value) hesaplar ve bu değere göre resmin iyileştirilmesini yapar.
- Daha iyi sonuçlar alabilmek için resim üzerinde bir takım dönüşümler (kesme, döndürme, ölçeklendirme... vb.) uygular.
- Yerleşim analizi işlemini gerçekleştirir.
- Taranmış resimdeki yazı içermeyen bileşenleri (frameletler, resimler... vb.) temizler.
- Karakterleri algılar.
- Algılanan bu karakterleri satırlar halinde gruplandırır.
- Karakter tanıma işlemini gerçekleştirir. (Her karakter için bir algoritma.)
- Hataları düzeltir. (1000 şeklindeki sonuçları 1000 haline getirilmesi.)
- Tanıma işleminden sonra çıktı formatını hazırlar ve çıktıyı oluşturur.

OCRAD'da Yapılan Değişiklikler

Tez çalışması sırasında bir takım ihtiyaçları gidermek için OCRAD projesine eklemelerde bulunulmuştur [30]. OCRAD projesinin standart halinde, sonuçlar karakter bazlı verilmektedir. Bu sonuçlar kelime bazlı olarak değiştirildi ve emin olmadığı kelimelerin alternatiflerinin verilmesi sağlandı.

OCRAD, karakterlerin keskinlik oranlarını veremediğinden dolayı asıl kelimelerin ve alternatif kelimelerin keskinlikleri hesaplanamadı. Bundan dolayı genel yapıyı etkilememek adına şöyle bir yöntem kullanıldı: Eğer bir kelimenin alternatifi yok ise bu kelime için %70, alternatif kelimelerin keskinlik oranlarını ise %65 olarak belirlendi. Kelimeleri resim halinde, orijinal resimden çıkartma işlemi eklendi. Bu değişiklikler neticesinde üretilen sonuçları XML formatında saklayabilme özelliği eklendi.

4.2.1.4. TESSERACT, OCRAD ve GOCR Karşılaştırılması

Tez uygulamasında kullanılan optik karakter okuyucu programlarının basit resimler üzerindeki tanıma oranlarının farkları incelenmiştir. Resmi farklı yazı tiplerine ve gri tonlarına çevirerek bir veri seti oluşturulmuştur. Tüm resimler şu özelliklere sahiptir;

- 300 DPI çözünürlükte olması.
- Hiç bozulmamış olması veya yazıların tek kolon olarak düzenlenmesi.
- Dil olarak İngilizce seçilmesi.
- Sadece ASCII-7 karakterleri içermesi.
- Hiç gürültü içermemesi.

Resimde bulunan orijinal yazı şu şekildedir:

“And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas
just above my chamber door;
And his eyes have all the seeming of a demon's that is
dreaming,
And the lamp-light o'er him streaming
throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted –
nevermore!”

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.12 Yazı Tipi Courier, Yazı Rengi Siyah Olan Metin Resmi

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.13 Yazı Tipi Courier, Yazı Rengi Gri Olan Metin Resmi

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.14 Yazı Tipi Justy, Yazı Rengi Siyah Olan Metin Resmi

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.15 Yazı Tipi Justy, Yazı Rengi Gri Olan Metin Resmi

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.16 Yazı Tipi Times, Yazı Rengi Siyah Olan Metin Resmi

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.17 Yazı Tipi Times, Yazı Rengi Gri Olan Metin Resmi

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.18 Yazı Tipi Verdana, Yazı Rengi Siyah Olan Metin Resmi

And the raven, never flitting, still is sitting, still is sitting
On the pallid bust of Pallas just above my chamber door;
And his eyes have all the seeming of a demon's that is dreaming,
And the lamp-light o'er him streaming throws his shadow on the floor;
And my soul from out that shadow that lies floating on the floor
Shall be lifted - nevermore!

Şekil 4.19 Yazı Tipi Verdana, Yazı Rengi Gri Olan Metin Resmi

Veri setindeki resimlerin tez çalışmasında kullanılan optik karakter okuyucu programlarından geçirildikten sonraki sonuçları Çizelge 4.1’de gösterildiği gibidir. OCR programlarının ürettiği sonuçları dwdiff programının istatistik çıktılarını orjinal yazı ile karşılaştırılarak belirlenmiştir. (Not: dwdiff programı Linux işletim sistemi üzerinde çalışır ve karşılaştırma işlemini satırlar yerine kelime bazlı yapar.) OCR programları tek başlarına yeteri kadar iyi sonuç veremediği için tez çalışmasında birden fazla OCR programı kullanılarak gerçekleştirim sağlanmıştır.

Çizelge 4.1 OCR Programlarının veri setine göre sonuçları

	GOCR	OCRAD	TESSERACT
Şekil 4.12	<i>% 67 (0.09s)</i>	<i>% 21 (0.02s)</i>	<i>% 81 (0.63s)</i>
Şekil 4.13	<i>% 67 (0.09s)</i>	<i>% 21 (0.03s)</i>	<i>% 81 (0.63s)</i>
Şekil 4.14	<i>% 31 (0.11s)</i>	<i>% 1 (0.02s)</i>	<i>% 15 (0.61s)</i>
Şekil 4.15	<i>% 31 (0.10s)</i>	<i>% 1 (0.02s)</i>	<i>% 15 (0.60s)</i>
Şekil 4.16	<i>% 76 (0.16s)</i>	<i>% 82 (0.03s)</i>	<i>% 92 (0.74s)</i>
Şekil 4.17	<i>% 76 (0.16s)</i>	<i>% 82 (0.03s)</i>	<i>% 92 (0.74s)</i>
Şekil 4.18	<i>% 98 (0.10s)</i>	<i>% 98 (0.02s)</i>	<i>% 98 (0.45s)</i>
Şekil 4.19	<i>% 98 (0.10s)</i>	<i>% 98 (0.02s)</i>	<i>% 98 (0.46s)</i>

4.2.2. ORACLE

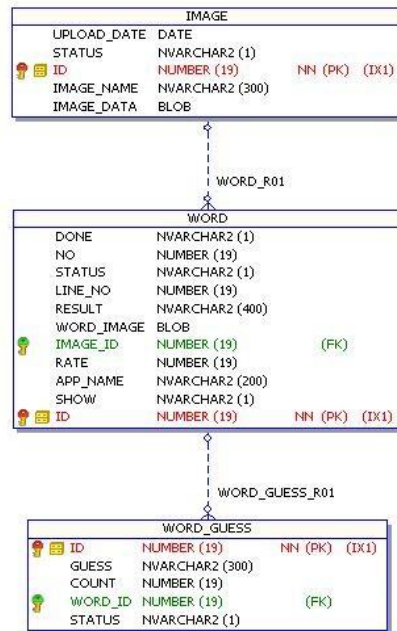
İlişkisel Veritabanı Yönetim Sistemleri (RDBMS) büyük miktarlardaki verilerin güvenli bir şekilde tutulabildiği, bilgilere hızlı erişim imkanlarının sağlandığı, bilgilerin bütünlük içerisinde tutulabildiği ve birden fazla kullanıcıya aynı anda bilgiye erişim imkanının sağlandığı yazılımlardır. Oracle veritabanı da bir ilişkisel veri tabanı yönetim sistemidir. Oracle veritabanının özellikleri şunlardır:

- Büyük miktarda veri tutabilmekte ve verilerin depolandığı alanları ayarlama imkanı vermektedir.
- Aynı anda çok sayıda kullanıcıya verilerin bütünlüğünü bozmadan hizmet verebilmektedir.
- İşletim sistemi, veri erişim dilleri ve ağ iletişim protokolleri standartlarıyla uyumludur. Hiç kapatılmadan çalışabilmektedir.
- Yetkisiz erişimleri engelleme ve kontrol edebilme imkanı sağlamaktadır.
- Bütünlüğü veritabanı düzeyinde sağlayabilmektedir, böylece daha az kod yazılmaktadır.
- İstemci/Sunucu mimarisinin bütün avantajlarını kullanabilmektedir.

ORACLE Kullanım Alanı

Çekirdek modülü tarafından oluşturulan sonuçların depolanmasında kullanılmıştır. Depolama işlemini gerçekleştirebilmek için uygulamanın kullandığı ilişkisel tablolar Çizelge 4.2’ de gösterilmiştir.

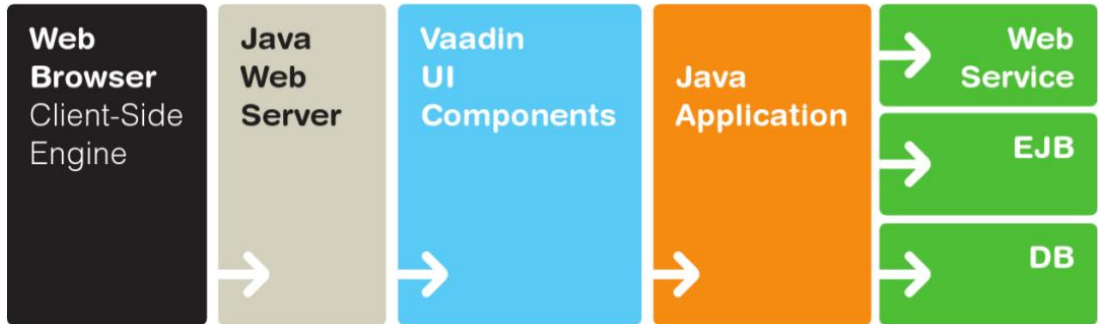
Çizelge 4.2 Veri Tabanı Şeması



4.2.3. VAADIN

Vaadin geliştiricilere Java ile yüksek kalitede kullanıcı ara yüzleri oluşturmasını sağlayan sunucu taraflı AJAX web uygulaması geliştirme çatısıdır. Vaadin uygulama çatısının çekirdeğini Java Kütüphanesi oluşturmaktadır. Bu sayede yüksek kalitedeki web tabanlı kullanıcı arayüzlerinin oluşturulması ve bakımını kolaylaştırır. Vaadin'in sunucu odaklı programlama modeli, geliştirme sırasında programcının web tarafını unutup AWT, Swing veya SWT gibi araçları kullanarak masaüstü uygulaması geliştirir gibi arayüz geliştirmeyi hedeflemektedir.

Geleneksel web programlama yeni web teknolojilerini öğrenmek için eğlenceli bir yol olsa da daha verimli olmak ve uygulama mantığı üzerine odaklanmak istenebilir. Sunucu odaklı programlama modeli ile Vaadin kullanıcı arayüzü yani tarayıcı ile sunucu arasındaki AJAX iletişimini yönetmektedir. Vaadin'in yaklaşımı ile HTML veya JavaScript gibi tarayıcı teknolojilerini öğrenmek zorunda kalınmamaktadır.



Şekil 4.20 Vaadin Genel Mimarisi

Şekil 4.20 Vaadin ile geliştirilmiş web uygulamalarının basit mimarisini göstermektedir. Vaadin sunucu taraflı çatı ve tarayıcı üzerinde JavaScript programı gibi çalışan, kullanıcı arayüzünü görüntülemeyi sağlayan, kullanıcıların işlemlerini sunucuya ileten istemci taraflı bir motordan oluşmaktadır. Uygulama Java uygulama sunucusunda Java Servlet oturumu gibi çalışır. İstemci taraflı motor, tarayıcıda JavaScript gibi çalıştırıldığından dolayı Vaadin kullanılarak geliştirilmiş uygulamaların kullanılması için herhangi bir tarayıcı eklentisine ihtiyaç duyulmamaktadır. Bu Flash, Java Appletleri veya diğer eklentilere ihtiyaç duyan

diğer uygulama çatılarına göre büyük kolaylık sağlamaktadır. Vaadin GWT desteğine dayandığı için geniş çapta tarayıcı desteği sağladığından dolayı geliştiricilerin tarayıcı desteği konusunda endişelenmelerine gerek kalmamaktadır.

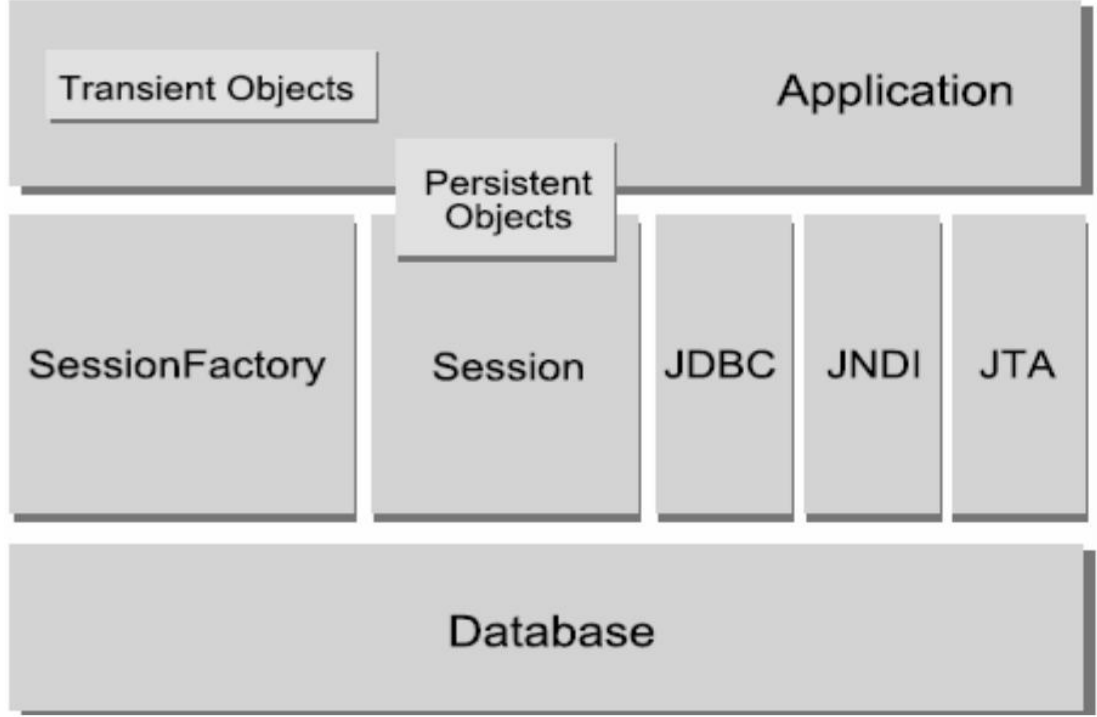
GWT programları Java kullanılarak yazılır fakat JavaScript olarak derlenir. Bundan dolayı geliştiricilerin JavaScript veya diğer tarayıcı teknolojilerini öğrenmesi zorunluluğu ortadan kalkar. Vaadin genişletilebilir şekilde tasarlanmıştır. Ve hatta Vaadin tarafından sunulan bileşenlere ek olarak herhangi bir üçüncü parti GWT bileşenleri kolayca kullanılabilir. GWT kullanılması yazılacak olan kodların saf Java ile yazılması gerektiği anlamına da gelmektedir.

VAADIN Kullanım Alanı

Tez kapsamında geliştirilen uygulamanın web tarafındaki kullanıcı arayüzlerinin oluşturulmasında kullanılmıştır.

4.2.4. HIBERNATE

Hibernate bir nesne/ilişkisel eşleme (ORM) aracıdır. Sadece Java sınıfları ve veritabanı tabloları arasında dönüşüm yapmamaktadır. Bu dönüşümlerle birlikte veri sorgulama (data query) ve veri çekme (data retrieval) işlemlerini de sağlamaktadır. Hibernate kullanımı olmadan tüm bu işlemler için SQL ve JDBC'nin olanaklarından faydalanılarak el ile (manual) veri işleme (data handling) gerçekleştirilmektedir. Bunların yanı sıra güçlü bir sorgulama dili (HQL) desteğiyle nesnelere veri tabanından çekmeyi, önbelleğe alma (caching) işlerini de etkin bir şekilde gerçekleştirmektedir.



Şekil 4.21 Hibernate Detaylı Mimarisi

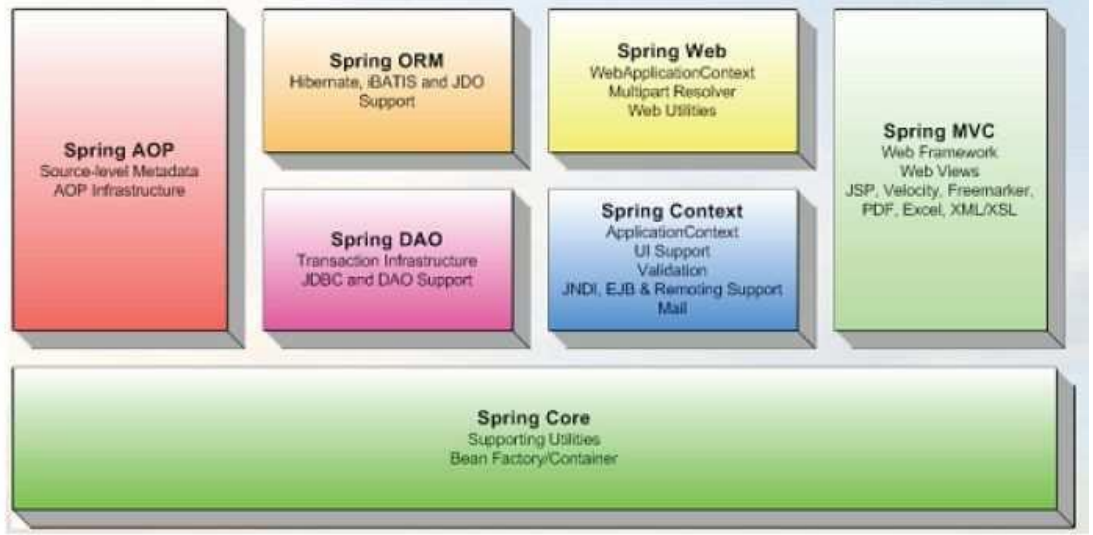
HIBERNATE Kullanım Alanı

Çizelge 4.21’de gösterilen ilişkisel tabloları uygulama tarafında modellemekte kullanılmıştır. İlgili model yardımı ile sistemde bulunan tablolara CRUD (Create, read, update and delete) işlemlerini üstlenmesi, ön bellekleme (caching) gibi özelliklerinden faydalanmak hedeflenmiştir.

4.2.5. SPRING

Spring, J2EE (Java 2 Enterprise Edition) uygulamalarını geliştirme süreçlerine alternatif bir uygulama geliştirme çatısı olarak kullanılmaktadır. Uygulama geliştirmeyi ve denetimi zorlaştıran ağır (heavyweight) çözümlere alternatif olarak hafif (lightweight) bir yapı sunmaktadır. Annotation güdümlü yapılandırma ve genişletilebilir XML yazmaya olanak sağlaması da Spring kullanımını kolaylaştırmaktadır. Java platformunun kullanımı, sunmuş olduğu detaylı API’lerden dolayı zorlaşmışken Spring kendi API’leri ile bu sorunu çözmektedir. Spring birbirinden bağımsız olan ve ihtiyaca göre kullanılacak birçok özellik

içermektedir. Bu özellikler Spring'i oluşturan yedi birim tarafından sağlanmaktadır. Bu birimler Şekil 4.22'de gösterilmektedir.



Şekil 4.22 SPRING Mimarisi

SPRING Kullanım Alanı

Geliştirilen sistem içerisindeki üçüncü parti yazılımların orkestrasyonunu sağlamaktadır.

5. DENEYSEL PERFORMANS ANALİZİ

İnsan hesaplama tekniklerine dayalı sistemlerin temel amacı her bir girdi (input) için ortak kanı (common sense) elde etmektir. Bu amaca ulaşabilmek için genel olarak kabul edilen düşünce geliştirilen insan hesaplama sistemlerinin belirli bir strateji ile tasarlanması ve bu sistemlerden maksimum oranda verim alınmasını sağlayacak etmenlerin dâhil edilmesidir.

5.1. Kullanılan Veri Kümelerinin Oluşturulması

Tez çalışması için geliştirilen uygulamanın temel amacı sisteme girmeye çalışanların insanlar olduğunu garanti etmeye çalışırken diğer yandan da arşiv sayısallaştırma işlemini gerçekleştirmektir.

Bu kapsamda örneklem alınacak havuz, sayısal hali bulunmayan, basım tarihi eski olan kitaplardan oluşmaktadır. Bu koşuldan dolayı TOBB ETÜ kütüphanesinde bulunan “Edebiyat ve Sanatta Sıkılğanlar” ve “Yaşama Felsefesi” adlı kitaplar kullanılmıştır.

Veri kümeleri oluşturulurken tarama işlemi sırasında kitaplara zarar vermemek için tam tarama yapılamadığından dolayı taranmış bölgelerin sorunsuz bölümleri seçilmiştir. Bunlardan iki tane veri kümesi oluşturulmuştur.

Veri Kümesi 1

Bu veri kümesi, “Edebiyat ve Sanatta Sıkılğanlar” kitabından tarayıcı vasıtası ile elde edilmiş on iki adet resimden oluşmaktadır. Bu resimlerden rastgele seçilmiş bir tanesi Şekil 5.1’ de gösterilmektedir. Şekil 5.1’ de gösterilen resmin çekirdek modülü tarafından oluşturulan sonuçları Çizelge 5.1’ de gösterilmiştir.

Edebiyatımızda küçük hikâye yazarı olarak önemli bir yer alan Ömer Seyfettin'in incelenmesine geçmeden önce, ona gelinceye kadar, edebiyatımızda küçük hikâyenin durumunu kısaca gözden geçirelim.

Şekil 5.1 Veri Kümesi 1'den Seçilen Rastgele Bir Resim

Çizelge 5.1 Şekil 5.1'nin OCR Tarafından Üretilen Sonuçları

Resim/OCR	GOOCR	OCRAD	TESSERACT
	Tahmin	Tahmin	Tahmin
Edebiyatımızda	E_ebi_____ (Oran : 66)	E_ebi_____ (Oran : 65)	Edebiyatımızda (Oran : 74)
küçük	_8_____ (Oran : 48)	_8_____ (Oran : 65)	küçük (Oran : 61)
hikâye	h_____ (Oran : 31)	h_____ (Oran : 65)	hikâye (Oran : 65)
yazarı	ya_____ (Oran : 46)	ya_____ (Oran : 65)	yazan (Oran : 75)
olarak	olarak (Oran : 98)	olarak (Oran : 65)	olarak (Oran : 69)

Veri Kümesi 2

Bu veri kümesi, “Yaşama Felsefesi” kitabından tarayıcı vasıtası ile elde edilmiş on iki adet resimden oluşmaktadır. Bu resimlerden rastgele seçilmiş bir tanesi Şekil 5.2’de gösterilmektedir. Gösterilen resmin çekirdek modülü tarafından oluşturulan sonuçlar Çizelge 5.2’de gösterilmiştir.

“Herkes seni, bizzat kendisi kadar tanır, Efruzcuğum! Bugün hiç kimse sana yabancı değildir, çünkü sen “hepimiz” değilsen bile ‘hepimizden’ bir parçasın.”

Şekil 5.2 Veri Kümesi 2’den Seçilen Rastgele Bir Resim

Çizelge 5.2 Şekil 5.2’nin OCR Tarafından Üretilen Sonuçları

Resim/OCR	GOCR	OCRAD	TESSERACT
	Tahmin	Tahmin	Tahmin
“Herkes	"_er___ (Oran : 62)	"er___ (Oran : 65)	"Herke_ (Oran : 72)
seni,	Be_i, (Oran : 65)	Be_i, (Oran : 65)	seni, (Oran : 75)
bizzat	b___t (Oran : 76)	b___t (Oran : 65)	Bizzat (Oran : 69)
kendisi	_end_i (Oran : 81)	_end_i (Oran : 65)	Kendisi (Oran : 78)
kadar	_d__ (Oran : 33)	_d__ (Oran : 65)	Kadar (Oran : 70)

5.2. Performans Unsurları

Tez çalışması kapsamında geliştirilen insan hesaplama sisteminde performansı etkileyecek bazı unsurlar ve bu unsurların sisteme olan etkileri aşağıda belirtilmiştir.

5.2.1. Resmin Bozulması

Karakterlerin veya kelimelerin şekillerinin bozulmaya uğratılmasının, kelime resimlerinin kullanılabilirliği üzerinde belirgin bir etkisi bulunmaktadır. Kelimeler üzerinde bozulma şu şekillerde sağlanabilmektedir. Bunlar;

- Çevirme (*Translation*) : Kelime resmindeki karakterlerin belirli bir oranda sağa, sola, yukarıya veya aşağıya doğru hareket ettirilerek sağlanmaktadır.
- Döndürme (*Rotation*) : Kelime resmindeki karakterleri saat yönünde veya aksi yönde döndürülerek sağlanmaktadır.
- Ölçeklendirme (*Scaling*) : Kelime resminde bulunan karakterlerin x veya y yönünde esnetilerek veya sıkıştırılarak sağlanmaktadır.
- Sarmalama (*Wrap*) : Farklı ölçekteki kelime resimlerinin elastik deformasyona uğratılarak sağlanmaktadır.

Kelime resimlerine bu yöntemlerden herhangi birinin uygulanması, kelime karakterlerinin karıştırılması (5 yerine S veya s harflerinin girilebilmesi) gibi bir takım problemleri beraberinde getirmektedir. Bozulmadan kaynaklanan problemler ile başa çıkabilmek için her kullanıcıya birden fazla tahmin yapabilmeye olanağı sağlanmalıdır. Geliştirilen insan hesaplama sistemine ait veri kümelerinin yapısı ve sistem entegre noktasının aynı resim için sadece bir tahmine izin vermesinden dolayı kelime resimlerinde bozulma yönteminin kullanılması sistemden elde edilecek verime olumsuz yönde etki sağlamaktadır.

5.2.2. Resim Boyutları

Kullanıcılara çözümlenmeleri için gösterilen kelime resimlerinin boyutlarının performans üzerinde önemli etkisi bulunmaktadır. Kelime resmindeki ayrıntılar ile boyutu arasında doğru orantı bulunmaktadır. Resmin boyutu ne kadar artarsa resimdeki ayrıntılar o oranda ortaya çıkmaktadır. Ayrıntıları yeterince belli olmayan kelime resimleri kullanıcılar tarafından yanlış tahminlerin oluşturulmasına yol açmaktadır. Uygulamada resim boyutlarından kaynaklanan olumlu etkileri kullanabilmek için gösterilen resimlerin boyutları mümkün olduğunca büyük kullanılmıştır. Şekil 5.3 ve Şekil 5.4’ de resim boyutunun sağladığı avantaj gösterilmektedir.



Şekil 5.3 Standart Kelime Resmi



Şekil 5.4 Boyutları Değiştirilmiş Kelime Resmi

5.2.3. Sonuçların Sınıflandırılarak Seçilmesi

Çekirdek modül tarafından oluşturulan sonuçlar, her bir optik karakter okuyucu uygulamasının kelime resimlerine ait tahminlerini ve bu tahminlere ilişkin tahmin oranlarını içermektedir. Bu oranlar 0 ile 100 arasındadır.

Kelime doygunluk eşik değeri (KDED) altında kalan kelimeler, doygunluğa ulaşmış kelimeler ile birlikte kullanıcılara gösterilerek çözümlenmeleri sağlanmaktadır. Bu çözümleme işlemi kelimelerin kelime doygunluk eşik değerine ulaşmaya kadar devam etmektedir. Çekirdek modül tarafından oluşturulan veri kümelerine ait elde edilen sonuçlar Çizelge 5.3’ de gösterilmiştir.

Çizelge 5.3 Veri Kümelerine Ait Sonuçlar

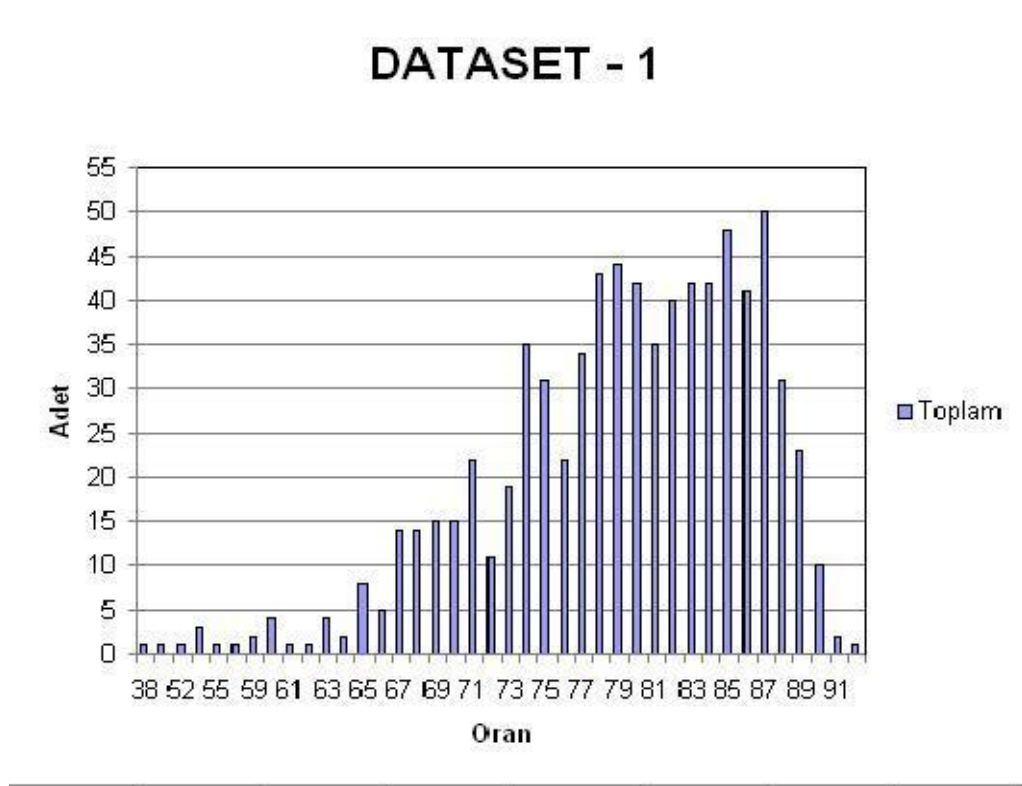
Veri Kümesi	Oran > KDED	Oran < KDED	Toplam	Başarım Oranı
Veri Kümesi - 1	52	870	922	% 5.63
Veri Kümesi - 2	54	460	514	% 10.50

Geliştirilen sistemden kısa sürede maksimum oranda verim almak için kullanıcılara gösterilen resimlerin belli bir metoda göre seçilmesi gerekmektedir. İnsan hesaplama sistemleri geliştirilirken verilerin seçilmesi ile alakalı en sık kullanılan yöntemler şu şekildedir:

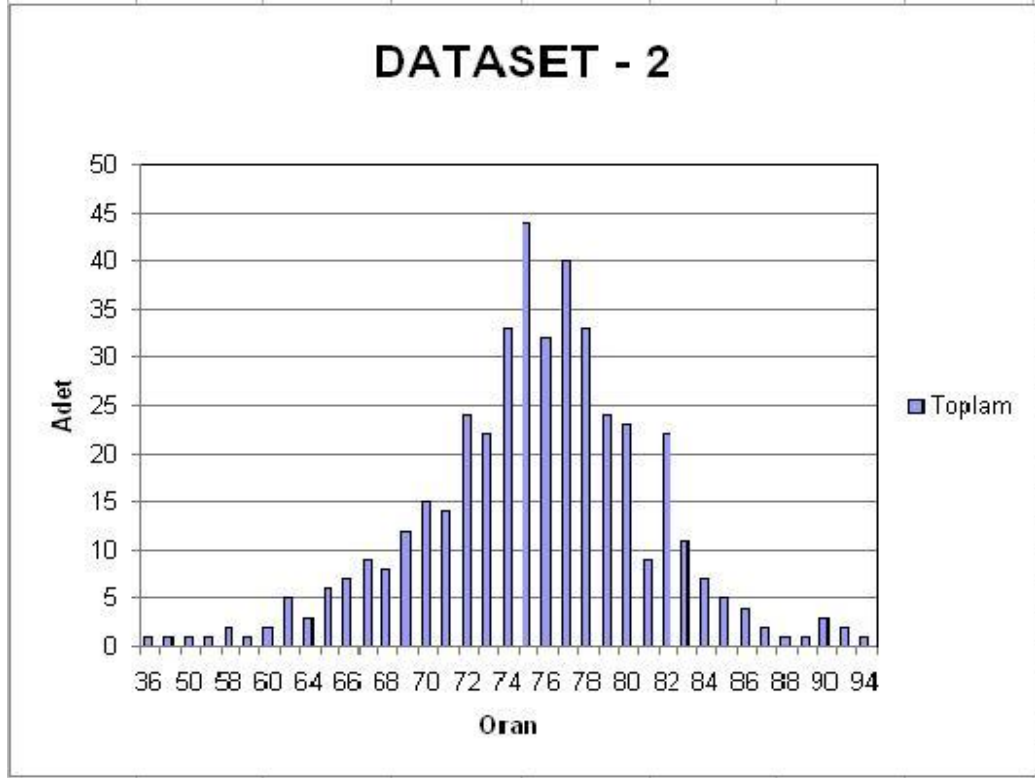
- The Random Puzzle Selection Algorithm (RPSA) ve The Fresh-first Puzzle Selection Algorithm (FPSA) : RPSA, her bir sefer için resim havuzundan rastgele resim seçerken, FPSA resimlerin oranlarına göre en az olanları seçerek çözümlenen resim sayısını arttırmaya yönelik bir yaklaşım sergilemektedir [32,33].
- The Optimal Puzzle Selection Algorithm (OPSA) : Maksimum verim elde etmek ve gerçekleştirimi kolay olduğu için genellikle tercih edilen yöntemler RPSA ve FPSA'dır. OPSA ile her bir resmin kaç kere çözümlenmesi gerektiği yani resimlerin gösterilme sayıları ve yukarıda bahsedilen iki yöntem optimize edilmeye çalışılmıştır [32,33].
- The Adaptive Puzzle Selection Algorithm (APSA) : Geleneksel yaklaşımlar her bir resim için yapılan ortak tahminlerin uygun sayısını tespit etmeye çalışırken resimlerin farklılıklarını görmezden geldikleri için sonuçların eşitliği sorunundan (equality of outcomes issue) etkilenebilirler. APSA (Adaptive Puzzle Selection Algorithm) yöntemi yaygın bulmaca çeşitlilik sorununu (puzzle diversity issue) tanımlamak ve fırsat eşitliğini teşvik ederek bulmaca çeşitliliği ile başa çıkmak için geliştirilmiştir. Bu yaklaşımda resmin bir önceki oyundaki geçmiş bilgisi kullanılarak WST (Ağırlık Sum Tree) ismi ile adlandırdıkları bir veri yapısı oluşturularak hesaplama karmaşıklığını azaltmak, uygulanabilirliğini arttırmak amaçlanmaktadır [31].

Tez kapsamında geliştirilen uygulamada kullanılan yöntem RPSA (Random Puzzle Selection Algorithm) algoritmasının geliştirilmiş halidir. Bu yöntem RPSA'nın aksine seçim yapılacak havuzun sınıflandırılarak kullanılmasını amaçlamaktadır. Sonuçların sınıflandırılması işleminin, kelimelerin tahmin oranlarının KDED'e yakınlıklarına göre yapılması gerekmektedir. Veri kümelerinden elde edilen sonuçlar dikkate alındığında KDED ile çözümlene için geçecek süre arasında doğru orantı

bulunmaktadır. Böylelikle KDED'e yakın olan kelimeler daha hızlı çözümlenerek doygunluğa ulaşmış kelime sayısı arttırılmaktadır. Arttırılan doygunluğa ulaşmış kelime sayısı ile birlikte doygunluğa ulaşmamış kelimeler farklı kombinasyonlar ile kullanıcıya gösterilerek ortak kanı oluşturma imkanı artmaktadır. Çekirdek modülü tarafından oluşturulan veri kümelerinin tahmin oranlarına göre dağılımları Şekil 5.5 ve Şekil 5.6' da gösterilmektedir.



Şekil 5.5 Veri Kümesi 1 İçin Tahmin Oranları Dağılımı



Şekil 5.6 Veri Kümesi 2 İçin Tahmin Oranları Dağılımı

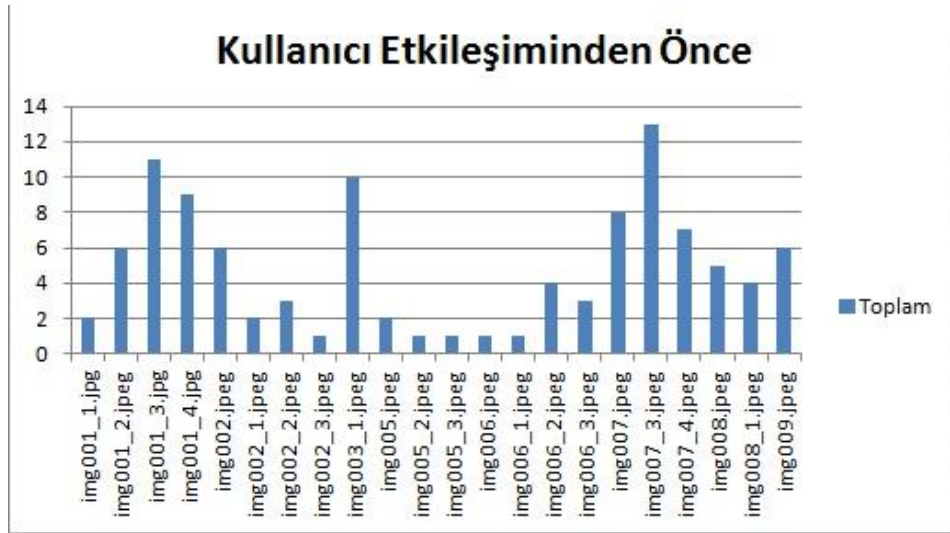
5.3. Gerçek Ortamda trCAPTCHA

TOBB ETÜ ortak eğitim modeli ile öğrenciler normal lisans öğrenimini sürdürürken, her akademik yılda 4 ay olmak üzere, devam ettiği lisans programı ile ilgili bir sanayi veya ticaret kuruluşunda gerçek iş ortamında tecrübe kazanma ve öğrendiklerini uygulama imkanı bulmaktadır.

Ortak eğitim modeli kapsamında öğrenciler ve firmalar tarafından yapılan işlemler (öğrenci tercihi, firma tercihi, yerleşim görüntüleme vb.) Uniform Bilgi Sistemi üzerinden gerçekleştirilmektedir. Bundan dolayı tez kapsamında geliştirilen uygulama, üniversite içerisindeki kullanıcı örnekleminin yanı sıra kullanım alışkanlıkları farklı olan firmalar tarafından da kullanılması amacıyla 2011-2012 Yaz dönemi ortak eğitim tercihleri esnasında devreye alınmıştır. Böylelikle uygulamayı kullanan kullanıcı sayısını arttırarak tahmin sayısını ve çözümlenen kelime sayısını arttırmak hedeflenmiştir. Fakat firmalardan gelen geri bildirimler sonucu uygulama

geri çekilmiş ve geliştirme modunda TOBB ETÜ Öğrenci İşlerine ve TOBB ETÜ Bilgi Teknoloji Müdürlüğünün kullanımına açılmıştır.

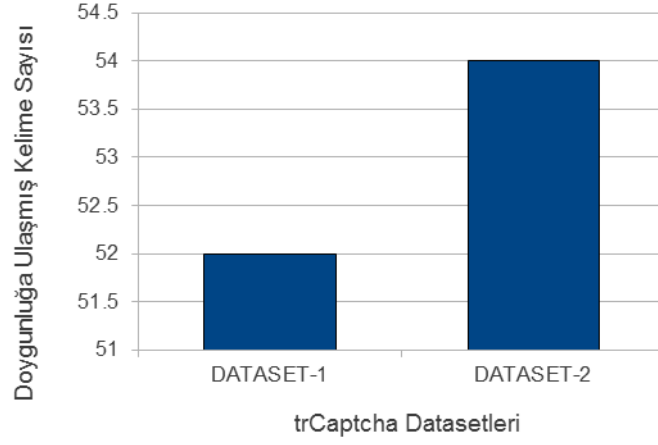
Tez kapsamında geliştirilen uygulama Uniform Bilgi Sistemi ile entegre olacak şekilde geliştirilmiş ve kullanıcıların bu sisteme girişleri sırasında karşılaştıkları görüntülerdeki kelimeleri tanımlamaları beklenildi. Böylelikle optik karakter okuyucu uygulamalarından elde edilen sonuçların kullanıcıların girdikleri verilerle birlikte daha fazla doyumluğa ulaştırılması hedeflenmiştir.



Şekil 5.7 Kullanıcı Etkisi Olmadan Doygunluğa Ulaşmış Kelime Sayısı

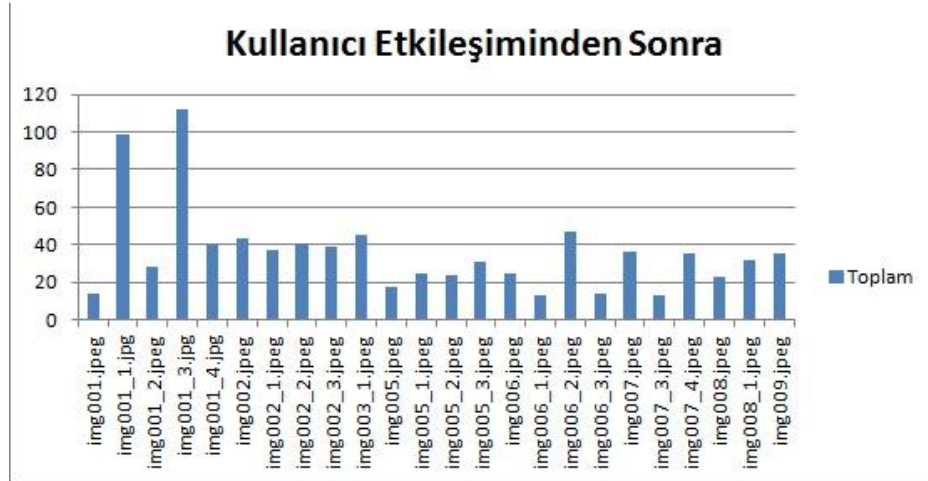
Şekil 5.7' de verilen görüntülerin doyumluğa ulaşmış kelime sayıları gösterilmektedir. Toplamda 106 kelime doyumluğa ulaşmıştır. Bu çizelge kelimelerin, kullanıcıların kullanımına açılmadan yalnızca optik karakter okuyucu uygulamaları kullanılarak elde edilen başlangıç durumuna ait kelime doyumluk oranlarını göstermektedir.

Dataset Doğunluk Durumu



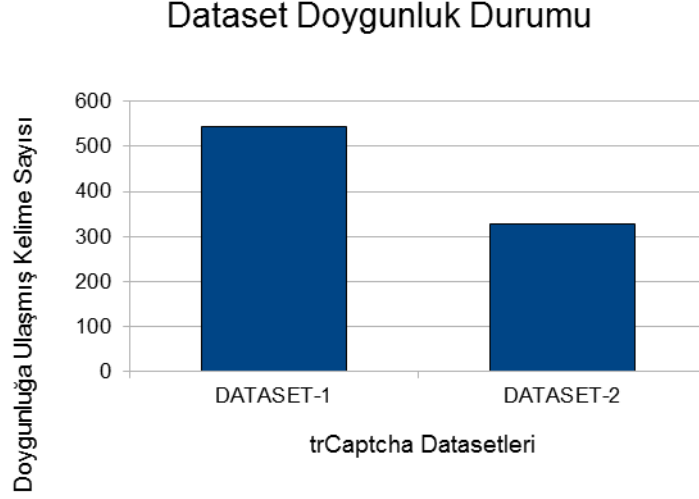
Şekil 5.8 Kullanıcı Etkisi Olmadan Veri Kümelerinin Doğunluk Durumu

Şekil 5.8’de kullanıcı etkeni olmadan yalnızca tez kapsamında kullanılan optik karakter okuyucu uygulamaları sonucunda veri kümeleri içerisindeki kelimelerin eriştiği doğunluk durumu ifade edilmiştir. Uygulama sonrasında elde edilen veri kümeleri doğunluk oranı ile karşılaştırılacaktır.



Şekil 5.9 Kullanıcı Etkisi İle Doğunluğa Ulaşmış Kelime Sayısı

Şekil 5.9’ da kullanıcı etkisinden sonra doygunluğa erişmiş kelime sayılarının Şekil 5.7’e göre arttığı görülmektedir.



Şekil 5.10 Kullanıcı Etkisi İle Veri Kümelerinin Doygunluk Durumu

Şekil 5.10’ de doygunluk durumunun dataset-1 için 52’den 520’ye, dataset-2 için 54’ten 340’a çıkmıştır. Görülmektedir ki uygulama insan etkileşimi ile gerçekleştirildiğinde görüntülerden elde edilen kelimelerin doygunluk durumu artmaktadır.

6. SONUÇ

Basılı kaynaklardaki metinlerin sayısal ortama aktarılması işlemi, sayısal ortamların ortaya çıkmasından günümüze süregelen süreçte farklı yaklaşımlarla gerçekleştirilmeye çalışılmıştır. Bu yaklaşımlardan ilk uygulananı, arşivlerin insanlar vasıtası ile el yordamına dayanarak sayısal ortama aktarılmasıdır. Fakat bu yaklaşım masraflı ve uzun zaman gerektirmektedir.

Bilgisayar dünyasında kaydedilen gelişmeler, bilgisayarların geniş hesaplama ve işlem gücü yeteneklerine sahip olması arşiv sayısallaştırma probleminin çözüm yönteminin insanlardan OCR yazılımlarına doğru kaymasına sebep olmuştur. OCR yazılımları arşiv sayısallaştırma yöntemlerinin etkinlik, verimlilik gibi etkenler ile değerlendirildiğinde başarısı çok yüksek bir yöntem olarak öne çıkmıştır. OCR yazılımlarının arşiv sayısallaştırılmasında önem kazanmasına rağmen sayısallaştırmak istenen arşivlerin özelliklerinden dolayı hatalar ile karşılaşılması kaçınılmazdır. İnsan gücü kullanılarak ve el yordamı ile gerçekleştirilmeye çalışılan ve başarısında insan faktörü olan yöntemler yerine OCR yönteminde otomatize edilebilir ve başarısı yüksek bir tanıma ve aktarma işlemi elde edilebilmektedir.

Bu tezde bahsi geçen yaklaşım ve metotlar veri sayısallaştırma ve bu verilerin sayısal arşivler oluşturmasında etkin çözüm ortaya sunmaktadır. Captcha yönteminin kullanıldığı yetkilendirme ya da doğrulama sistemlerinde görüntü olarak sunulan karakterler ile hem doğrulama sağlanmakta hem de OCR yöntemi ile elde edilen kelimeler sayısal ortama aktarılmaktadır.

Karmaşık örüntü tanıma algoritmaları kurmak, bu işlemlerin çözümü için yüksek hesaplama, işlemci gücü ve dolayısıyla elektrik harcamak yerine kelimeleri soyut olarak da anlamlandırabilen insan kontrolünü kullanmak ve bunu doğrulama ve yetkilendirme için kullanılan zorunlu bir sistem içerisinde gerçekleştirerek arşiv sayısallaştırma işinin etkili bir şekilde sağlanabileceği düşünülmektedir.

Böyle bir çalışma sayesinde sadece bir OCR yazılımı yerine bir kaç farklı OCR yazılımından görüntüler metne çevrilerek başarı oranının yükselme olasılığı ortaya koyulmuş olacaktır. Ayrıca benzeri arşiv sayısallaştırma uygulamalarındaki oyun yaklaşımlarının aksine güncel işlerini gerçekleştirmeleri için uygulamayı kullanmak zorunda olmaları sebebiyle kullanıcıların dikkatini yoğunlaştırmalarını gerektirmektedir. Bu durumda tezde bahsi geçen uygulama ile insan etkeninin olumlu bir kaynağa dönüştürebileceği gösterilmektedir.

6.1. Gelecek Çalışmalar

Tez çalışması kapsamında geliştirilen trCAPTCHA uygulaması daha önce de bahsedildiği gibi Türkçe arşivlerin sayısallaştırmasına yardımcı olmayı hedeflemektedir. Uygulamanın kullanıcılara açılmasının ardından elde edilen sonuçlar analiz edildiğinde bu hedefin başarıldığı görülmektedir. Bu başarıyı daha üst seviyeye çıkartmak için trCAPTCHA uygulamasını temel alarak bir takım iyileştirmeler ve eklemeler yapılabilir. Bu ekleme ve iyileştirmelerden bazıları şunlardır;

- trCAPTCHA ve OCR yazılımları arasında asimetrik bir bağ bulunmaktadır. Yani trCAPTCHA uygulaması OCR yazılımları tarafından oluşturulan sonuçları kullanmaktadır. Fakat trCAPTCHA uygulaması OCR yazılımlarına bir katkıda bulunmamaktadır. Bundan dolayı çözümlenen kelimelere ait karakterlerden oluşan bir veri tabanı oluşturularak OCR yazılımlarının okuyamadığı karakterleri bu veri tabanından kontrol ederek daha güçlü sonuçlar üretilmesi sağlanabilir.
- trCAPTCHA uygulaması tez çalışması kapsamında Uniform Bilgi Sistemi'ne entegre olacak şekilde geliştirilmiştir. Farklı türlerde arşivlerin de sayısallaştırılabilmesi amacıyla farklı sistemlere kolay entegrasyonu sağlayan bir servis geliştirilebilir.
- Geliştirilen uygulama kapsamında “Kullanılan Yazılım ve Teknolojiler” bölümünde de belirtildiği üzere 3 adet OCR yazılımı kullanılmıştır. Uygulama mimarisi yeni OCR yazılımlarını kolay destekler halde tasarlanmıştır. Bu mimari avantajı kullanarak yeni OCR yazılımları eklenebilir. Bu eklemeler sayesinde var

olan OCR yazılımlarının yaptığı hatalar daha aza indirilebilir. Kullanılan OCR yazılımlarının artmasının bir başka avantajı ise taranan kelimelere ait alternatif kelimelerin sayısının artması böylelikle sonuca daha hızlı ulaşılması sağlanabilir.

- Uniform Bilgi Sistemine sağlanan entegrasyon ile kullanıcılar sisteme giriş esnasında trCAPTCHA uygulaması ile karşılaşmaktadırlar. Bazı kullanıcılar (öğrenciler ve ortak eğitim firmaları gibi) trCAPTCHA uygulaması ile daha az etkileşime geçmektedir. Bundan dolayı TOBB ETÜ’de farklı sistemlere entegre olarak kullanım daha da arttırılabilir. Kullanımın artması beraberinde daha fazla kelimenin ve arşivin sayısallaşmasını sağlayabilir.

7. KAYNAKLAR

- [1] E. Horvitz and T. Paek. Complementary computing: Policies for transferring callers from dialog systems to human receptionists. *User Modeling and User Adapted Interaction*, 17, 2007. DOI: 10.1007/s11257-006-9026-1
- [2] Advantage Research Incorporated, Data Collection Services Quality Control. Erişim Tarihi: 11 Eylül 2011.
Erişim Adresi: www.advantageresearchinc.com/dcsqualitycontrol.htm.
- [3] "reCAPTCHA: Stop Spam, Read Books".
Erişim Tarihi : 10 Temmuz 2010.
Erişim Adresi : <http://www.google.com/reCAPTCHA/learnmore>
- [4] "Wikipedia, the free encyclopedia".
Erişim tarihi: 10 Temmuz 2010.
Erişim Adresi : <http://tr.wikipedia.org/wiki/Captcha>
- [5] "The Official CAPTCHA Site".
Erişim tarihi : 10 Temmuz 2010.
Erişim Adresi : <http://www.captcha.net/>
- [6] Pinkas, B. and Sander, T. Securing Passwords Against Dictionary Attacks. In *Proceedings of the ACM Computer and Security Conference (CCS' 02)*, pages 161-170. ACM Press, November 2002.
- [7] L. von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using Hard AI Problems for Security. In E. Biham, editor, *Advances in Cryptology — EUROCRYPT 2003*, volume 2656 of *Lecture Notes in Computer Science*, chapter 18, page 646. Springer Berlin / Heidelberg, Berlin, Heidelberg, May 2003. ISBN 978-3-540-14039-9. doi: 10. 1007/3-540-39200-9_18. URL http://dx.doi.org/10.1007/3-540-39200-9_18, 3
- [8] E. Law and L. von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the 27th international conference on Human factors in computing systems, CHI '09*, pages 1197–1206, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518881. URL <http://dx.doi.org/10.1145/1518701.1518881>. 21, 29, 30
- [9] Luis von Ahn, *Human Computation*, December 7, 2005, CMU-CS-05-193
- [10] von Ahn, L. and L. Dabbish, "Labeling images with a computer game", *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, 2004.
- [11] Kuan-Ta Chen, Chien-Wei Lin, Ling-Jyh Chen and Irwin King (2010). *Human Computation Games and Optimization of Their Productivity*, *Web Intelligence and Intelligent Agents*, Zeeshan-UI-Hassan Usmani (Ed.), ISBN: 978-953-7619-85-5, InTech, <http://www.intechopen.com/articles/show/title/human-computation-games-and-optimization-of-their-productivity>
- [12] Ingmar Weber, Stephen Robertson, Milan Vojnovi, "Rethinking the ESP Game", *Technical Report MSR-TR-2008-132*, September 2008

- [13] von Ahn, L., R. Liu, and M. Blum, "Peekaboom: a game for locating objects in images", CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, New York, NY, USA, 2006.
- [14] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A Game for Locating Objects in Images. In Proceedings of the SIGCHI conference on Human Factors in computing
- [15] Lenat, D. B. CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38(11), 1995, pages 32-28.
- [16] Mindpixel Project.
Erişim Tarihi: 23 Aralık 2010.
Erişim Adresi: <http://www.mindpixel.org>
- [17] The Open Mind Common Sense Project.
Erişim Tarihi. 22 Şubat 2010.
Erişim Adresi: <http://www.kurzweilai.net/articles/art0371.html>
- [18] Instructions, Taboo.
Erişim Tarihi: 23 Kasım 2011
Erişim Adresi: <http://www.hasbro.com/common/>
- [19] von Ahn, L., M. Kedia, and M. Blum, "Verbosity: a game for collecting common-sense facts", CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, New York, NY, USA, 2006.
- [20] Kim, Y.E., Schmidt, E., and Emelle, L. MoodSwings: A collaborative game for music mood label collection. In Proc. 9th Intl. Conf. on Music Information Retrieval (Philadelphia, September 14-18). 231-236, 2008.
- [21] Lamere, P. Social tagging and music information retrieval. Journal of New Music Research. 37(2):101-104, 2008.
- [22] Law, E., von Ahn L., Dannenberg, R., and Crawford M. TagATune: a game for music and sound annotation. In Proc. 8th Intl. Conf. on Music Information Retrieval. 361-364, 2007.
- [23] Duranti, L. (2000). The InterPARES Project: InterPARES Symposium Proceeding, February 2000 – The complete proceedings from the February 2000, InterPARES Symposium.
Erişim Tarihi: 2 Ekim 2009. Erişim Adresi:
http://www.interpares.org/display_file.cfm?doc=ip1_symposium_2000.pdf
- [24] Runardotter, M. (2007). Information technology, archives and archivists- an interacting trinity for long- term digital preservation.
Erişim Tarihi: 29 Eylül 2009.
Erişim Adresi: <http://epubl.luth.se/1402-757/2007/08/LTU-LIC-0708-SE.pdf>
- [25] Türkiye İstatistik Kurumu (TÜİK) Hanehalkı Bilişim Teknolojileri Kullanım Anketi (2010).
Erişim Tarihi: 24 Ekim 2010.
Erişim Adresi: http://www.tuik.gov.tr/PreTablo.do?tb_id=60&ust_id=2
- [26] United Nations Archives and Records Management Section. (2006). Guideline on records digitisation.
Erişim Tarihi: 26 Ekim 2010.
Erişim Adresi:
<http://archives.un.org/unarms/en/unrecordsmgmt/unrecordsresources/guideline%20on%20records%20digitisation.htm>

- [27] AccesIT Türkiye projesi sunumları.
Erişim Tarihi: 23 Kasım 2011.
Erişim Adresi: <http://www.accessit.hacettepe.edu.tr/>
- [28] Tesseract OCR.
Erişim Tarihi: 24 Eylül 2011.
Erişim Adresi: <http://code.google.com/p/tesseract-ocr/>
- [29] Gocr.
Erişim Tarihi: 24 Eylül 2011.
Erişim Adresi: <http://jocr.sourceforge.net/>
- [30] Ocrad.
Erişim Tarihi: 24 Eylül 2011.
Erişim Adresi: <http://www.gnu.org/software/ocrad/>
- [31] “Wikipedia, the free encyclopedia”.
Erişim Tarihi: 11 Ocak 2011.
Erişim Adresi: http://en.wikipedia.org/wiki/Amazon_Mechanical_Turk
- [32] Yu-Song Syu, Hsiao-Hsuan Yu, Ling-Jyh Chen, “Exploiting Puzzle Diversity in Puzzle Selection for ESP-like GWAP Systems”, Research Center for Information Technology Innovation, Academia Sinica, Institute of Information Science, Academia Sinica, November 2010
- [33] Ling-Jyh Chen, Bo-Chun Wang, Kuan-Ta Chen, Irwin King and Jimmy Lee, “An Analytical Study of Puzzle Selection Strategies for the ESP Game”, Institute of Information Science, Academia Sinica, Department of Computer Science & Engineering, Chinese University of Hong Kong, 2008
- [34] J. Ross, L. Irani, M.S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers? shifting demographics in mechanical turk. In CHI Work-in-Progress, pages 2863–2872, 2010. DOI: 10.1145/1753846.1753873 Cited on page(s) 46, 47, 74
- [35] Amazon mechanical turk. <https://www.mturk.com/mturk/welcome>. Cited on page(s) 46
- [36] P. Ipeirotis. Analyzing the amazon mechanical turk marketplace. XRDS, 17(2), 2010. DOI: 10.1145/1869086.1869094 Cited on page(s) 45, 46, 61
- [37] W. Mason and S. Suri. Conducting behavioral research on amazon’s mechanical turk. In Social Science Research Network Working Paper Series, 2010. Cited on page(s) 46, 47
- [38] F. A. Long, Am. J. Econ. Sociol. 52, 223 (1993).
- [39] Crevier 1993, p. 50

8. ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : GÜMÜŞ, İbrahim
Uyruğu : T.C.
Doğum tarihi ve yeri : 18.10.1983 Fransa
Medeni hali : Bekar
Telefon : 0 (312) 292 40 47
Faks : 0 (312) 292 40 91
e-mail : igumus@etu.edu.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Lisans	KTÜ/İstatistik ve Bilgisayar Bil.	2006

İş Deneyimi

Yıl	Yer	Görev
2010 -	TOBB ETÜ BTM	Yazılım Müh.
2007 - 2009	Onmap - Bilgi Sistemeleri, Ankara	Yazılım Müh.

Yabancı Dil

İngilizce

Yayınlar