BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAYISAL VİDEO VE ÇİZİM VERİLERİNDE ANLAMSAL KAVRAM TANIMA

EMEL BOYACI

YÜKSEK LİSANS TEZİ
2017

SAYISAL VİDEO VE ÇİZİM VERİLERİNDE ANLAMSAL KAVRAM TANIMA

SEMANTIC CONCEPT RECOGNITION IN DIGITAL VIDEO AND SKETCH DATA

EMEL BOYACI

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2017

“SAYISAL VİDEO VE ÇİZİM VERİLERİNDE ANLAMSAL KAVRAM TANIMA” başlıklı bu çalışma, jürimiz tarafından, 11/08/2017 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI 'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan	Prof. Dr. Adnan YAZICI
Üye (Danışman)	Yrd. Doç. Dr. Mustafa SERT
Üye	Prof. Dr. Hasan OĞUL

ONAY

..../..../.....

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

BAŞKENT ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS / DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 08/09/2017

Öğrencinin Adı, Soyadı : Emel Boyacı

Öğrencinin Numarası : 21310038

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Bilgisayar Mühendisliği Tezli Yüksek Lisans

Danışmanın Unvanı/Adı, Soyadı : Yrd. Doç. Dr. Mustafa SERT

Tez Başlığı : Sayısal Video ve Çizim Verilerinde Anlamsal Kavram Tanıma

Yukarıda başlığı belirtilen Yüksek Lisans/Doktora tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 49 sayfalık kısmına ilişkin, 08/09/2017 tarihinde tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %4'dır.

Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:…….

Onay

… / … / 20…

Öğrenci Danışmanı Unvan, Ad, Soyad,

Yrd. Doç. Dr. Mustafa SERT

TEŞEKKÜR

Bu çalışmamın gerçekleştirilmesinde, tez konusunu seçerken isteklerimi göz önünde bulundurup bana yardımcı olan, rehberlik eden, yol gösteren ve desteğini hiç bir zaman esirgemeyen tez danışmanım Sayın Yrd. Doç. Dr. Mustafa SERT'e teşekkürlerimi sunarım.

Tüm hayatım boyunca benden maddi ve manevi desteklerini esirgemeyen ve her zaman yanımda olan sevgili aileme teşekkürlerimi bir borç bilirim.

Ayrıca iş arkadaşlarım Sayın Mete EZER, Işık AYRANCI KIVRAK, Büşra TANGIÇ, Habip Kenan ÜSKÜDAR ve dostlarıma çalışma sürecinde bana gösterdikleri anlayış ve desteklerinden ötürü teşekkür ederim.

Son olarak her zaman destekçim ve dostum olan Sayın Semih MALKOÇ'a tez boyunca desteklerinden dolayı minnettarım.

ÖZ

SAYISAL VİDEO VE ÇİZİM VERİLERİNDE ANLAMSAL KAVRAM TANIMA

EMEL BOYACI
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Günümüzde çoğul ortam aracı olan video verileri günlük hayatımızda önemli bir rol oynamakta ve eğitim, iletişim, sağlık, eğlence alanlarında oldukça yaygın olarak kullanılmaktadırlar. Günlük yaşantıda, anlık paylaşımların çoğalması, video veri kullanımını büyük ölçüde artırmıştır. Sonuç olarak, bu verilerin yönetimi, sınıflandırılması, sezimi ve geri getirim yöntemlerine ihtiyaç duyulmaktadır. Bu gibi işlevleri mümkün kılmak için atılması gereken en önemli adım, bu verilerin anlamsal kavramlarını tahmin etmek ve anlamaktır. Video anlamsal kavram sezimi, son yıllarda çoklu ortam endüstrisi tarafından önemli bir araştırma problemi olarak görülmektedir. Sınıflandırma, kavram sezimi için kullanılan en kabul gören yöntem olup, sınıflandırma sisteminin çıktısı anlamsal kavramlar olarak yorumlanmaktadır. Bu kavramlar otomatik dizinleme, video nesnelerinin aranması ve geri getirim (retrieval) için kullanılabilmektedir. Bununla birlikte, kullanılan özniteliklerin boyutları yüksek olup, mevcut sınıflandırıcılarla kavram tespiti yüksek hesaplama karmaşıklığına maruz kalmaktadır. Bu tez çalışmasında, derin evrişimsel sinir ağ (Convolutional Neural Network) modelleri üzerinde, öznitelik etkinliğini artırmak amacıyla öznitelik seçimi ve veri kaynaştırma tekniklerine dayalı video içeriklerinden kavram sezim yöntemleri önerilmektedir. Eğitim maliyetini azaltmak amacıyla Temel Bileşen Analizi (PCA-Principal Component Analysis) tekniği öznitelik düzeyinde probleme uygulanmıştır. Farklı derin evrişimsel sinir ağlarından elde edilen kaynaşmış öznitelik vektörlerinin sınıflandırılmasında Destek Vektör Makineleri (SVMs-Support Vector Machines) kullanılmıştır. Video kavram sezimi için önerilen yöntemde geçmiş çalışmalarda tercih edilen TRECVID 2013 SIN video veri kümesi (38 kavram) kullanılmıştır. Geliştirilen sınıflandırma yönteminin etkinliğini değerlendirmek amacıyla, önerilen video kavram sınıflandırma yöntemi çizim tanıma problemine de uygulanmıştır. Elde edilen sonuçlara göre, öznitelik seviyesinde veri kaynaşım tanıma başarımlarını artırmıştır.

ANAHTAR SÖZCÜKLER: Video Kavram Sezimi, Çizim Kavram Tanılama, Evrişimsel Sinir Ağları (CNN), Veri Kaynaştırma, PCA, DVM.

Danışman: Yrd. Doç. Dr. Mustafa SERT, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

SEMANTIC CONCEPT RECOGNITION IN DIGITAL VIDEO AND SKETCH DATA

EMEL BOYACI

Başkent University Institute of Science and Engineering

Computer Engineering Department

Nowadays, video data, which is a multimedia tool, plays an important role in our life and being used commonly in the fields of education, communication, health and history. Furthermore, the proliferation of instant sharing has greatly increased the use of video data in daily life. As a result, there is a need for new techniques in the field of computer vision for the management, classification, detection and retrieval of these data. The most important step to make such functions possible is to estimate and understand the semantic concepts of these data. In recent years, video concept detection has been seen as an important research problem by the multimedia industry. Classification is the most accepted method for concept detection and the outputs of the classification system are interpreted as semantic concepts. These concepts can be used for automatic indexing, search and retrieval of video objects. However, the dimensions of the features used are high and the concept detection with existing classifiers is subject to high computational complexity. In this thesis study; video concept detection based on feature selection and data fusion techniques are proposed in order to enhance the feature effectiveness on the Convolutional Neural Network models for video contents. In order to reduce the cost of training, Principal Component Analysis (PCA) technique is applied at the feature level. Support Vector Machines (SVMs) were used to classify fused feature vectors obtained from different deep convolution neural networks. TRECVID 2013 SIN video data set (38 concepts), which is preferred in past studies, is used in the proposed method for video concept detection. Proposed method for video concept detection is also applied to sketch recognition problem in order to measure the effectiveness of the developed classification method. According to the results obtained, these systems which proposed the data fusion method at the feature level, increased the recognition success.

KEYWORDS: Video Concept Detection, Sketch Concept Recognition, Convolutional Neural Networks (CNNs), Data Fusion, PCA, SVM.

Supervisor: Asst. Prof. Dr. Mustafa SERT, Başkent University, Computer Engineering Department.

İÇİNDEKİLER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

CNN	Convolutional Neural Network
MLP	Multi-Layer Perceptron
PCA	Principal Component Analysis
DCA	Discriminant Correlation Analysis
FV	Fisher Vektor
MKL	Multiple Kernel Learning
DT	Distance Transform
BoW	Bag-of-Words
DVM	Destek Vektör Makinası
RTF	Radyal Tabanlı Fonksiyon
OVA	One-Versus-All
Scene	Sahne
Shot	Kayıt
Frame	Çerçeve
SOAP	Simple Object Access Protocol
REST	Representational State Transfer
XML	Extensible Markup Language
XSD	XML Schema Definition
WSDL	Web Services Description Language
JSON	Java Script Object Notation
TRECVID	TREC Video Retrieval Evaluation
SIN	Semantic Indexing
ILSVRC	Large Scale Visual Recognition Challenge
IEEE	Institute of Electrical and Electronics Engineering

1 GİRİŞ

Sayısal görüntü ve videoların hızla artan bir şekilde yaygınlaşması, çoklu ortam veritabanlarında içerik tabanlı arama giderek önemli bir problem haline gelmektedir. İşlenmesi gereken büyük miktarda veri ve buna bağlı yüksek donanım gereksinimleri nedeniyle, son yıllarda bilgisayar görü alanında görüntü ve video analizi önemli bir yere sahiptir. Etkili görüntü ve video araması için bir ön şart, çoklu ortam veri içeriğini otomatik olarak dizinlemektir.

Sayısal video yakalama cihazlarının ve çevrimiçi arama motorlarının yaygın kullanımı ile çoğu kullanıcı, büyük video kaynaklarıyla etkileşim halinde olan basit arayüzlere ihtiyaç duymaktadır. Videolar, genellikle ilişkili metin anahtar kelimeleri tarafından tanımlanmayan yapılandırılmamış bilgileri içerdiğinden, ham video verilerinden üst düzey bilgi çıkarmak için anlamsal kavram sınıflandırma ve sezim işlemlerine ihtiyaç duyulmaktadır. Anlamsal kavram tespiti, video dizisinde bir veya birden fazla anlamsal kavramın varlığını belirten bir video ataması yapmak veya bir veya daha fazla etiket (kavram) atamak görevi olarak tanımlanmaktadır. Anlamsal kavram tespit sistemleri, içerik veya anahtar kelime tabanlı video arama, içerik tabanlı video özetleme, robotik uygulamalar gibi geniş bir uygulama alanlarına destek sağlayan, otomatik indeksleme ve büyük çoklu ortam bilgisinin organizasyonunu mümkün kılmaktadır.

Bu nedenlerden dolayı, video ve görüntü (çizim vb.) için kavram sınıflandırma gereksinimine ihtiyaç duyulmaktadır.

1.1 Videolarda Anlamsal Kavram Sınıflandırması

Videolar görüntülerden oluşur ve bu görüntüler bir video kayıtının çekimlerini oluşturur. Bir video kayıtının ilgili çekimleri video sahnelerini oluşturur ve her sahne anlamsal bir bütünlük taşır. Bir video sahnesindeki anlamsal bütünlük, video sezim, dizinleme ve arama gibi birçok alanda önemli bir yere sahiptir. Video sahnelerinin otomatik olarak açıklanması maliyetli bir iş olduğundan, araştırmacılar, video verileri üzerinde anlamsal kavram yöntemleri üzerinde çalışmaktadırlar. Video kavram sınıflandırmasının bir görüntünün anlamsal kategorisini tanıma problemiyle karşı karşıyadır. Video kavram sınıflandırması, nesne tanıma veya sahne tanıma seviyesinde olabilir (Van De Sande et al., [42]).

1.2 Çizimlerde Anlamsal Kavram Sınıflandırması

Tarih öncesi çağlardan beri çizim, eşsiz bir iletişim yöntemi olmuştur. Günümüzde, dokunmatik ekranlı cihazların (dokunmatik yüzeyler ve dokunmatik telefonlar) artan popülaritesi ile çizim, insan-bilgisayar etkileşiminde önemli bir yere sahip olmuştur. İnsanlar, akıllı telefonda çizdikleri bir sahneye benzer görüntüleri almak istemektedirler. Özellikle okul öncesi çocuklar için etkili bir şekilde bilgisayarlarla iletişim kurmak için çekici bir yöntemdir. Dolayısıyla, bilgisayarların insan eliyle çizilmiş resimlerin anlamasını sağlamak son derece önemlidir.

Etkili bir çizim tanıma sistemi, bir bilgisayarın insan ile etkileşime girmesine, etkili çizim tabanlı arama yapmasına, çocuk eğitimlerinin bilgisayarlarla geliştirmesine ve oyun tasarımını geliştirmesine olanak tanımaktadır.

1.3 Problem Tanımı

Çoklu ortam içeriğinden (resim, ses, video) anlamsal bilgilerin çıkarılması uzun zamandır zorlu ve popüler bir araştırma alanı olmuştur. Ayrıca bu verilerin büyük boyutta olmaları analiz, sınıflandırma, geri getirim maliyetlerini de büyük ölçüde artırmıştır. Görsel kavram sezimi, büyük karmaşıklık ve değişkenlik içerdiğinden dolayı kavramların tespit edilmesi zor bir görevdir. Ayrıca kavram tespiti alanındaki en büyük sorun, başarılı kavram sezim sistemlerinin temelini oluşturan etkin özniteliklere sahip olmamasıdır. Bu nedenle, öznitelik çıkarımı, özniteliklerin etkin kullanımı önemli bir problem haline gelmiştir.

Son zamanların en popüler konularından biri olan ve çoklu ortam verileri üzerindeki başarımlarının oldukça yüksek olan Evrişimsel Sinir Ağları, (CNN -Convolutional Neural Network) (LeCun et al., [63]) son yıllarda bilgisayarla görü alanında önemli bir kullanıma sahiptir. Bilgisayar donanımındaki ilerlemelerle birlikte, daha büyük makine öğrenme sistemlerini eğitmek daha kolaydır. Özellikle paralel mimarilerin daha kolay erişilebilirliği, veri eğitimde kullanılacak büyük veri setlerinin (Trecvid, ImageNet) ortaya çıkmasıyla bu sistemlerin uygulanabilirliğini artırmıştır.

Özetle, bu tez çalışmasında çoklu ortam koleksiyonlarında sınıflandırma ve sezimlemeyi desteklemek için etkili öznitelik çıkarım tekniklerine duyulan ihtiyaç hızla artmaktadır. Üst düzey öznitelik çıkarma (high-level feature extraction) veya anlamsal dizinleme olarak da bilinen görsel kavram seziminde amacımız etkin

öznitelikleri kullanarak ve eğitim maliyetini mümkün olan en düşük seviyede tutarak başarımları artırmaktır.

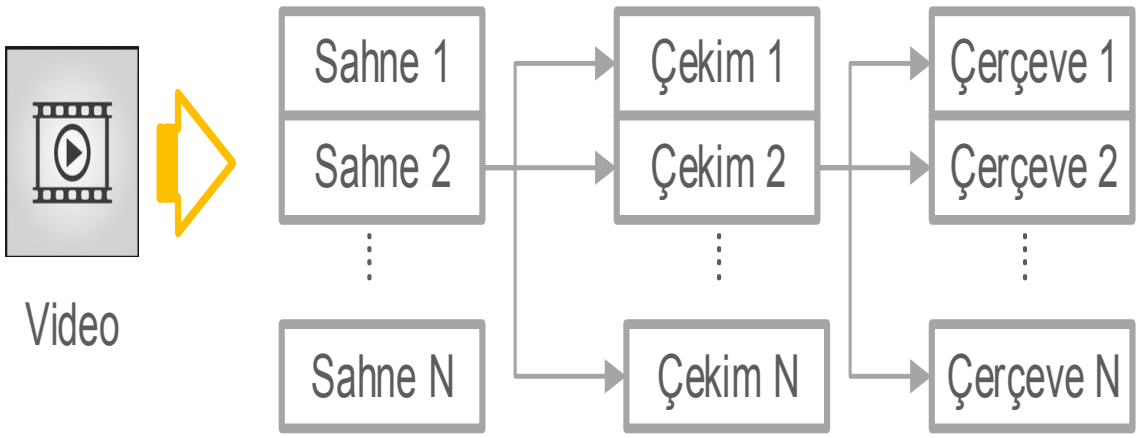
1.4 Tanımlamalar

Video: Bir veya birden fazla sahneden oluşmaktadır. Her sahne kendi içinde çekim dediğimiz benzer çerçevelerin bir araya gelerek oluşan yapıdır.

Çekim (Shot): Tek bir kamera tarafından tek bir sürekli hareketle yakalanan bir dizi karedir. Bir çekim sınırı iki çekim arasındaki geçiştir (Basanth Kumar, [64]).

Sahne (Scene): Çekimlerin anlamsal bir birime mantıksal olarak gruplandırılmasıdır (Basanth Kumar, [64]).

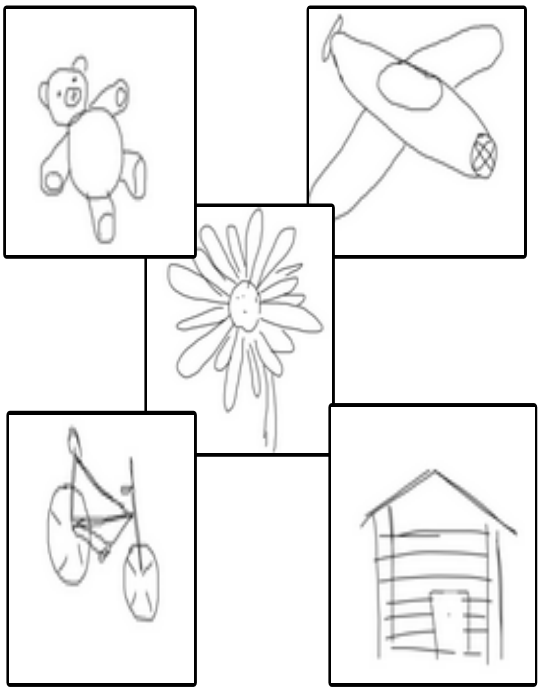
Çerçeve (Frame): Bir çekim içerisindeki benzer görüntü dizinidir.



Şekil 1.1 Video bölümlerinin genel yapısı

Bu çalışmada, video verilerinden çıkarılarak elde edilen çerçeveler üzerinde video kavram sınıflandırması yapılmıştır. Video verisinin mantıksal yapısı Şekil 1.1'de gösterilmiştir.

Çizim: Duygu ve düşünceleri ifade etmenin doğal bir yoludur. Metin kullanarak açıklanan bilgilere göre daha çok şey ifade edebilmektedir. Aynı zamanda çocuklar ya da okuma yazma bilgisi olmayan insanlar için uygun bir iletişim aracıdır. İnsan-bilgisayar etkileşimi daha kolay ve daha üst düzey dillere doğru ilerledikçe, çizim her türlü uygulamada yerini almaya devam edecektir. Şekil 1.2'de örnek çizimler verilmiştir (Eitz et al., [14]).

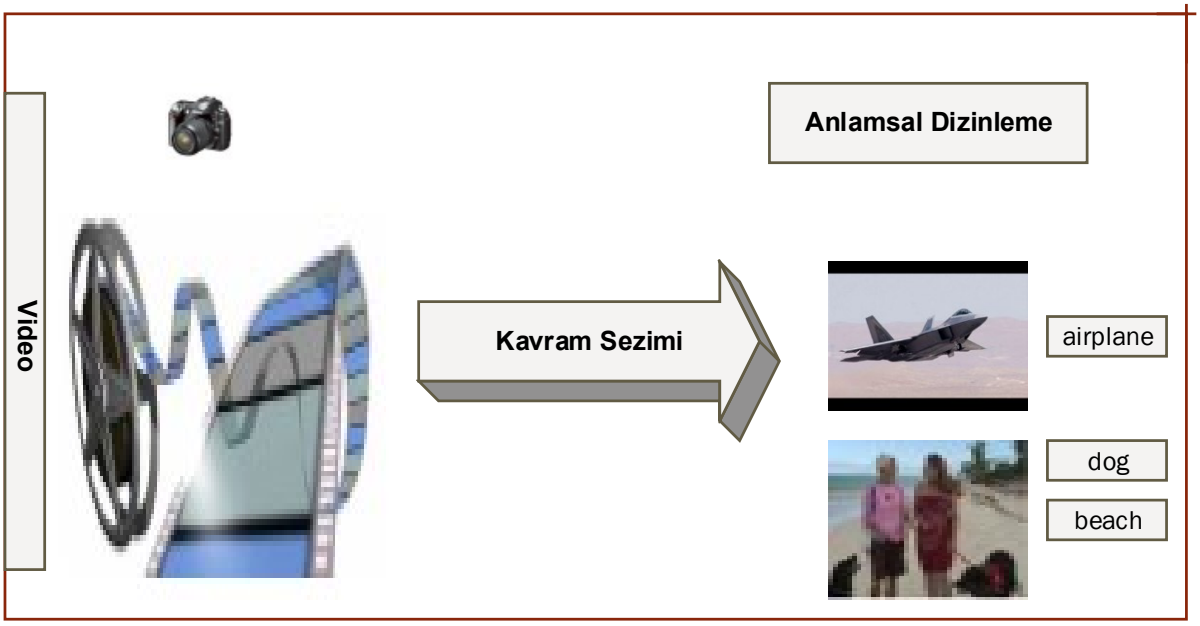


Şekil 1.2 Çizim örnekleri (Sangkloy et al., [22])

1.5 Motivasyon

Teknolojinin hızla gelişmesiyle birlikte, video verisinin sosyal medya (İnstagram, Facebook, Snapchat), sağlık ve eğitim alanında kullanımı giderek artmaktadır. Video içeriklerinden arama yapmak, sezimlemek ve onu daha erişilebilir hale getirmek için, doğru bir şekilde sınıflandırılması gerekmektedir. Kullanıcılar, cihazlarında veya uygulamalarında bu verileri verimli bir şekilde organize etmek ve araştırmak istemektedirler. Bu tür işlevleri mümkün kılmak için bilgisayarla görü ve bilgi alma tekniklerine (sezimleme, sınıflandırma) ihtiyaç duyulmaktadır. Kavram sezimi, anlamsal kavramların video çekimlerine otomatik olarak atama görevini temsil etmektedir (Şekil 1.3).

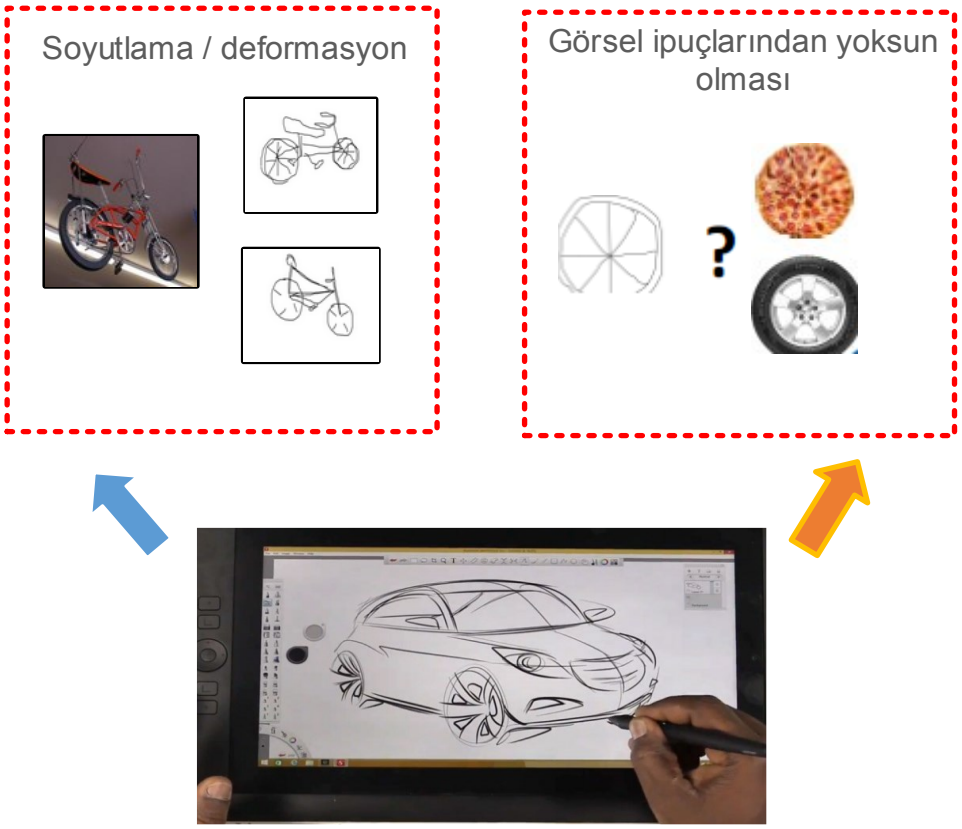
Kavramların sezimi, görsel kavramların ortaya çıkışındaki büyük karmaşıklık ve değişkenlik nedeniyle çok zor bir görevdir. Özellikle, farklı hedef alanlara uygulanan öğrenilmiş kavram modellerinin genelleme kabiliyeti kavram sezimleme alanında ciddi bir sorundur. Çünkü bazı durumlarda anlamsal kavramların görsel görünümü, ilgili resim veya video kaynağının alanına güçlü bir şekilde bağlıdır. Örneğin televizyon haberleri ile kullanıcı tarafından üretilen YouTube videolarının arasındaki fark kolayca görülebilir.



Şekil 1.3 Kavram sezimi

Çizim tanımada, ana zorluklardan biri çizimlerin alan ve zamana göre kısıtlı olmamasıdır. Ayrıca, soyutlanmış / deformasyon olan veya görsel ipuçlarından yoksun olan çizimlerin tanımlanması da göz ardı edilememektedir (Şekil 1.4). Çizimler eğitim, karikatür, eğlence (oyun oynama), yüz tanıma sistemleri gibi birçok alanda kullanılmaktadır. Bu nedenle, çizim tanıma için otomatik ve bilgisayar tabanlı yöntemlere ihtiyaç duyulmaktadır. Günümüzde akıllı telefonlar, artan bilgi işlem gücü sayesinde günlük yaşamımızda geniş bir kullanıma sahiptir ve çizim tanıma için akıllı telefonlar ideal bir platform olarak görülmektedir. Ayrıca, akıllı telefonların yaygın kullanımı nedeniyle, üniversite, okul ya da plaj gibi yerlerde insanlar, çizim tanıma uygulamasını kullanma ihtiyacı duymaktadırlar.

Bu çalışmadaki ana motivasyonumuz, çizim ve video tanıma probleminde CNN mimarilerinin farklı katmanların tanıma kabiliyetini kullanmaktır. Safadi et al., [66] çalışmasında, CNN mimari katmanlarından (soft-max katman öncesi) elde edilen özniteliklerin CNN-DVM iletim yöntemi ile kullanımından, soft-max sınıflandırmasına göre daha yüksek başarım elde etmişlerdir. Bu amaçla, güçlü CNN mimarilerinin (AlexNet (Krizhevsky et al., [1]), VGG19 (Simonyan et al., [2]), ResNet101 (He et al., [4]), GN-Triplet (Sangkloy et al., [22]) vb.) farklı katmanlarından elde edilen öznitelikler ile farklı kaynaşım yöntemleri araştırılmıştır. Son olarak, akıllı telefonlar için istemci-sunucu uygulamasını temel alan ve CNN öznitelikleri ile birleşim yöntemleri içinde en iyi performans gösteren çizim tanıma uygulaması geliştirilmiştir.



Şekil 1.4 Çizim tanıma problemleri (Sangkloy et al., [22])

1.6 Tez Organizasyonu

Bu tezin organizasyonu şöyledir; çalışmada kullanılan temel bilgiler Bölüm 2'de anlatılmaktadır. Video ve çizim tanıma ile ilgili geçmiş çalışmalar Bölüm 3, kullanılan yöntemler Bölüm 4, geliştirilen uygulamalar Bölüm 5, Bölüm 6'ta elde edilen deneysel sonuçlar yer almaktadır. Bölüm 7'de ise sonuçlar ve gelecek çalışmalar sunulmaktadır.

1.7 Katkılar

Bu tez çalışmasındaki amaç, video ve çizim verileri üzerinde, farklı CNN mimarilerini üzerinden elde edilen öznitelik performanslarının kıyaslanması ve etkin bir şekilde kaynaşım yöntemleri ile video ve çizim tanıma başarımlarının artırımını sağlamaktır. Aynı zamanda eğitim maliyeti ve sınıflandırma performansı göz önünde bulundurulmuştur. Çalışmalar sırasıyla TU-Berlin (Eitz et al., [14]), Sketchy (Sangkloy et al., [22]) ve Trecvid 2013 SIN veri kümeleri üzerinde gerçekleştirilmiştir.

Bu çalışmanın katkıları aşağıdaki maddelerde sunulmaktadır:

- Popüler olan çizim ve video veri setlerinde CNN mimarilerin öznitelik performans değerlendirmesi.
- Etkin kaynaşım yöntem analizlerinin gerçekleştirilmesi.
- Eğitim ve hesaplama maliyetini azaltmak için (öznitelik vektör boyutların büyük olması) boyut indirgeme yöntemlerin uygulanması.

Bu tezde aşağıda sunulan yayın çalışmaları yapılmıştır:

- Boyaci, E., Sert, M., Feature–level fusion of deep convolutional neural networks for sketch recognition on smartphones, In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE2017), 8-10 Ocak, Las Vegas, Nevada- USA, s.485-486, 2017.
- Boyaci, E., Sert, M., Video Classification Based on ConvNet Collaboration and Feature Selection, IEEE 25th Signal Processing and Communications Applications Conference (SIU 2017), Antalya, Turkey, s.Tbd.
- Boyaci, E., Sert, M., International Journal of Internet Technology and Secured Transactions (IJITST), 2017 (Küçük revizyon alınmıştır).

2 İLGİLİ ÇALIŞMALAR

2.1 Video Kavram Sınıflandırma ile İlgili Çalışmalar

Video kavram sınıflandırma problemi üzerinde bir çok çalışma yapılmaktadır. Bu çalışmalar global, yerel ve CNN tanımlayıcılar olmak üzere üç başlık altında incelenmektedir.

2.1.1 Global tanımlayıcılar

Renk, doku veya kenar histogramları [51; 52; 53], tutarlılık vektörleri (coherence vectors) (Vailaya et al., [54]), korelogramlar (Liu et al., [55]), bant geçiren filtrelerden dokular (Manjunath et al., [56]) ve renk momentleri (Yanagawa et al., [57]), büyük ölçekli görüntü ve video koleksiyonlarını sınıflandırmada sıkça kullanılmıştır [57; 58]. Görüntüdeki uzamsal bilgileri daha iyi kullanmak için, düzen histogramı ve çoklu çözünürlük (multi-resolution) histogramı (Hadjidemetriou et al., [59]), ve bağlam içinde kullanılmış histogram (Ni et al., [60]) gibi öznitelikler geliştirilmiştir.

Global tanımlayıcıların en büyük avantajı, basitlik ve verimliliktir. Nedenleri bölge bölütleme veya nesne çıkarmanın gerekli olmamasıdır. Bu nedenle, genel tanımlayıcılar, özellikle çok sayıda görüntüyü içeren senaryolarda, görüntü sınıflandırması için hala yaygın olarak kullanılmaktadır. Global görsel özelliklerin en büyük dezavantajı, görüntülerdeki nesnelerin tek tek modellenememesidir. Sonuç olarak, genel görsel özellikler, "dog", "beach" gibi nesne yönelimli kavramların sınıflandırılmasında tatmin edici bir performans sunmayabilir [59; 60; 65].

2.1.2 Yerel tanımlayıcılar

Yerel tanımlayıcılar görüntü sınıflandırması ve nesne tanıma için sıkça kullanılmaktadır. Genellikle yerel bir tanımlayıcı, yerel bir ilgi noktasına merkezlenen bir yerel bölge veya yama (patch) çıkartmaktadır. Yerel ilgi noktası farklı olan ve genellikle yoğunluk, renk ve doku gibi belirli görüntü özelliklerinin bir değişikliği ile ilişkilendirilen bir görüntü modelidir (Jiang, [65]).

Geçmiş çalışmalarda yerel tanımlayıcılar kullanılan görüntü sınıflandırma performansları, kelime kümesi (BoW- Bag-of-Words) şeklinde yerel tanımlayıcıların çoklu tiplerini kullanarak elde edilmektedir.

Chang et al., [61], SIFT (Scale Invariant Feature Transform) (Nowak et al., [37]) tanımlayıcıları TRECVID video kıyaslama verilerindeki multimedya kavramlarını tespit etmek için kullanmışlardır. Dong ve Chang [62], çekirdeğin zamansal akışların BoW gösterimine dayalı olarak tanımlandığı, çok seviyeli zamansal eşleme kullanılarak hesaplanan benzerlik metriğine sahip olan haber videolarındaki kapsamlı jenerik olayları saptamak için çekirdek temelli ayrımcı sınıflandırmayı kullanmaktadır. Zhou et al., [63], her video kayıtında geçici eşleme olarak kodlanan SIFT-BoW tabanlı bir çerçeve önermektedir. Bütün bu yöntemler, videolarda zor olan kavram seziminde başarılı sonuçlar vermiş fakat daha etkin öznitelik çıkarma araştırmalarına devem edilmektedir.

2.1.3 Derin CNN tanımlayıcılar

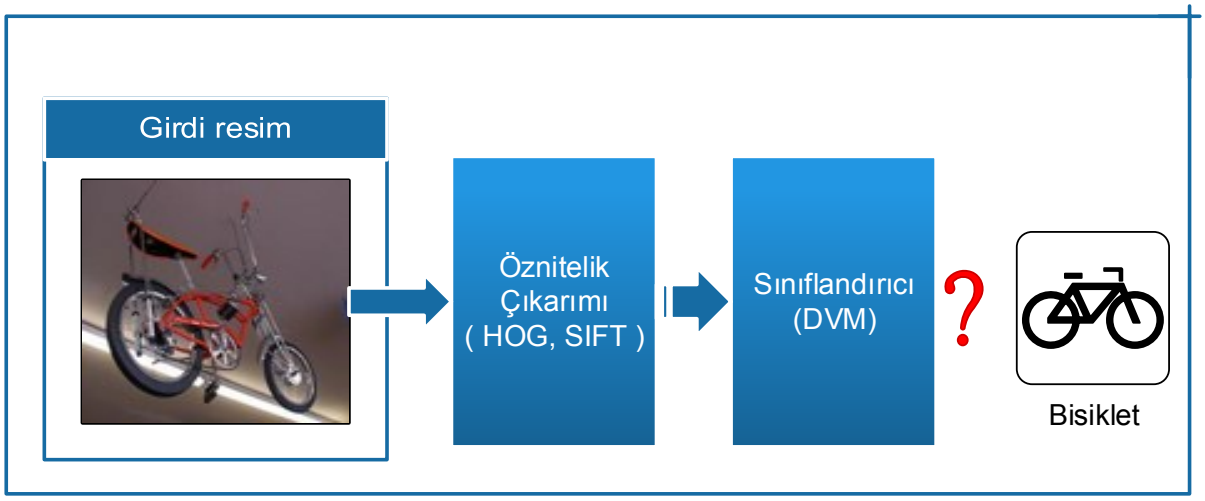
Son yıllarda, CNN'ler bilgisayar görü uygulamaları içerisinde büyük performans getirileri sağlamıştır. Bu başarının çoğunun temel nedeni iki faktördür. Bunlar, paralel işlem mimarileri ve daha geniş görüntü veri setlerinin kullanılabilirliğidir [41; 22; 8]. ILSVRC gibi büyük boyutlu veri setlerinin oluşturulması derin evrişimsel ağlar kurmayı mümkün kılmıştır. Ayrıca, daha yüksek hesaplama kaynaklarının erişilebilirliği, araştırmacıların daha fazla parametre ile daha derin ağlar tasarlamalarına izin vermiştir. Örneğin, ILSVRC yarışmasında 2014'teki en iyi performans gösteren ağlarda 100 milyondan fazla parametre kullanılmıştır. Bu iki faktörün daha iyi eğitim algoritmaları ile birleştirilmesi (Zha et al., [41]), hemen hemen her bilgisayar görü alanında en gelişmiş sonuçları veren oldukça başarılı mimarilere yol açmıştır.

CNN'lerin en önemli özelliklerinden biri genelleme kabiliyetidir. Tamamen farklı bir veri setine veya uygulama alanına önceden eğitimli bir ağ uygulanabilmektedir. Ayrıca performansı yüksek başarımlar elde etmek mümkündür. CNN'ler, resim sınıflandırması, nesne algılama, duygu sezimi, olay tespiti, hareket tanıma, yüz tanıma, trafik işareti tanıma, ve kanser tespit gibi bunlarla sınırlı olmamak üzere çeşitli görüntü işleme görevlerine uyarlanmıştır [10; 11].

CNN mimarilerinin video sınıflandırma alanına uygulanması da kapsamlıdır. CNN modellerinden elde eldilen özniteliklerin, birleşme tekniklerini kullanarak video kavramları anlamsal olarak sınıflandırılabilmektedir (Ergun et al., [20]).

2.2 Çizim Tanıma

Çizim tanıma işleminin ana amacı, belirli bir veriyi daha önceden tanımlanmış sınıflar arasında uygun bir şekilde sınıfa atama işlemidir. Bu sınıflandırmayı gerçekleştirim çalışmaları, verilen çizimlerden kullanışlı ve etkin öznitelikler ile olmaktadır. Geçmiş çalışmalarda, çizim tanımada en yaygın olarak kullanılan düşük seviye gösterimleri HOG (Histogram of Oriented Gradient) (Dalal and Triggs, [35]), GIST (General Iterative Shrinkage and Thresholding) (Oliva and Torralba [36]), SIFT ve kelime kümesi (BoW) tabanlı yerel öznitelikler ile sınıflandırma işlemleridir ve genel yaklaşım Şekil 2.1' de gösterilmiştir.



Şekil 2.1 Resim anlamada genel yaklaşım

Son araştırmalarda [15; 17; 43] insan çizim veri kümelerindeki uçak, martı gibi çizim etiketlerini tahmin etmek için çeşitli girişimlerde bulunulmuştur.

Çoklu çekirdek öğrenme (MKL) ve Balıkçı Vektörleri (FV), yerel özelliklerden (HOG ve SIFT gibi) ve BoW temsiline göre, daha yüksek performans elde edilmiştir. Li et al., [43] bütüncül yapıyı yakalamak ve en gelişmiş tanıma performansını elde etmek için yeni bir yapılandırılmış sunum taslağı önermişlerdir. Bir kaç öznitelik birleşimine dayalı MKL - DVM kullanarak %65.81 oranında başarım elde etmişlerdir. Fisher Vector yöntemi (Schneider et al., [17]) %68.9'luk bir doğruluk seviyesine ulaşmakta ve MKL yönteminden daha iyi performans göstermektedir. Bununla birlikte, FV yöntemi, daha yüksek boyutsallığına bağlı olarak bellek karmaşıklığının dezavantajına sahiptir.

Son yıllarda, Erişimsel Sinir Ağı (CNN), görsel-işitsel veriden anlam bilgilerini öğrenmede mükemmel performans sergilemiştir [19; 7; 1; 2]. Bu başarı, CNN mimarilerinin genelleme yeteneği ile açıklanabilir.

Alexnet (Krizhevsky et al., [1]), VGG19 (Simonyan et al., [2]), GN-Triplet (Sangkloy et al., [22]) gibi önceden eğitilmiş CNN modellerini farklı veri setlerine veya uygulama alanlarına uygulamak ve yüksek başarımlar elde etmek mümkündür. Bu, fotoğrafik resimler üzerinde önceden eğitilmiş CNN modellerinin çizim tanıma problemine uygulanabilir olup olmadığı sorusunu gündeme getirmektedir. Ayrıca, CNN'lerin dahili katmanları, farklı anlamsal seviyeler taşır ve bu katman özellikleri, tanımayı gerçekleştirmek için bir sınıflandırıcı tarafından kullanılabilir.

Mevcut ImageNet veri kümesi ile önceden eğitim almış CNN mimarilerini (örneğin AlexNet ve VGG19) çizim tanıma probleminde farklı katman özelliklerini birleştirerek kullanmaya yönelik bir girişim, veri tanımadaki performansları oldukça başarılıdır [13; 25; 28]. Bu tez çalışmasında, analizlere dayalı olarak AlexNet ve VGG19 mimarilerinin FC6 ve Pool5 katmanları kullanmış ve bağımsız CNN katman özelliği doğruluğu %69.175 oranında geliştirilmiştir. Guo et al. [34] çizim tanıma problemi için Sketch-20K veri setinde %56'lık bir doğruluk elde etmişlerdir. FD-SIFT ve BoW gösterimleri birleştirerek birleşik şekil göstergelerini (FD-SIFT + BoW) DVM ile eğitim yapmışlardır. Ayrıca, ImageNet'de AlexNet, VGGNet ve GoogLeNet olarak adlandırılan üç evrişimsel sinir ağlarını kullanarak Sketch-20K üzerinde ve ek veri ile sonuç almışlardır. Sonuç olarak %77.6'lık bir performans sonucu elde etmişlerdir. Bununla birlikte, harici veri özellikleri hakkında bilgi verilmemiştir.

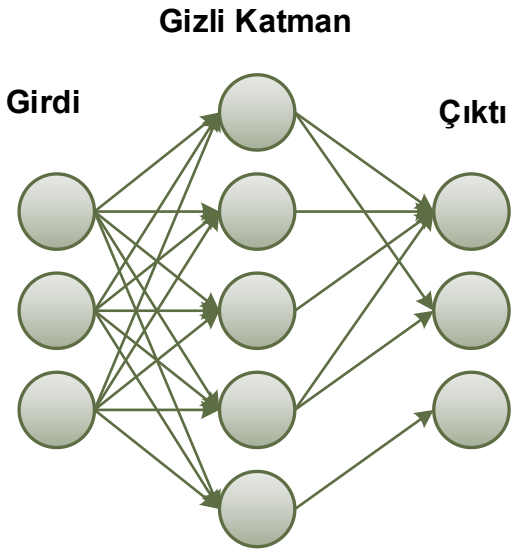
Bilgisayar sistemlerinin kullanımı masaüstü/diz üstü bilgisayardan mobil cihazlara geçtiğini göz önüne alarak akıllı telefonlarda çizim tanıma için etkili algoritmalara ihtiyaç duyulmaktadır. Bu yönde Quick Draw! adlı çizim tanımı uygulaması geliştirmişlerdir (Jongejan et al., [29]). Bu, makine öğrenimi ile oluşturulmuş web tabanlı bir çizim oyunudur. Özellikle, uygulama kullanıcıların sinir ağı kullanarak neler çizdiğini tahmin etmeye çalışır. Tseng et al., [27] düşük hafıza tüketiminden dolayı mobil cihazlarda kullanılabilen bir çizim tabanlı görüntü alma aracı ayrıntılı olarak açıklanmıştır. Ham görsel tanımlayıcıları küçültmek için görsel karma bitlerinden faydalanmayı önermişlerdir. Yüksek boyutlu mesafe dönüştürme (DT) özelliklerini kullanmışlardır. Xiao et al. [38], akıllı telefonlarla çekilen taslak

görüntüleri PowerPoint'te sayısal akış şemalarına dönüştüren bir PPTLens sistemi önermektedir.

Onların inme çıkarma yöntemi beyaz tahtanın veya kağıdın kenarlarını tanımlamakta ve daha sonra kırpılmakta ve düzeltilmektedir. Son olarak, görüntü Stroke Width Transform ve Markov Random Field (MRF) optimizasyonunu kullanarak ikili biçimde gösterilir.

3 TEMEL BİLGİLER VE YARARLANILAN ARAÇLAR

Çok katmanlı Algılayıcı (MLP - Multi-Layer Perceptron), ileri-beslemeli yapay sinir ağlarının en yaygın türlerinden birisidir. MLP, giriş katmanı, çıktı katmanı ve bir veya daha fazla gizli katmandan oluşur. MLP'nin her katmanı, önceki ve sonraki katmanlardan gelen nöronlarla yönlü olarak bağlantılı olan bir veya daha fazla nöron içerir. Şekil 3.1'deki örnek, üç girdi, iki çıktı ve beş nöron gizli katmanı olan üç katmanlı bir algılayıcıyı temsil etmektedir.



Şekil 3.1 Çok Katmanlı Algılayıcı örneği

CNN'ler, biyolojik olarak ileri-besleme tipi yapay sinir ağı mimarileri olup, çok katmanlı algılayıcıların (MLP) seçenekleri olarak sınıflandırılabilir. Son zamanlarda, derin CNN'ler, farklı alanlardaki birçok bilgisayarla görü görevlerinde başarılı sonuçlar vermektedir. CNN, 1980'lerin başında tanıtıldı ve el yazısı, rakam tanıma gibi basit ve küçük görü tanıma görevlerini çözmek için uygulanmıştır (LeCun et al., [47]). Ayrıca CNN mimarileri otomatik yüz tanıma sistemleri, video kavram sezimi gibi bilgisayar görünün birçok alanında uygulanmaktadır [9; 18].

Bu bölümde video kavram sınıflandırması yöntemlerinde kullanılan evrişimsel sinir ağ yapısı ve içerdikleri katmanlar hakkında temel bilgiler sunulmaktadır.

3.1 Evrişimsel Sinir Ağlar

Adından da anlaşılacağı gibi sinir ağları, beyin yapısından sonra modellenen bir makine öğrenme tekniğidir.

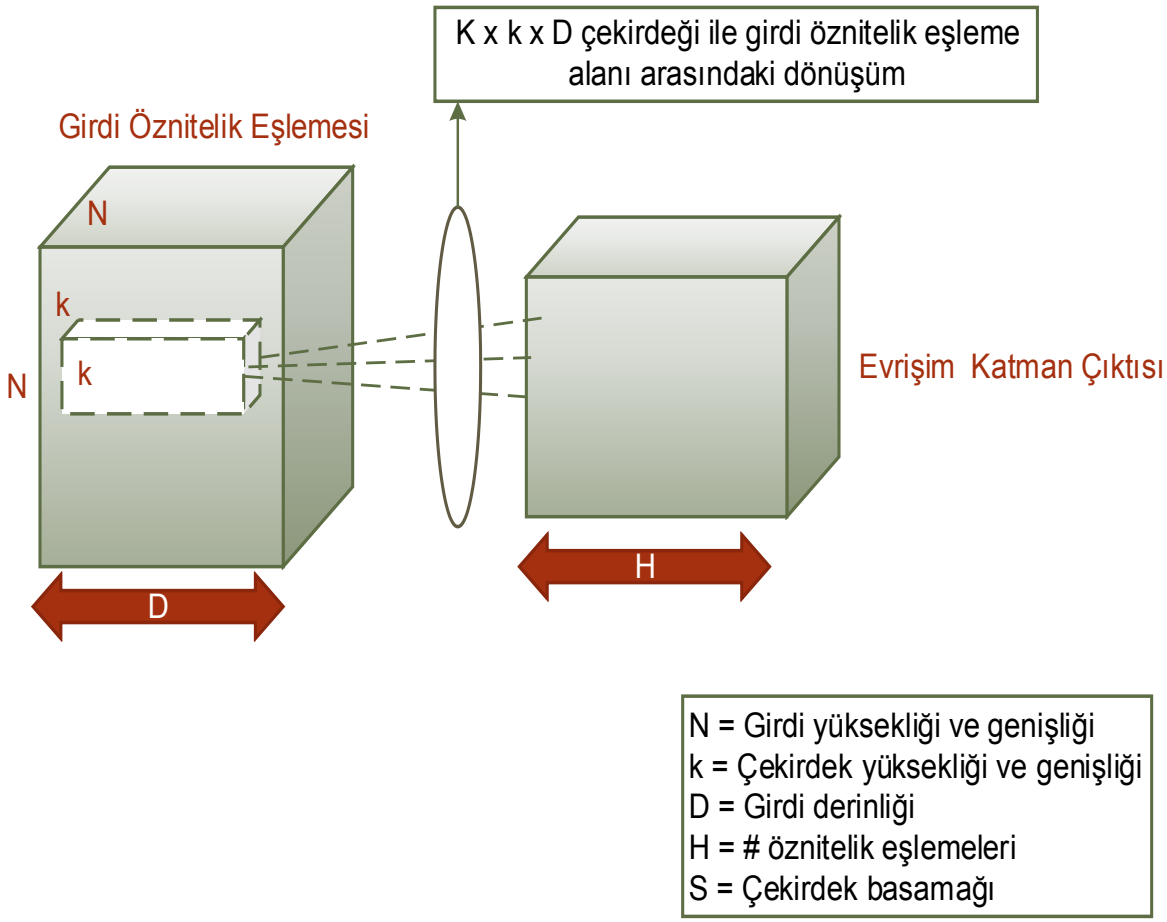
Nöron adı verilen öğrenme birimlerinin bir ağından oluşur. Bu nöronlar, girilen sinyallere (örneğin bir uçağın resmi) karşılık gelen çıkış sinyallerine (örn. "Uçak" etiketi) nasıl dönüştürüleceğini öğrenecek ve otomatik tanıma temelini oluşturacaktır. Geleneksel sinir ağlar gibi, CNN öğrenilebilir ağırlık ve bias içeren nöronlardan oluşmaktadır. Her bir nöron bazı girdileri alır, bir nokta ürünü yapar ve isteğe bağlı olarak doğrusal olmayan bir şekilde takip eder. Bütün olarak CNN, bir ucundaki ham resim piksel değerlerinden sınıflandırma skorlarına kadar devam eden bir fonksiyon olarak da ifade edilebilir.

CNN ayrıca öğrenme modelini eğitmek için optimizasyon yöntemleri ile birlikte öğrenme ağırlıklarını ayarlamak için ileri ve geriye geçişten oluşan geri yayılımı kullanmaktadır. Mimarilerin sonunda bulunan ve bir sonlandırma fonksiyonu olan soft-max, tam bağlı katmandan sonra eklenmektedir.

Derin CNN modeller temel olarak dört farklı katman tipi olan evrişim, veri birleştirme, doğrusal olmayan ve tam bağlı katmanlara sahiptirler.

3.1.1 Evrişim katmanları

Evrişim işlemi girdinin farklı özelliklerini çıkarır. İlk evrişim katmanı, kenarlar, çizgiler ve köşeler gibi düşük seviye özellikleri çıkarır. Üst düzey katmanlar daha üst düzey özellikler çıkarır. Şekil 3.2, CNN'lerde kullanılan 3 boyutlu evrişim işlemini göstermektedir. Girdi, *N* yükseklik ve genişliğinde ve *D* derinliğinde ise, *N x N x D* boyutundadır ve her biri *k x k x D* (k - çekirdek yükseklik ve genişliği) boyutundan ayrı ayrı *H* çekirdekleri ile evrişim işlemi yapılmıştır. Bir girdinin bir çekirdekle evrişimi bir çıkış özelliğini üretir ve *H* (öznitelik eşlemeleri) çekirdekleri bağımsız olarak *H* özellikleri üretir. Girişin sol üst köşesinden başlayarak, her çekirdek soldan sağa, birer birer de birer elemana taşınır. Sağ üst köşeye ulaşıldığında, çekirdek bir öğe aşağıya doğru hareket ettirilir ve çekirdeği bir kerede bir elemandan, soldan sağa kaydırılır. Bu işlem, çekirdek sağ alt köşeye ulaşıncaya kadar tekrarlanır. *N = 32* ve *k = 5* olduğunda, soldan sağa 28 benzersiz konum ve çekirdeğin üstten alta kadar 28 benzersiz konumu vardır. Bu konumlara karşılık gelen çıktıdaki her özellik 28x28 (yani, *(N-k + 1) x (N-k + 1))* elemanları içerecektir. Sürgülü bir pencere işleminde çekirdeğin her konumu için çekirdeğin giriş ve *k x k x D* elemanları *k x k x D* elemanları çarpılarak biriktirilir. Bu nedenle, bir çıktı özelliğinin bir elemanı oluşturmak için, *k x k x D* çarpma işlemi gereklidir (Ovtcharov et al., [48]).



Şekil 3.2 Evrişim süreci

Genişliği *W*, yüksekliği *H* olan bir girdi için evrişim çıktısının tam boyutunu *genişlik* = F_w, *yükseklik* = F_h boyutundaki bir filtre ile hesaplamak için aşağıdaki (3.1) ve (3.2) denklemleri kullanılmaktadır. Burada, S_w ve S_h sırasıyla evrişimin yatay ve düşey eğrisidir ve *P*, görüntünün sınırına eklenen sıfır dolgu miktarıdır.

$$çıktı\ genişliği = \frac{W - Fw + 2P}{\mathrm{Sw}} + 1 \tag{3.1}$$

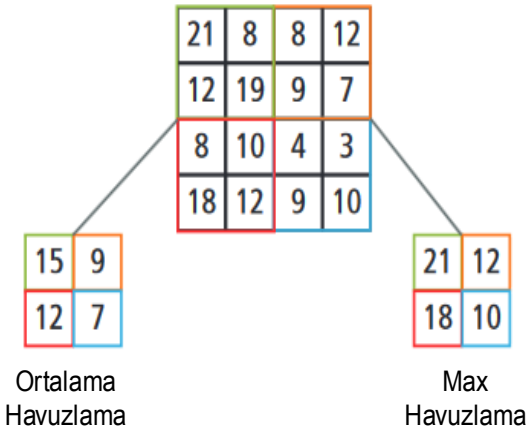
$$çıktı\ yüksekliği = \frac{H - Fh + 2P}{Sh} + 1 \tag{3.2}$$

3.1.2 Veri birleştirme katmanları

Evrişim katmanlarına ek olarak evrişim sinir ağları ayrıca veri birleştirme katmanları içermektedir.

Toplama yani veri birleştirme katmanları evrişim katmanlardan hemen sonra kullanılır. Toplama katmanlarının yaptığı iş evrişim katmanların çıktılarını sadeleştirmektir.

Maksimum (max) ve ortalama veri birleştirme en yaygın olarak kullanılan veri birleştirme yöntemleridir. Ortalama veri birleştirme için bölgedeki dört değerin ortalaması hesaplanır. Maksimum veri birleştirme için, dört değerin maksimum değeri seçilir. Şekil 3.3'te veri birleştirme süreci ayrıntılı bir şekilde irdelemektedir. Girdi 4x4 boyutundadır. 2x2 alt örnekleme için, 4x4 bir görüntü, boyut 2x2 olan örtüşmeyen dört matrise ayrılmıştır. Maksimum veri birleştirme durumunda, 2x2 matristeki dört değerin maksimum değeri çıktı olur. Ortalama veri birleştirme halinde, dört değerin ortalaması çıktıdır (ortalama sonucu en yakın tam sayıya yuvarlanmıştır).



Şekil 3.3 Ortalama ve Max veri birleştirme yöntemleri

3.1.3 Doğrusal olmayan katmanlar

Bir x değeri girdi olarak verildiğinde, ReLU katmanı çıktıyı $x > 0$ ise x olarak, $x <= 0$ ise negatif eğilimli $*x$ olarak hesaplanmaktadır. Negatif eğim parametresi verilmediğinde, standart ReLU işlevinin $max\ (0, x)$ değerine karşılık gelmektedir.

$$f(x) = max\ (0, x) \tag{3.3}$$

3.1.4 Tam bağlı katmanlar

Tam bağlanmış bir katmanda, her bir nöron bir önceki katmandaki her bir nörona bağlanır ve her bağlantının kendi ağırlığı bulunur. Bu, tamamen genel amaçlı bir bağlantı modelidir ve verilerdeki özellikler hakkında hiçbir varsayım yapmaz. Bellek (ağırlık) ve hesaplama (bağlantılar) açısından da çok maliyetlidir.

3.2 Popüler CNN Mimarileri

3.2.1 Alexnet CNN

AlexNet ilk olarak (Krizhevsky et al., [1]) makalesinde görüntü işleme alanına CNN uygulanarak bir ilkin gerçekleştirilmesiyle ortaya çıkmıştır.

Çizelge 3.1 AlexNet ağ mimarisi

Katman Adı	**Katman Boyutu**
Resim girdisi	(227x227x3)=154587
conv1	(55x55x96)=290400
pool1	(27x27x96)=69984
norm1	(27x27x96)=69984
conv2	(27x27x256)=186624
pool2	(13x13x256)=43264
norm2	(13x13x256)=43264
conv3	(13x13x384)=64896
conv4	(13x13x384)=64896
conv5	(13x13x256)=43264
pool5	(6x6x256)=9216
fc6	(1x1x4096)=4096
fc7	(1x1x4096)=4096
fc8	(1x1x1000)=1000
prob	(1x1x1000)=1000

AlexNet 5 evrişim katmanından ve 3 tam bağlı katmandan oluşmaktadır. Bu yapıda sırasıyla ilk, ikinci ve beşinci katmanlardan sonra üç tane max toplama katmanı vardır. Giriş imgesinin boyutu 227x227 büyüklüğüne normalize edilmiştir. İlk evrişim katmanının çekirdek boyutu 11 x 11 'dir. İkinci katmanın ki ise 5x5 boyutundadır. Geri kalan evrişim katmanları için çekirdek boyutu 3x3 'dür. AlexNet 'in son öznitelik boyutunun büyüklüğü 1000'dir. Bu listede olmayan softmax katmanı son katman olarak AlexNet'te bulunmaktadır. Softmax katmanı 1000 ImageNet ve ILSVRC veri

setlerinin anlamsal sınıflandırılması için ayarlanmaktadır. Çizim tanıma için literatürde yüksek başarımlar elde edilmiştir. Çizelge 3.1' de Alexnet nöral ağ mimarisinin katman isimleri ve katman boyutları gösterilmiştir.

3.2.2 VGG19 CNN

VGG19, CNN mimarisinin derinliğini AlexNet'in 8 katmanından 19 katmana çıkarmıştır ve bu da ayrımcı gücünü büyük ölçüde geliştirmiştir.

Çizelge 3.2 VGG19 ağ mimarisi

Katman	Katman Boyutu
conv1_1	(224x224x64)=3211264
conv1_2	(224x224x64)=3211264
pool1	(112x112x64)=802816
conv2_1	(112x112x128)=1605632
conv2_2	(112x112x128)=1605632
pool2	(56x56x128)=401408
conv3_1	(56x56x256)=802816
conv3_2	(56x56x256)=802816
conv3_3	(56x56x256)=802816
conv3_4	(56x56x256)=802816
pool3	(28x28x256)=200704
conv4_1	(28x28x512)=401408
conv4_2	(28x28x512)=401408
conv4_3	(28x28x512)=401408
conv4_4	(28x28x512)=401408
pool4	(14x14x512)=100352
conv5_1	(14x14x512)=100352
conv5_2	(14x14x512)=100352
conv5_3	(14x14x512)=100352
conv5_4	(14x14x512)=100352
pool5	(7x7x512)=25088
fc6	(1x1x4096)=4096
fc7	(1x1x4096)=4096
fc8	(1x1x1000)=1000
prob	(1x1x1000)=1000

VGG19 modeli (Simonyan et al., [2]) bir çok evrişimsel katmandan oluşmaktadır. Bu katmanları üç tane tam bağlı (FC) katman izlemektedir. İlk ikisinin her biri 4096 kanal içermektedir. Üçüncüsü yani son olan FC katmanı 1000 boyutludur ve ILSVRC sınıflandırması yapmaktadır, dolayısıyla her bir sınıf için 1000 kanal içermektedir. Son katman soft-max katmanıdır. Çizelge 3.2'de VGG19 mimarisinin katman isimleri ve boyutları yer almaktadır. Buna ek olarak, çok küçük (3x3) evrişimsel filtre kullanan VGG19 girdi görüntülerindeki ayrıntıları yakalama yeteneğine sahiptir.

3.2.3 GoogleNet CNN

Video kavram sınıflandırmasında kullanılan ikinci CNN modeli olan GoogleNet, 2014 ILSVRC yarışmasında galibiyet kazanmıştır (Szegedy et al., [3]). GoogleNet, AlexNet'den daha derin bir ağ olup, veri birleştirme katmanları hesaba katılmazsa 22 kattan oluşmaktadır. Çizelge 3.3, GoogleNet'in genel katman yapısını özetlemektedir. GoogleNet mimarisinin tüm katman çizelgede ayrıntılı olarak gösterilmemektedir.

Çizelge 3.3 GoogleNet ağ mimarisi

Katman	Katman Boyutu
conv1_1/7x7_s2	(112x112x64)=802816
pool1/3x3_s2	(56x56x64)=200704
pool1/norm1	(56x56x64)=200704
conv2/3x3_reduce	(56x56x64)=200704
conv2/3x3	(56x56x192)=602112
conv2/norm2	(56x56x192)=602112
pool2/3x3_s2	(28x28x192)=150528
inception_3a/1x1	(28x28x64)=50176
inception_3a/3x3_reduce	(28x28x96)=75264
inception layers	... (inception katmanları tekrarlanmaktadır)
inception_5b/pool_proj	(7x7x128)=6272
inception_5b/output	(7x7x1024)=50176
pool5/7x7_s1	(1x1x1024)=1024
loss3/classifier	(1x1x1000)=1000
prob	(1x1x1000)=1000

3.2.4 ResNet 101 ve GN – Triplet CNN

2015 yılında yayınlanan Resnet-101 modeli 101 katmandan oluşmaktadır (He et al., [4]). 2015 yılında yapılan ImageNet yarışmasında nesne sezimlemede %3.57 oranında hata payıyla kazanmışlardır. İnsanların sezimlemede ki hata payının %5 oranında olması nedeniyle büyük bir başarım kazanmışlardır.

Kısaca, AlexNet, VGG19, GoogleNet'den daha yeni ResNet'e, bu mimarilerin evrimindeki bir eğilim ağın derinleştirilmesidir. Artan derinlik, hedef fonksiyona daha iyi yaklaşması için bir ağın kurulmasını sağlar ve daha yüksek ayırt edici güç ile daha iyi öznitelik sunumları üretmektedir.

GN-Triplet CNN mimarisi, 22 katmandan oluşmaktadır (Sangkloy et al., [22]). GN-Triplet mimarisi, AlexNet ve VGG19 ile karşılaştırıldığında nispeten yeni bir mimaridir. GN-Triplet, üçlü ve sınıflandırma kaybı ile eğitilmiş GoogleNet ile tasarlanmıştır.

3.3 Kullanılan Yazılım Kütüphaneleri

Tez çalışması boyunca, uygulamalarımızda çeşitli açık kaynak kütüphanelerinden yararlanılmıştır.

Görsel öznitelik çıkarımı

Temel bileşen analizi (PCA), öznitelik kaynaşım yöntemi için Matlab (MATLAB, [46]) aracı kullanılmıştır.

Derin öğrenme

Özniteliklerin elde edildiği hazır CNN modellerinin uygulanabilmesi için Caffe çatısı kullanılmıştır (Jia et al., [6]).

Sınıflandırıcı

Sınıflandırıcı olarak Destek Vektör Makinesi (DVM) kütüphanesi (Cortes et al., [45]) tercih edilmiştir.

4 VİDEO KAVRAM SINIFLANDIRMA

Görsel kavramların karmaşık ve değişken bir yapıya sahip olmaları nedeniyle kavramların tespit edilmesi zorlu bir görevdir. Özellikle, farklı hedef alanlara uygulanan öğrenilmiş kavram modellerinin genelleme kabiliyeti kavram sezim alanında ciddi bir sorun oluşturmaktadır. Çünkü bazı durumlarda anlamsal kavramların görsel görünümü, ilgili resim veya video kaynağının alanına bağlıdır. Bu duruma örnek olarak, televizyon haberinin ve kullanıcı tarafından üretilen YouTube videolarının dağınıklığı gösterilebilir. Kavram algılama alanındaki büyük bir sorun, başarılı kavram sezimleme sistemlerinin temelini oluşturan güçlü ve etkin özniteliklere sahip olmaktır. Geçmiş çalışmalarda kullanılan yaklaşımlar ağırlıklı olarak SIFT (Scale Invariant Feature Transform) tanımlayıcıları veya hızlandırılmış güçlü öznitelikler (SURF-Speeded-Up Robust Features) gibi anahtar noktalara dayalı yerel görsel öznitelikler üzerine odaklanmaktadır [49; 50].

CNN mimarileri son yıllarda, görsel kavram sınıflandırma ve sezim alanlarına önemli katkılarda bulunmaktadır [13; 25; 28; 44]. Bu nedenle tez çalışmasında video içeriklerinin anlamsal olarak seziminde, CNN mimarilerinin farklı katmanlarından elde edilen özniteliklerin gürbüzlüğü analiz edilerek başarımı yüksek olan kullanılmıştır. Son çalışmalardaki [1; 2; 3; 4; 7] başarımlarından dolayı, AlexNet, VGG19, GoogleNet, ResNet, GN-Triplet CNN modelleri tez çalışmasında kullanılmıştır. Öznitelik çıkarmada AlexNet, VGG19, GoogleNet, ResNet, GN-Triplet modellerin katman sonuçları incelenmiş ve farklı kaynaşım yöntemleri ile başarımlar irdelenmiştir. Ek olarak eğitim maliyeti de göz önünde bulundurulmuş ve boyut indirgeme yöntemi olan PCA uygulanmıştır. Öznitelik etkinliğini artırabilmek amaçlı öznitelik ve skor seviyesinde kaynaşım yöntemleri incelenmiştir.

Tez çalışmamızda önerdiğimiz gürbüz öznitelik analizi, öznitelik ve skor seviyesinde kaynaşım yöntemleri ve tasarlanan DVM sınıflandırıcısı izleyen alt bölümlerde açıklanmaktadır.

4.1 Öznitelik çıkarımı

Bir çok bilgisayar görü probleminde olduğu gibi, öznitelik çıkarma teknikleri görsel tanımada önemli bir rol oynamakla birlikte sınıflandırma doğruluğunu da önemli ölçüde etkilemektedir.

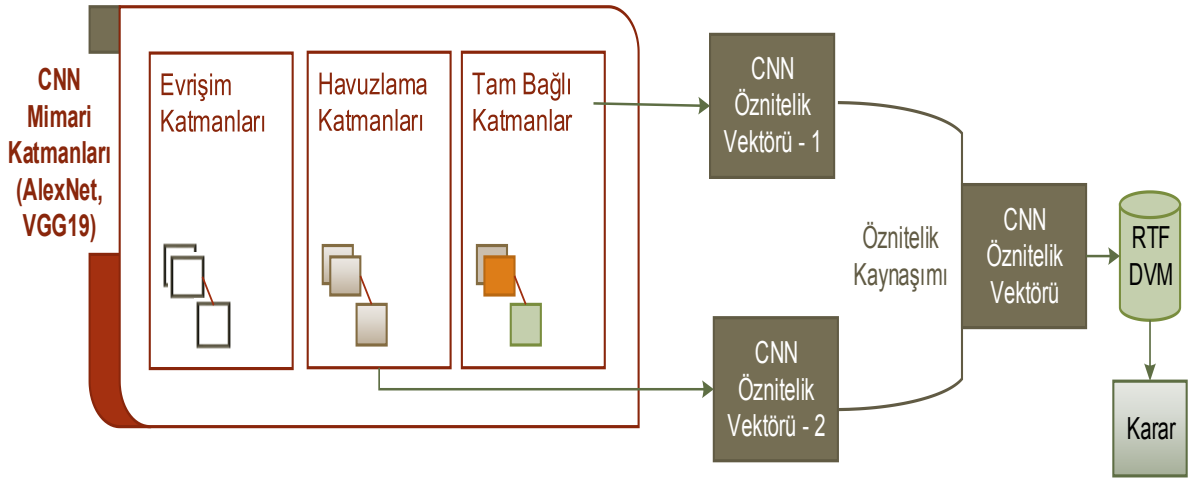
Son yıllardaki araştırmalar göstermektedir ki, CNN mimarileri basit bir görüntü tanımlayıcı olarak kullanılabilmektedir ve bilgisayar görü uygulamalarında iyi performans sonuçları vermektedir (Razavian et al., [33]). CNN mimarisini öznitelik tanımlayıcısı olarak kullanmanın en yaygın yolu, bir görüntüyü mimariye sunmak ve tam bağlantılı katmanlardan birini görüntü özniteliği (tanımlayıcı) olarak kullanmaktır. Bu nedenle, tam bağlantılı ve veri birleştirme katmanlarından elde edilen derin CNN öznitelikleri ile veri kaynaşım yöntemi uygulanmıştır. CNN modellerini eğitmek karışık ve nispeten masraflı bir iş olduğundan, sırasıyla AlexNet, VGG19, GoogleNet, ResNet ve GN-Triplet olarak beş farklı önceden eğitimli ağ kullanılmıştır. Derin öğrenme modellerinden CNN özniteliklerini çıkarmak için Caffe (Jia et al., [6]) çerçevesi kullanılmıştır.

Çalışmada, farklı CNN modellerin son katmanlarından elde ettiğimiz öznitelik vektörlerine *L2* normalizasyonu uygulanmıştır:

$$L_2(x, y) = \sqrt{\sum_{i}^{m} (x_i - y_i)^2} \tag{4.1}$$

Burda, $x = \{x_1, x_2, x_3, \dots, x_m\}$ m boyutlu örnek girdiyi, $y = \{y_1, y_2, y_3, \dots, y_m\}$ ise m boyutlu ileri geçiş çıkışı (forward pass output) ifade etmektedir.

İlk yöntemde, AlexNet ve VGG19 modellerinden elde edilen öznitelikler incelenmiştir (Şekil 4.1). Evrişim, veri birleştirme ve tam bağlı (FC) gibi tipik CNN katmanları, öğrenilen kavramlarla ilgili farklı düzeylerde bilgi taşımaktadır. FC katmanlarının öznitelik olarak kullanılması, evrişim ve veri birleştirme katmanlarına kıyasla görsel kavram tanıma uygulamalarında daha iyi bir doğruluk sağladığı gösterilmiştir [3; 7]. Buna ek olarak, son katmanlar evrişim katmanlarına göre daha az boyuta sahiptir, bu da; geliştirilecek sistemlerin akıllı telefon gibi kısıtlı kaynaklara sahip cihazlardaki bellek ve zaman karmaşıklığı bakımından bir artıdır. AlexNet modelinin FC6 ve VGG19 modelinin Pool5 katmanları daha iyi tanıma hassasiyetine sahip olduğu (Çizelge 6.6) için şemamızda bu katmanlardan elde edilen öznitelikler kullanılmıştır.



Şekil 4.1 Öznitelik düzeyli kaynaşım blok şeması

İkinci çalışmada çizim veri kümesi kullanılarak, bilgisayar görü görevlerindeki başarılarından dolayı AlexNet, VGG19 ve GNTriplet olmak üzere üç gürbüz CNN mimarisi kullanılmıştır. Bu mimarilerin ayrıntıları, öznitelik çıkarma ve veri kaynaşım yöntemi aşağıdaki bölümlerde açıklanmaktadır. Özetlemek gerekirse, öznitelikler sırasıyla CNN mimarileri olan VGG19, AlexNet ve GN-Triplet'in Pool5, FC6 ve Pool5 katmanlarından elde edilmiştir. Bu katmanların boyutları Çizelge 4.1'de verilmektedir. Kaynaşım işleminden önce, farklı CNN modellerinin çeşitli katmanlarından üretilen öznitelik vektörlerine *L2* normalizasyon yöntemi uygulanmıştır.

Çizelge 4.1 Kullanılan CNN mimarileri katmanlarının boyutları

Katman Adı	CNN Model	Boyut	Çıktı Geometrisi
FC6	AlexNet	4096	1x1x4096
FC7	AlexNet	4096	1x1x4096
FC8	AlexNet	1000	1x1x1000
FC6	VGG19	4096	1x1x4096
POOL5	VGG19	25088	7x7x512
POOL5	GN-Triplet	1024	1x1x1024

Örüntü tanımadaki başarımlarından dolayı, CNN mimarisinin soft-max sınıflandırıcısı yerine Destek Vektör Makinesi (DVM) tercih edilmiştir. DVM tasarımı amacıyla, doğrusal ve radyal taban fonksiyonunun (RTF) olmak üzere farklı çekirdek fonksiyonları analiz edilmiştir. RTF çekirdek fonksiyonunun parametre eniyilemesi için ızgara arama (grid search) algoritması kullanılmıştır. RTF denklemi (4.2) eşitliğinde verilmiştir.

$$rbf(x, y) = e^{-\gamma \sum_i (xi - yi)^2} \tag{4.2}$$

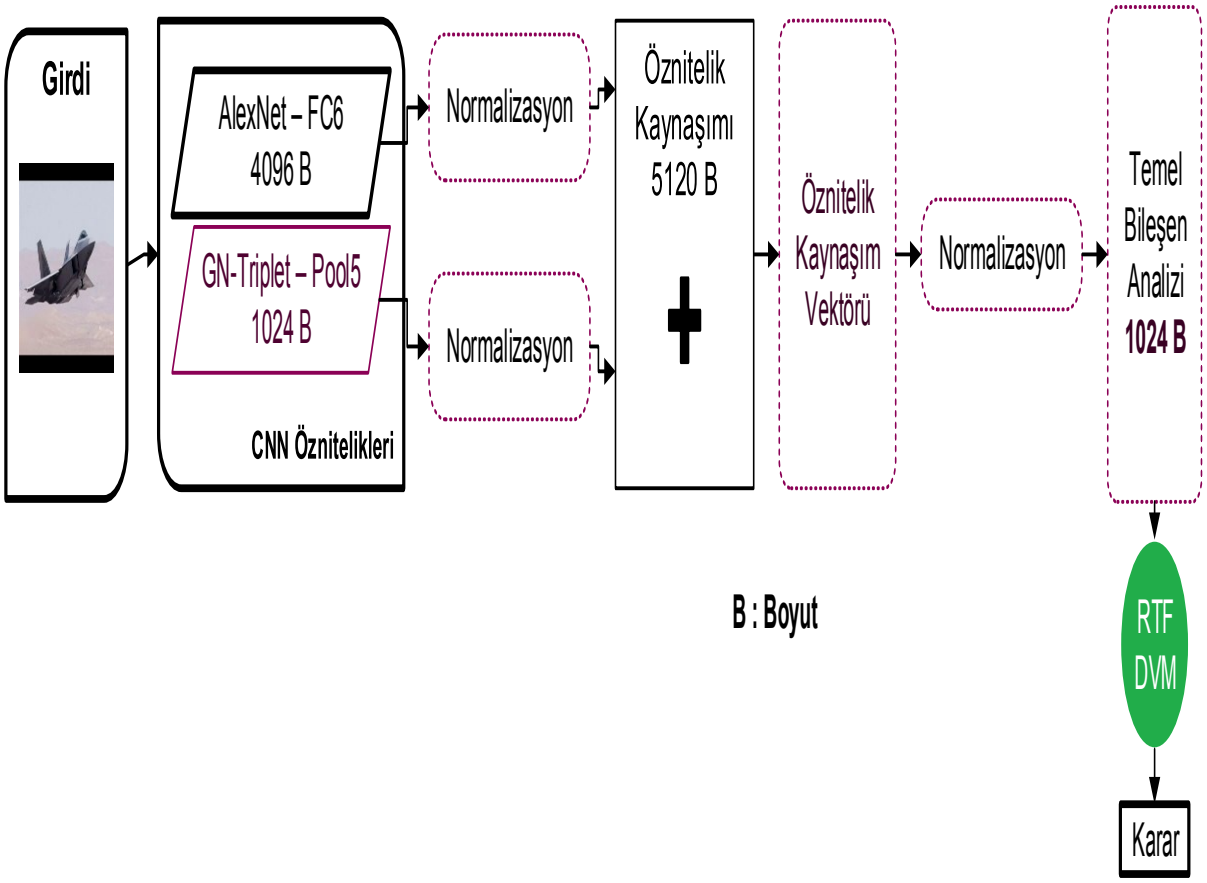
Tez çalışmasında 3. Yöntem olarak, dört farklı CNN mimarisinden elde edilen öznitelikler, öznitelik kaynaşım, Temel Bileşen Analizi (PCA-Principal Component Analysis) ve Ayırtaç İlinti Analizi (DCA) (Haghighat et al., [9]) yöntemleri Trecvid veri seti üzerinde uygulanmış ve elde edilen yeni öznitelik temsilleri ile Destek Vektör Makinesi (DVM) sınıflandırıcısı tasarlanmıştır. Sistemimizin genel şeması Şekil 4.3'de verilmiştir. Öznitelik çıkarım işleminden önce videolar üzerinde çekim sezimleme işlemi yapılmıştır. Bu işlem için Trecvid tarafından açıklanan çekim süre bilgileri kullanılmıştır. Elde edilen çekimlerden, ortadaki imge anahtar çerçeve olarak seçilmiş ve işlemler bu çerçeve üzerinden gerçekleştirilmiştir.

Öznitelik seviyesi kaynaşım stratejimiz Şekil 4.3'de gösterilmektedir. Evrişim, örnekleme ve tam bağlı (FC) gibi tipik CNN katmanları, girdi resim karesi ile ilgili farklı düzeylerde bilgi taşırlar.

Tam bağlı son katmanların doğruluk oranlarının görece yüksek olması (Ergun ve Sert, [7]) ve diğer katmanlara göre daha küçük boyutlu olması sebebiyle, öznitelik çıkarımı modellerin son katmanlarından elde edilmiştir. Farklı CNN modellerin (AlexNet, VGG19, GoogleNet, ResNet101) son katmanları için gerçekleştirdiğimiz analiz sonuçları Çizelge 6.2'de sunulmuştur. Buna göre, RTF çekirdek, doğrusal çekirdek'ten daha başarılı sonuçlar vermektedir. Bu nedenle, sonraki aşamalarda yapılan testlerde de RTF çekirdek kullanılmıştır.

4.2 Öznitelik Düzeyli Kaynaşım

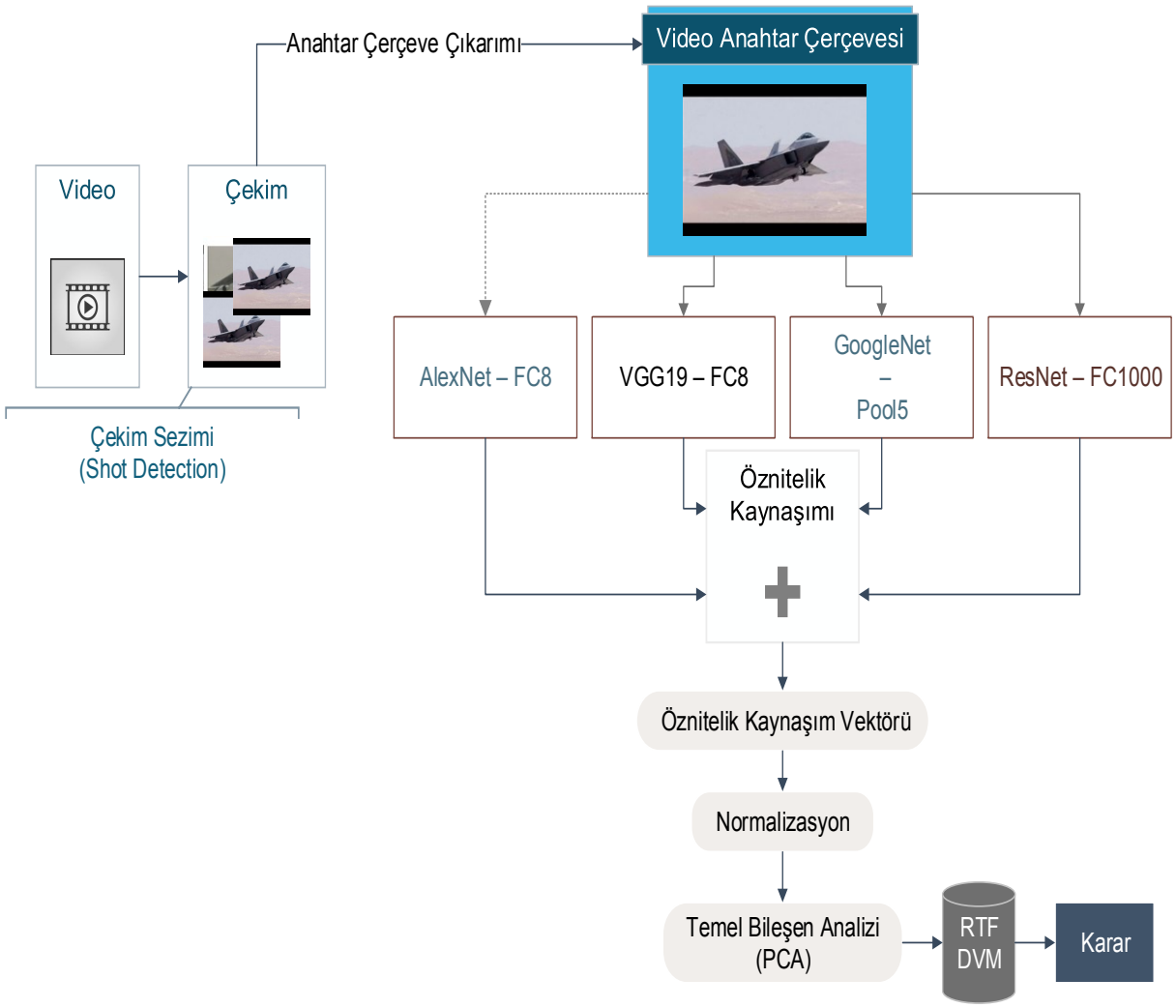
Veri kaynaştırma, birden fazla kaynaktan gelen verilerin işlenerek veya ilişkilendirilerek bir araya getirilmesidir. Veri kaynaştırma öznitelik düzeyinde ve model düzeyinde (late) olabilmektedir. Bir sınıflandırıcının çıktı kararı veya eşleşen değerinden, daha zengin bilgi içermesi dolayısı ile öznitelik seviyesinde kaynaşımın daha etkili olduğuna inanılmaktadır (Ergun vd., [13]). Bu nedenle, öznitelik kaynaşım yöntemlerinden olan DCA (Discriminant Correlation Analysis) ve art arda bağlama işleçleri öznitelik kaynaşım tekniği olarak kullanılmıştır.



Şekil 4.2 Öznitelik seçimine dayalı öznitelik kaynaşım yöntemi

Art arda bağlama işleci, öznitelik vektörlerinin birbirinin ardı sıra eklenmesi olarak uygulanmaktadır. Örneğin, x ve y sırasıyla p ve q uzunluğunda iki öznitelik vektörü ve $||$ kaynaştırma işleci olmak üzere, (4.3) eşitliğinde gösterildiği üzere $(p+q)$ uzunluğundaki z öznitelik vektörü elde edilmektedir.

$$x = \{x_1, x_2, x_3, \dots, x_p\}$$
$$y = \{y_1, y_2, y_3, \dots, y_q\} \quad (4.3)$$
$$z = x \,||\, y$$



Şekil 4.3 Önerilen video kavram sınıflandırma sistemi

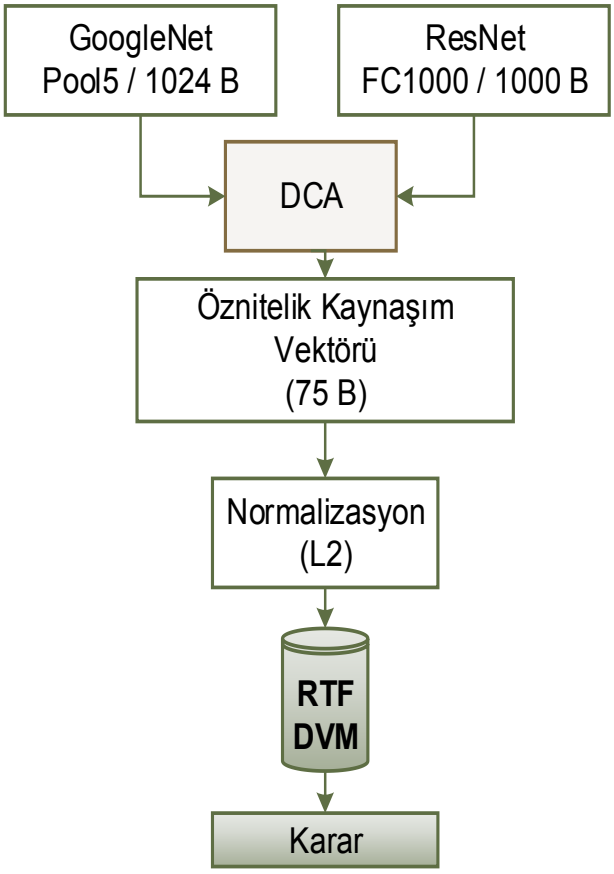
Bu işleç boyut artımına neden olmakla birlikte, daha fazla veri içermesi ve basitliği nedeniyle tercih edilmiştir. Çalışmalarda, öznitelik kaynaşım yönteminde aşağıda belirtilen tasarım kriterleri baz alınmıştır:

- Şekil 4.1'de gösterilen yöntemde, AlexNet ve VGG19 modellerinden başarımlarının diğer katmanlara göre yüksek olması sebebi ile FC6 ve Pool5 katman öznitelikleri kullanılmıştır.
- Şekil 4.2'de gösterilen yöntemde, AlexNet-FC6 katmanı ile GN-Triplet modelinin Pool5 katman öznitelikleri, VGG19 modelinden daha yüksek başarım sağladığından dolayı tercih edilmiştir.
- Şekil 4.3'de son katman öznitelikleri kullanılarak 4 farklı CNN (AlexNet, VGG19, GoogleNet, ResNet) modeli kullanılmıştır.
- Şekil 4.4'te geliştiren yöntemde ise Şekil 4.3'te kullanılan modellerden en

yüksek başarım sağlayan 2 farklı CNN modeli (GoogleNet, ResNet) üzerinde DCA kaynaşım yöntemi uygulanmıştır.

Art arda ekleme kaynaşımı ile boyutu artan vektör için çekim yöntemi olan PCA uygulanmış ve böylelikle boyut indirgeme ve öznitelik etkinliği sağlanmıştır. Elde edilen sonuçlar Çizelge 6.4'te gösterilmiştir.

DCA, klasik ilişkileri öznitelik kümelerinin ilinti analizine dahil eden bir öznitelik düzeyli kaynaşım tekniğidir. Ayrıca DCA, iki özellik kümesindeki çift yönlü ilintilerini en üst düzeye çıkararak, ilintileri ortadan kaldırarak ve ilintileri sınıflar arasında sınırlandırarak etkili bir öznitelik birleşimi gerçekleştirir.



Şekil 4.4 DCA öznitelik kaynaşım yöntemi

Bu yöntem, tek bir yöntemden çıkarılmış farklı öznitelik vektörlerinden ayıklanmış özellikleri birleştirmek için model tanıma uygulamalarında kullanılabilir. DCA'nın öznitelik sınıf yapısını göz önüne alan ilk teknik olması dikkate değerdir.

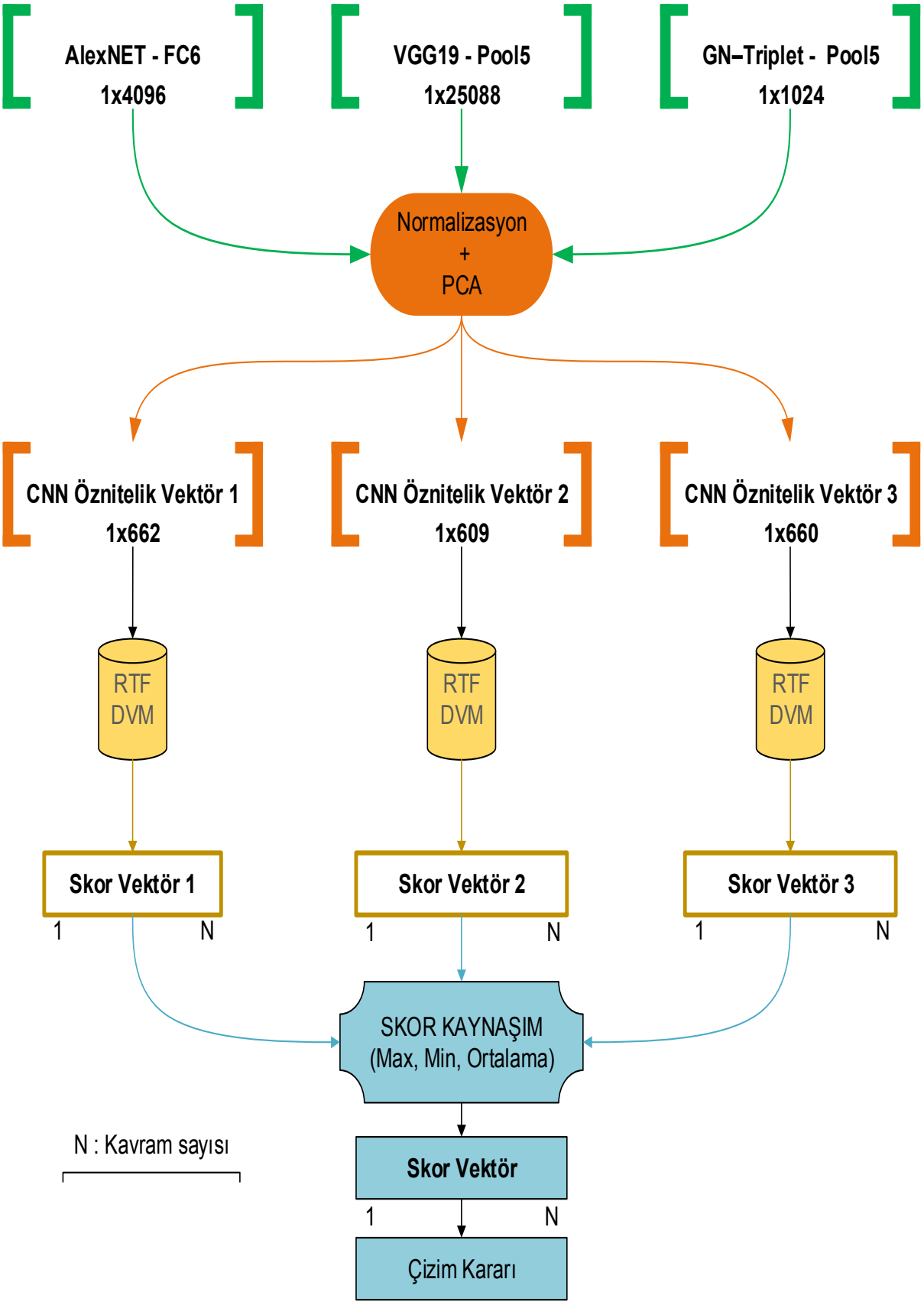
Ayrıca, çok düşük hesaplama karmaşıklığına sahiptir ve gerçek zamanlı uygulamalarda kullanılabilir. Bu avantajlarından dolayı, DCA kaynaşım yöntemi Şekil 4.4'de gösterildiği gibi uygulanmış ve elde edilen sonuçlar Çizelge 6.2 ve Çizelge 6.3'de gösterilmiştir.

4.3 Skor seviyesinde kaynaşım

Öznitelik kaynaşımında, hedef sınıflandırıcı tarafından işlenmeden önce farklı katmanların öznitelikleri entegre edilmiştir. Başka bir deyişle, farklı kaynaklardan elde edilen öznitelikler tek bir öznitelik vektörüyle birleştirilir. Diğer yandan, skor kaynaşımında, bu kaynakları birleştirmeden önce her bilgi kaynağından kavram öğrenme ayrı ayrı gerçekleştirilmiştir.

Skor kaynaşım yöntemi Şekil 4.5'de gösterilmektedir. Kullanılan yöntemde öncelikle en iyi performans gösteren katmanlardan olan AlexNet FC6, VGG19 Pool5 ve GN-Triplet Pool5 öznitelikleri çıkartılmıştır. Elde edilen özniteliklere *L2* normalizasyon ve boyut azaltmayı sağlayan PCA yöntemi uygulanmıştır. Böylelikle AlexNet FC6 katman boyutu 4096 olan vektör boyutundan 662, VGG19 Pool5 için 25088 olan boyut için 609 ve son olarak GN-Triplet Pool5'den elde edilen boyut 1024 iken 660 boyuta indirgenmiştir. Geçmiş çalışmalarda, skor kaynaşım yöntemlerinde soft-max sınıflayıcı katmanını kullanmak çok yaygındır. Bununla birlikte, soft-max katmanı, önceden eğitimli veri kümesi için özel olarak optimize edilmiştir ve büyük miktarda eğitim verilerinin mevcut olmadığı alanlar için iyi olmayabilir. Bu nedenle, yaygın kullanımın aksine, CNN-DVM iletimindeki kaynaşım için DVM çıktısını kullanma önerilmiştir.

DVM sınıflandırıcılarının çıktıları, her katman için skor vektörlerini temsil etmektedir. V değeri $\{v_1, v_2, ..., v_n\}$ değerlerine sahip bir skor vektörü olsun; burada n, çizim veri kümesindeki kavramların sayısıdır ve v_i, i. kavramının frekansı (tf - $term\text{-}frequency$) 'dir. Kaynaşım operatörleri olan max, min ve $ortalama$ kullanarak çıkış vektörleri (v_i ve v_j) üzerinde skor birleştirme işlemi gerçekleştirildi.



Şekil 4.5 Skor seviyesinde kaynaşım yöntemi

Örneğin kaynaşım operatörü olan max operatörü, iki vektörü girdi olarak alır ve bir vektör üretir; burada vektör elemanları, vektörün her bir karşılık gelen elemanının en yüksek değeri olarak seçilir. Kavramsal karar V'nin tf değerlerine göre gerçekleştirilir, burada V'deki maksimum tf değerinin endeksi, kararı tanımlamaktadır.

4.4 Boyut indirgeme ve sınıflandırıcı tasarımı

Öznitelik seviyeli kaynaşım yöntemleri, geç yani model bazlı yöntemlere kıyasla belirli avantajlara sahiptir. Çünkü farklı öznitelik vektörleri bazı modellerin farklı karakteristik özelliklerini sergilemektedir ve bu öznitelikleri birleştirilen bir formda kullanmak, eldeki veriler hakkında etkili ve ayrımcı bilgileri içermektedir.

Bu yöntemde, en iyi performans gösteren CNN mimarilerinden elde edilen özniteliklerin birleştirilmiş ve *L2* normalizasyonu uygulanarak PCA uygulanmıştır. Performans ve doğruluk arasındaki dengeyi temel alarak indirgenen öznitelik boyutu 1024 olarak seçilmiştir. Örneğin, AlexNet ve GN – Triplet modellerinden elde edilen 5120 boyut yerine 1024 boyutlu öznitelik vektörleri elde edilmiştir.

Önerilen sistemler, CNN-DVM iletimi şeklinde tasarlanmıştır. Diğer bir deyişle, seçilen CNN mimarilerinden çıkarılan öznitelikler, kaynaştırılmış ve PCA yöntemiyle boyut indirgemesi yapıldıktan sonra DVM sınıflandırıcıya verilmektedir. DVM algoritması için LibSVM (Chang et al., [40]) kütüphanesi kullanılmıştır. Çok sınıflı sınıflandırma sorununun üstesinden gelmek için bire-karşı-hepsi (OVA) tekniği kullanılmıştır.

DVM'nin çekirdeği olarak radyal taban fonksiyonu (RTF) ve çekirdek parametrelerini optimize etmek için ızgara arama (grid search) algoritması uygulandı.

Bununla birlikte, önerilen sistemlerde kullandığımız birleştirme operatörü gibi stratejilerin bir dezavantajı, birleştirme operatörünün son özellik boyutunu kullanması ve bunun sonucunda öğrenme algoritması için boyutsallık sorunlara neden olmaktadır. Bu sorunun üstesinden gelmek ve aynı zamanda farklı özelliklerini korumak ve sınırlı hesaplama gücü olan cihazlar için önemli bir sorun olan öznitelik boyutunu azaltmak için Temel Bileşen Analizi (PCA) kullanılmıştır.

PCA yöntemi, tanıma, veri sınıflandırma, görüntü sıkıştırma alanlarında kullanılan bir tekniktir. PCA verideki gerekli ve etkin bilgileri ortaya çıkarmaktadır.

Boyutu fazla olan verilerdeki genel özellikleri bularak boyut sayısının azaltılmasını ve verinin sıkıştırılmasını sağlamaktadır. PCA yöntemindeki temel mantık çok boyutlu bir veriyi, verideki temel özellikleri yakalayarak daha az sayıda değişkenle göstermektir. Boyutun indirgenmesi ile verideki bazı özelliklerin kaybedilmesine rağmen kaybolan özellikler veri hakkında daha az bilgi içermektedir.

PCA yaklaşımı sözde kod ile açıklanacak olursa:

PCA Sözde Kod Algoritması

GİRDİ: X ← girdi veri seti matrisi.

ÇIKTI: **y** ← çıktı veri seti matrisi.

X veri setinden sınıf etiketlerini çıkart

kovaryans matrisini hesapla

ortalama vektör $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$

kovaryans matris $\Sigma = \frac{1}{n-1}((\mathbf{X} - \overline{\mathbf{x}})^T(\mathbf{X} - \overline{\mathbf{x}}))$

kovaryans matrisinden özvektörleri(eigenvectors) ve özdeğerleri(eigenvalues) elde et

$\Sigma v = \lambda v$

özvektör v

özdeğer λ

özdeğerleri azalan düzende sırala

k en büyük özdeğerlerine karşılık gelen k özvektörlerini seç

$d \times k$ boyutlu W özvektör matrisi oluştur

X boyutlu bir öznitelik alt uzayını elde etmek için orijinal veri kümesini dönüştür

$y = W^T \times x$

5 UYGULAMALAR

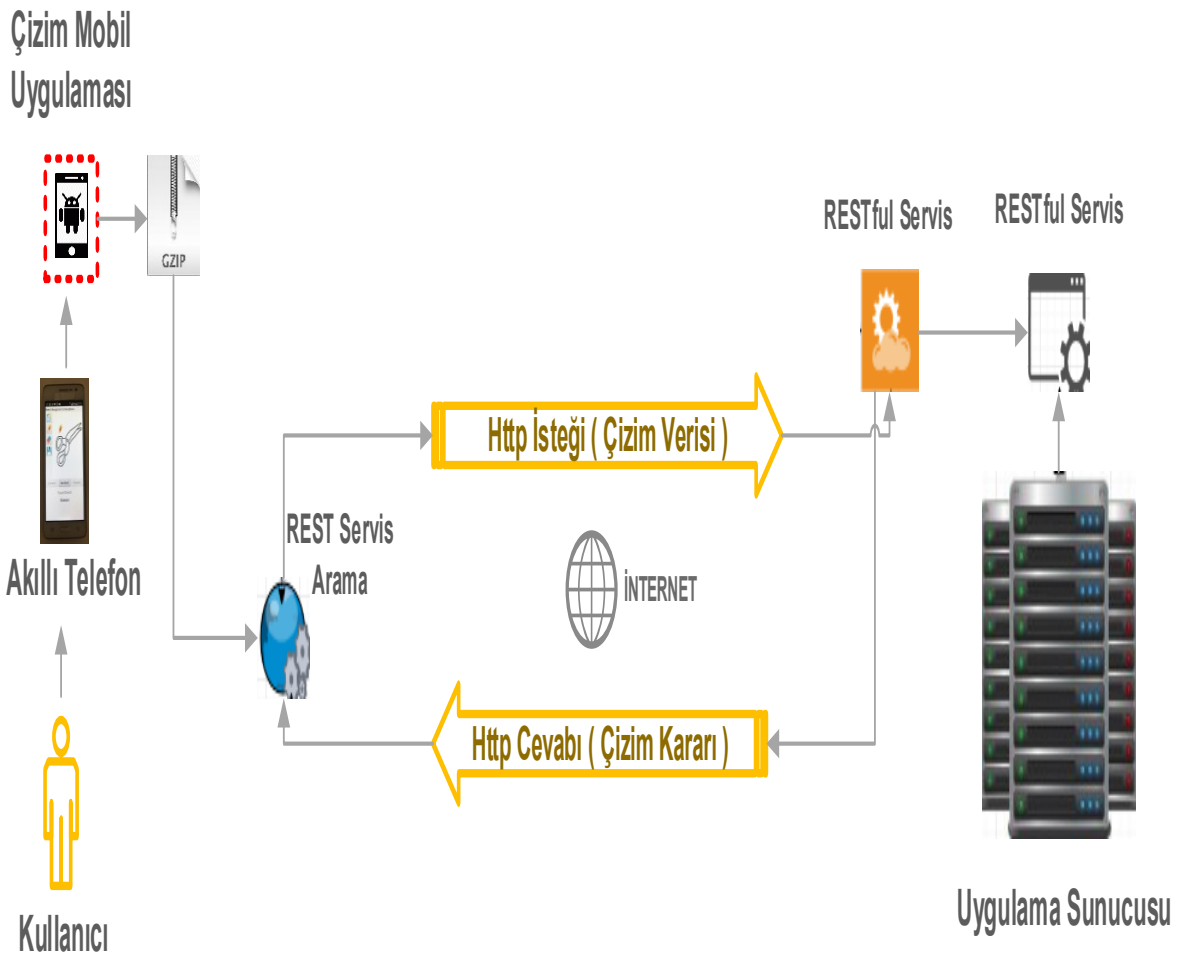
5.1 Akıllı Telefonlar için Servis Tabanlı Çizim Tanıma Uygulaması

Akıllı telefonlar için çizim tanıma uygulaması tasarlanmıştır. Çizim tanıma uygulamasının mimarisi Şekil 4.2'de gösterilmektedir. Sistem iki modülden oluşmaktadır. Bunlar Çizim Tanıma Servis ve Mobil Çizim Uygulaması. Çizim Tanıma Servisi, çizim işleme ve tanıma görevlerini gerçekleştiren ve bir uygulama sunucusu tarafından gerçekleştirilen web servis uygulamasıdır.

Çizim Mobil Uygulaması, kullanıcıdan aldığı çizimi web servis tarafından servis tarafına ulaştırılmaktadır. Çizim Tanıma Servisi, tanımlama görevi sırasıyla, servis önce kullanılan CNN mimarilerinin katmanlarından öznitelikleri çıkarır, *L2* normalizasyonu gerçekleştirir, kaynaşım operatörünü uygular, kaynaştırılmış özelliklerde tekrar *L2* normalizasyonu gerçekleştirir, boyut indirgeme uygulanır (PCA), OVA yöntemi ile daha önce eğitilmiş DVM modellerini kullanarak çizim kavramını öngörür ve sonuçları mobil uygulamaya geri göndermektir. Sonuç olarak, çizim kavramı, Mobil Çizim Uygulama arayüzünde kullanıcıya gösterilmektedir.

Çizim tanıma servis mimarilerinde, iki yaygın olarak kullanılan web servisler Basit Nesne Erişim Protokolü (SOAP) ve Temsili Durum Transferi (REST) gibi servis mimarileridir.

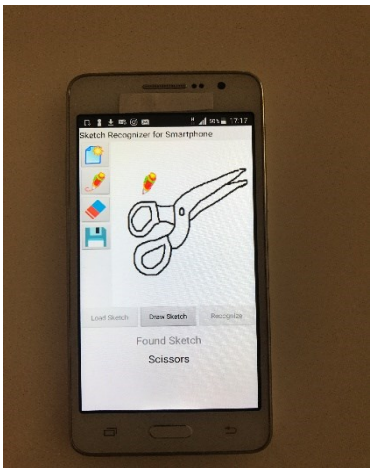
SOAP, uzun süredir web servis arayüzleri için yaygın olarak kullanılsada, REST mimari kullanımı giderek yaygınlaşmaktadır [23; 24]. REST mimarisi, özellikle mobil uygulamalarda SOAP'a göre bazı avantajlara sahiptir; örneğin, SOAP kullanan servislerin değişimi, genellikle istemci tarafında karmaşık bir kod değişikliği anlamına gelmektedir. Buna ek olarak, web servislerinden SOAP istemci tarafı kod üretimi ve uygulaması Açıklama Dili (WSDL) ve XML Şeması Tanımı (XSD) karmaşık olabilmektedir.



Şekil 5.1 Geliştirilen çizim tanıma uygulama mimarisi

Bu nedenle, mobil uygulamada uygulama güncelleme problemi ortaya çıkmaktadır. Ayrıca REST uygulanması ve bakımı daha kolay olan esnek bir mimariye sahiptir. SOAP servisi her zaman XML döndürürken, REST servisi döndürülen verilere göre esneklik sağlar. REST servislerinden gelen veri yükleri için standartı Java Kod Nesne Yazılımı'dır (JSON). JSON yükleri genellikle XML karşılıklarına göre daha küçüktür. Kısaca SOAP veri iletimi için daha geniş bant genişliğine ihtiyaç duymaktadır. RESTful servislerinin JSON biçimindeki verilerle kullanılması, istemci cihaz platformunun iOS veya Android işletim sistemi kullanan mobil cihaz uygulamaları için daha iyi bir seçimdir.

Sonuç olarak, RESTful servisi, çizim tanıma uygulanmasının gerçekleştirilmesinde tercih edilmiştir. Mobil cihazlarda daha sınırlı ağ bant genişliği kullanımını en aza indirgemek için uygulamaların düşük gecikme süresi ile sorunsuz çalışmasını sağlamak için tanıma servisi sunulmaktadır.



Şekil 5.2 Akıllı telefon üzerinde çizim tanıma uygulaması

Son olarak, akıllı telefonlar için istemci-sunucu uygulama mimarisini temel alan en iyi performans gösteren öznitelik seviyesinde kaynaşım şemasını kullanarak bir çizim tanıma uygulaması geliştirilmiştir (Şekil 4.2).

5.2 Video Kavram Sezimi için Web Uygulaması

Video içeriklerini anlamak, sezimlemek ve onu daha erişilebilir hale getirmek için, video içeriğinde bulunan kavramlarının sınıflandırılması gerekmektedir. Kullanıcılar, bu verileri verimli bir şekilde organize etmek ve araştırmak için video kavram sezimleme uygulamalarına ihtiyaç duymaktadırlar. Bu ihtiyaçlar göz önünde bulundurularak, tez çalışmasında, video kavram içeriklerin otomatik sezimi için web uygulaması tasarlanmıştır.

Analiz

Video kavram sezim uygulaması temel olarak kullanıcının video yükleyerek, yüklediği videodan çekim sezimi, anlamsal kavramların sınıflandırılmasını ve sezimini sağlayan bir web uygulamasıdır.

Kullanıcı arayüz gereksinimleri:

- ❖ Kullanıcılar mp4 uzantılı video yükleyebilmelidir.
- ❖ Yüklenen videonun anlamsal sınıflandırılması yapılabilmelidir.
- ❖ Kullanıcı seçtiği kavram ile ilgili sonuçları uygulama içerisinde görebilmelidir.

Tasarım

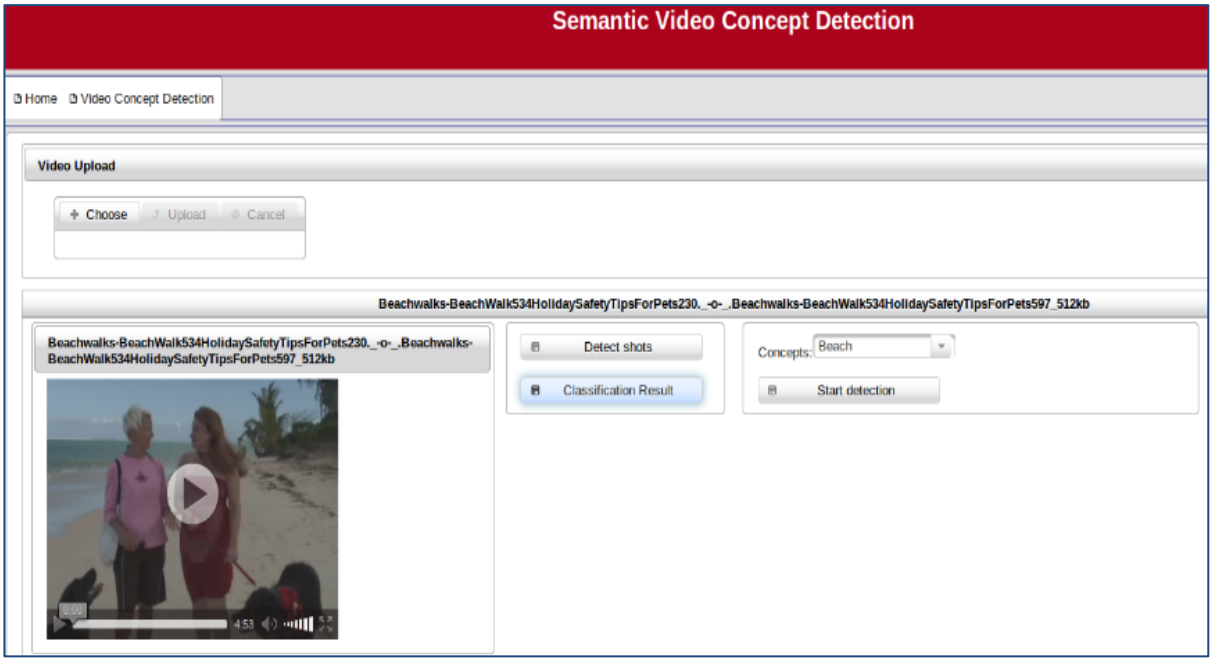
Uygulama mimarisi çok katmanlı Java web uygulaması olarak tasarlanmıştır. Önyüz teknolojisi olarak Primefaces, Icefaces bileşen kütüphaneleri kullanılmıştır. MVC (Model View Controller) tasarımına uygun olarak geliştirim yapıldı.

Video sınıflandırması ve kavram sezimi için bash betikleri oluşturulmuştur. Video çekimlerini çıkarabilmek için Python ile yazılmış betik kullanıldı.

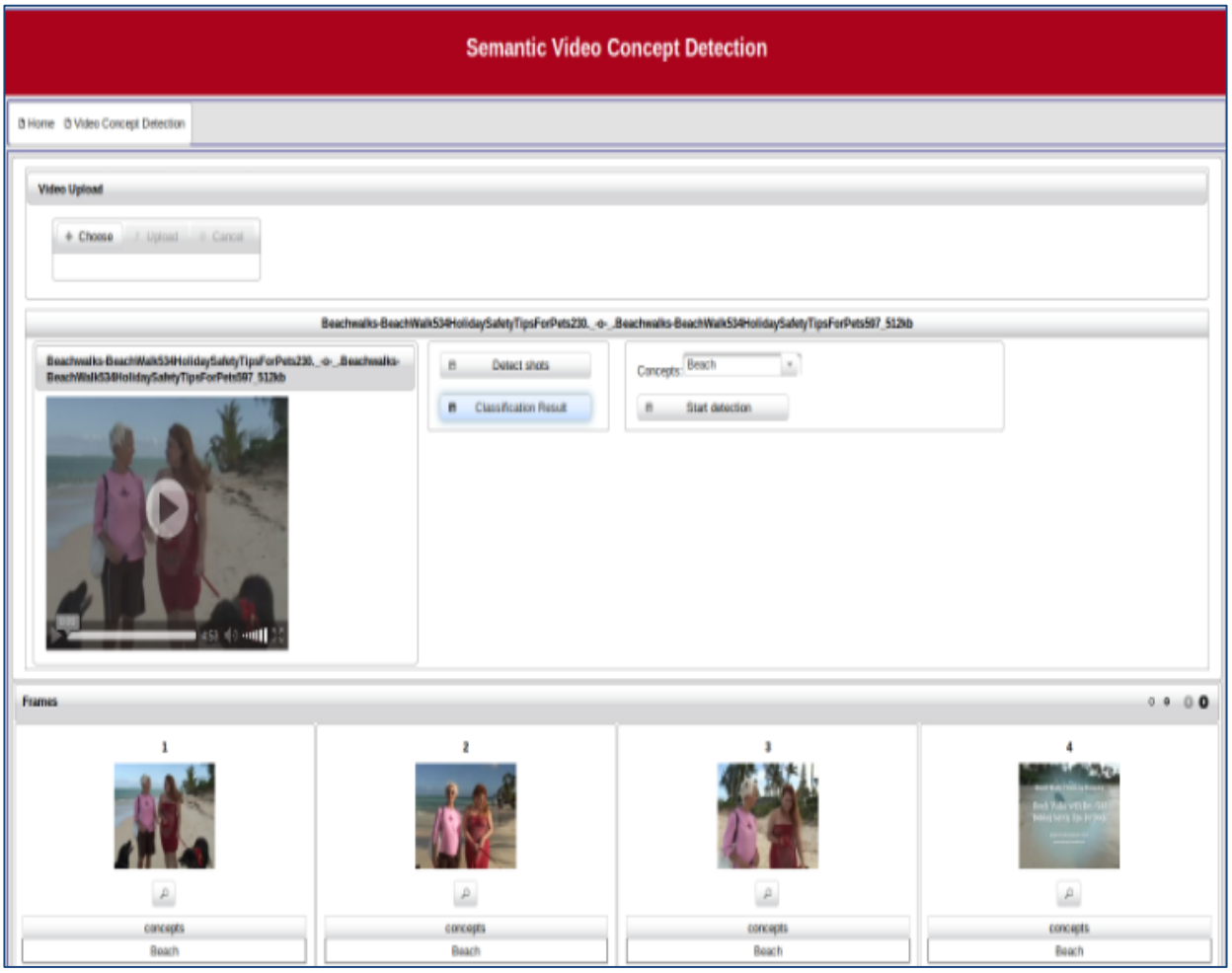
Uygulama içerisinden yüklenen video için bu betiklerin çağırılması ile dosya sisteminde oluşan sonuçlar önyüzde kullanıcıya gösterilecek şekilde tasarım gerçekleştirildi.

Tasarlanan uygulamada kullanıcı 3 farklı işlem yapabilmektedir (Şekil 5.3). Bunlar sırasıyla:

- İşlem yapmak istenen video verisinde çekim sezimlemesi (Şekil 5.4).
- Çekimlere ait anahtar çerçevelerin sınıflandırılması (Şekil 5.4).
- Şekil 5.5'te gösterildiği üzere seçilen kavramı içeren çekim bilgisinin görüntülenmesi.



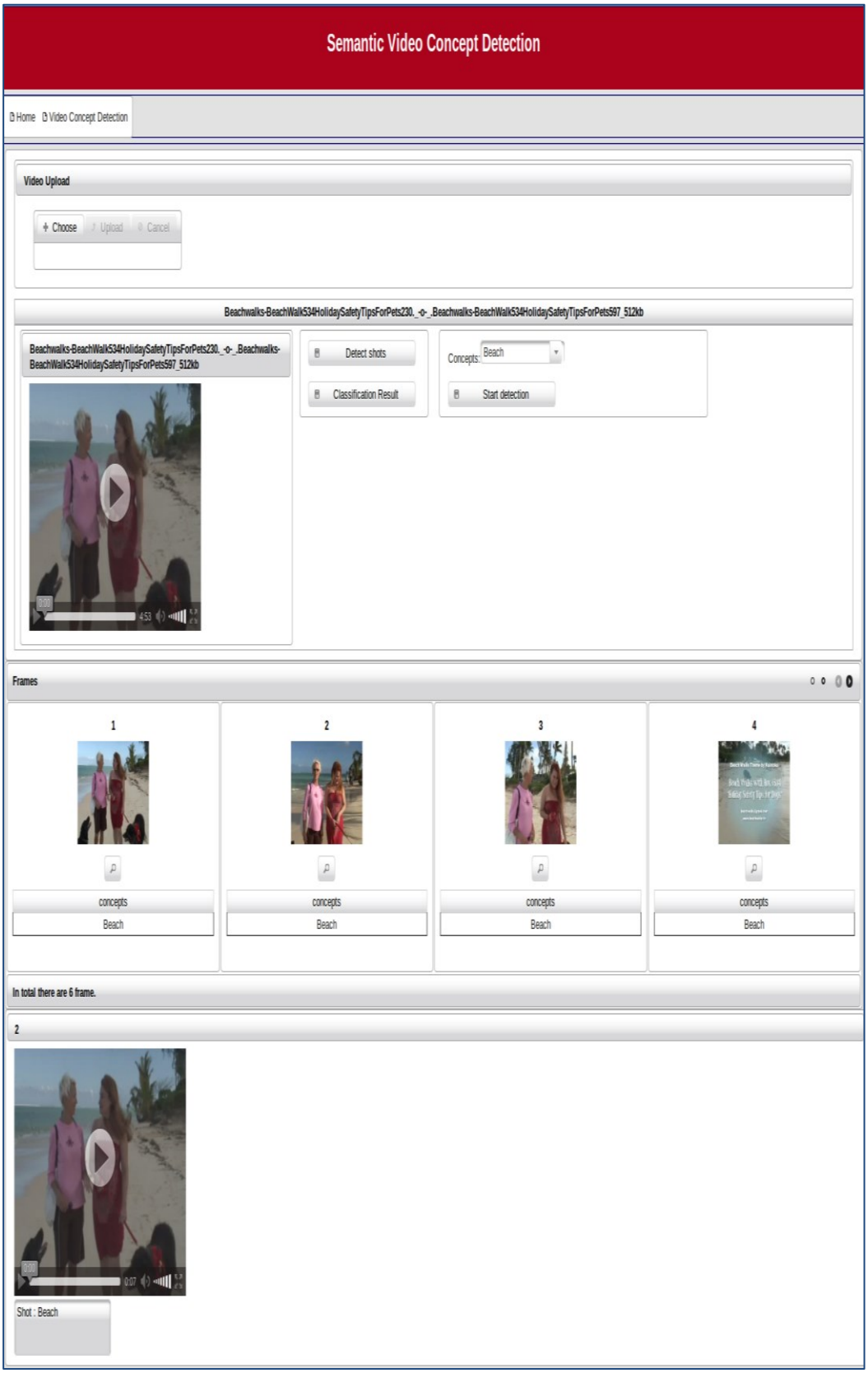
Şekil 5.3 Video Kavram Sezimi uygulama işlevleri



Şekil 5.4 Video kavram sınıflandırma işlemi

Uygulamada Şekil 4.3'te önerilen yöntem kullanılmıştır. Özetle, uygulamada aşağıda belirtilen maddeler son kullanıcı tarafından gerçekleştirilebilmektedir. Uygulamadan bir örnek Şekil 5.5'te gösterilmiştir.

- Kullanıcı, sezim yapmak istediği video verisini yükleyebilmektedir.
- Yüklenmiş olan video veriye ait çerçeve ve çekim bilgilerini "*Detect Shots*" isimli düğmeye basarak elde edilmektedir.
- Herhangi bir çerçeve görseli üzerine basarak ilgili çekimi görüntülenmektedir.
- "*Classification Result*" düğmesi ile sınıflandırma sonuçları, çerçeveler altında yer almaktadır.
- Video kavram sezimi için algılanması istenen kavram seçildiğinde, kavrama ait çekimler görüntülenmektedir.



Şekil 5.5 Video Kavram Sezim uygulaması

6 DENEYSEL SONUÇLAR

6.1 Kullanılan Veri Kümeleri

6.1.1 Trecvid 2013 SIN

Tez çalışmasında, geçmiş çalışmalarda yaygın olarak kullanılan veri kümesi TRECVID 2013 SIN (anlamsal dizinleme) video veri kümesi kullanılmıştır. Eğitim veri kümesi 2010, 2011 ve 2012 Trecvid SIN görevinde kullanılan test veri kümelerinin birleşiminden ve yeni video verilerinden oluşmaktadır. Eğitim veri kümesi IACC.1 kolleksiyonundan videolar içermektedir. Eğitim veri kümesinin içerdiği videoların toplam süresi yaklaşık olarak 600 saattir ve her bir videonun süresi 10 saniye ile 3.5 dakika arasındadır. (IACC.1.A, IACC.1.B, IACC.1.C). IACC.2.A Test veri kümesi IACC.2 kolleksiyonundan elde edilen ve toplam 200 saatlik video süresi olan bir veri kümesidir. Bu veri kümesinde her bir videonun süresi 10 saniye ile 6 dakika arasındadır.

Ayrıca her kolleksiyon çekim sınırlarını ve ana çekim sınırlarını içermektedir. Yapılan çalışmada 38 kavram (sınıf) kullanılmıştır. Bu kavramlar Çizelge 6.1'de verilmiştir. Veri setinden alındanörnek anahtar çerçeveler Şekil 6.1'de gösterilmiştir.

Şekil 6.1 Trecvid 2013 SIN veri kümesinden alınan örnekler

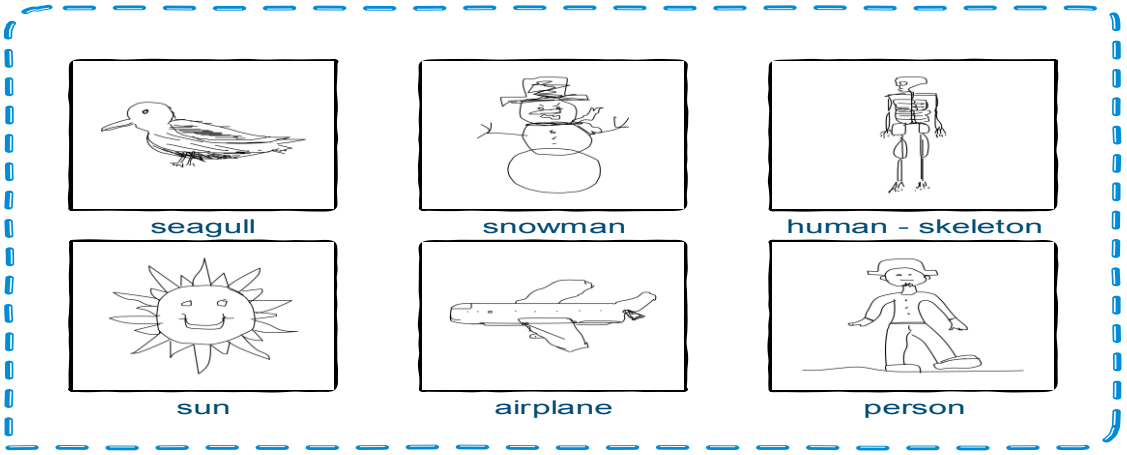
Çizelge 6.1 Trecvid-2013 SIN veri kümesinde kullanılan kavramlar

Video Kavram İsimleri			
Airplane	Computers	Kitchen	Throwing
Anchorperson	George_Bush	Motorcycle	Baby
Animal	Explosion_Fire	New_Studio	Flowers
Beach	Female-Human- Face-Closeup	Old_People	Fields
Boat_Ship	Door_Opening	Girls	Flags
Boy	People_Marching	Running	Forest
Bridges	Government_Leader	Singing	Dancing
Bus	Military_Airplane	Sitting Down	Hand
Chair	Instrumental Musician	Telephones	Quadruped
Skating	Studio_with Anchorperson		

6.1.2 TU-Berlin ve Sketchy Veri Kümeleri

Tez çalışmasında iki farklı çizim veri seti kullanılmıştır. Bunlardan ilki, TU-Berlin (Eitz et al., [14]), veri setidir.

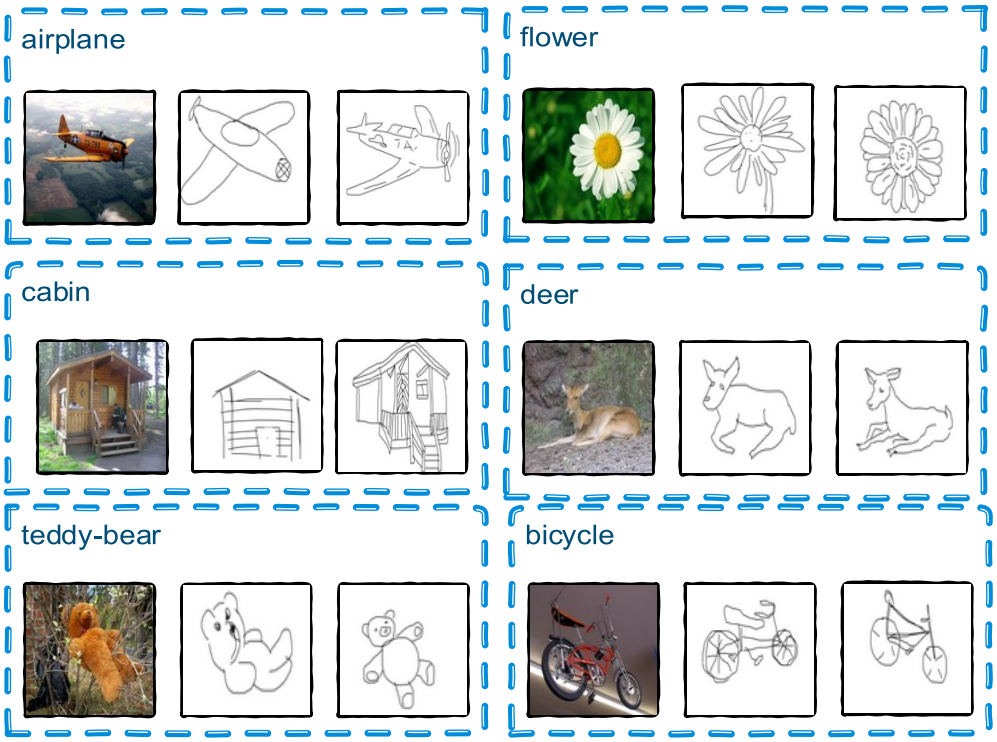
Bu veri seti toplamda 2000 çizim ve 250 kategoriden oluşmaktadır. Kategorilerden örnek olarak seagull, sun, and snowman verilebilir. Her bir kategoride 80 çizim bulunmaktadır. Çizimler, Amazon Mechanical Türk tarafından 1.350 katılımcıdan toplanmıştır. Tüm değerlendirmeler, alternatiflerle karşılaştırmayı mümkün kılmak için 3 bölümlü çapraz doğrulama (3-fold cross validation) kullanılarak gerçekleştirilmiştir. Veri setinden alınan örnek çizimler Şekil 6.2'de gösterilmiştir.



Şekil 6.2 TU-Berlin veri kümesinden alınan örnekler

İkinci veri kümesi, çizim-fotoğraf çiftlerinden oluşan büyük ölçekli bir koleksiyondur (Sangkloy et al., [22]).

Veri seti, 12.500 nesnenin fotoğrafı ve 125 kategoride 75.471 insan çizimi ile oluşturulmuştur. Sketchy veri seti, çizim ve resim anlayışının geliştirilmesi için açık bir yapıdadır. Bu veri setinden alınan örnekler Şekil 6.3'de gösterilmiştir.



Şekil 6.3 Sketchy veri kümesinden alınan örnekler

6.2 Video Kavram Sınıflandırma Sonuçları

Bu bölümde, video kavram sezim çalışmasında veri kümesi olarak kullanılan Trecvid SIN sonuçları, aşağıda yer alan alt başlıklar halinde incelenmiştir.

6.2.1 Öznitelik analiz sonuçları

Analizlerde, seçilen AlexNet, VGG19, GoogleNet ve Resnet101 modellerin soft-max katmanından bir önceki katman olan katmanlarının bağımsız olarak testleri yapılmıştır. Sonuçlar doğrusal ve RTF çekirdek kullanılarak, DVM başarım sonuçları elde edilmiştir. Sonuçlara göre, son katmanlar, seçilen modeller içinde RTF çekirdek – DVM yaklaşımı ile daha iyi performans göstermektedir. Çizelge 6.2' de Resnet101 RTF çekirdek – DVM ile %45.34 sonucunu alırken, doğrusal çekirdek - DVM ile %44.29 sonucu alınmıştır. Ayrıca ResNet101 modelinin diğer modellere göre daha yüksek başarım gösterdiği sonucuna varılmıştır. Bunun sonucunda, mimarilerin

katman derinliği arttıkça, daha etkin özniteliklerin elde edildiği gözlenmiştir.

Çizelge 6.2 CNN model katman sonuçları

CNN Model Adı ve Katman	Doğruluk (%)	
	Doğrusal Çekirdek DVM	RTF Çekirdek DVM
AlexNet - FC8	38.25	39.28
VGG19 - FC8	40.94	42.16
GoogleNet – Pool5	42.45	42.67
ResNet – FC1000	**44.29**	**45.34**

6.2.2 Art arda ekleme yöntem sonuçları

4 farklı modelin art arda ekleme yöntemiyle öznitelik kaynaşımından %46.92 doğruluk sonucu elde edilmiştir (Çizelge 6.3). Fakat, kaynaşım öznitelik vektörünün boyut olarak büyük olması nedeniyle eğitim maliyeti yüksektir. Alexnet-FC8, VGG19-FC8, GoogleNet-Pool5 ve Resnet101-FC1000 katmanlarından çıkarılan öznitelikler sırasıyla 1000, 1000, 1024, 1000 boyutludur. Öznitelik kaynaşımı sonucu toplam boyutu 4024 olan vektör elde edilmiştir. Elde edilen sonuçlara göre, özniteliklerin elde edildiği CNN modellerin katman derinliği arttıkça başarımın yükseldiği gözlenmiştir. Katman derinliğinin artmasıyla, girdilerden daha anlamsal bilgi içeren ve spesifik öznitelikler elde edildiği bulgusuna varılmıştır.

6.2.3 Boyut indirgeme sonuçları

Art arda ekleme yöntemi, modeller üzerindeki bağımsız testlere göre elde edilen başarımı artırtığı gözlenmektedir (Çizelge 6.3).

Fakat eklenme ile artan öznitelik vektör boyutu, eğitim maliyetini artırmıştır. 4024 boyutunda elde edilen öznitelik vektörü 1000 boyuta indirgenerek %65.16 oranında eğitim maliyeti azaltılmıştır. Boyut indirgeme ve öznitelik seçiminde sıkça kullanılan PCA yöntemi normalizasyon sonrası uygulanmıştır.

PCA işlemi ile 1000 boyutlu öznitelik vektörü elde edilmiştir. PCA yöntemi ile vektör boyutunun indirgenmesine rağmen, başarım sonuçlarında artış gözlenmiştir.

Bunun nedeni ise etkin öznitelik seçimi yapmasıdır. Sonuçlara göre, öznitelik kaynaşımı ile elde edilen %46.92 doğruluk oranı PCA ile %50.27'ye yükseltmiş ve öznitelik vektör boyutunun indirgenmesi ile eğitim maliyeti ve karmaşıklığı azalmıştır.

6.2.4 DCA yöntem sonuçları

Öznitelik seviyesinde kaynaşım yöntemi olan DCA ile sonuç Çizelge 6.3' de gösterilmiştir. DCA ile kaynaşım 2 farklı modelin kaynaşımı ile elde edilmiştir. Kaynaşım, öznitelik seviyesinde ve en iyi performans sonucu veren GoogleNet-Pool5 ile ResNet101-FC1000 arasında yapılmıştır. Art arda bağlama yöntemiyle elde edilen 2024-boyutlu öznitelik vektör boyutu 75-boyuta indirgenmiş ve RTF-DVM yaklaşımı uygulanmıştır. Sonuç olarak %47.47 doğruluk oranı elde edilmiştir. DCA yöntemi ile boyut indirgenme sağlanmasına rağmen art arda ekleme ile PCA yöntemi sonucuna göre daha az başarım göstermektedir.

Çizelge 6.3 Art arda ekleme, DCA ve PCA yöntem Trecvid sonuçları

CNN Model Adı ve Katman	**Uygulanan Yöntem**	**Doğruluk (%)**
		RTF Çekirdek DVM
AlexNet – FC8, VGG19 - FC8, GoogleNet – Pool5, ResNet – FC1000	Art Arda Bağlama İşleci (Öznitelik Kaynaşım)	46.92
GoogleNet – Pool5, ResNet – FC1000	DCA (Öznitelik Kaynaşım)	47.47
AlexNet – FC8, VGG19 - FC8, GoogleNet – Pool5, ResNet – FC1000	**Art Arda Bağlama İşleci + PCA**	**50.27**

6.2 Çizim Kavram Sınıflandırma Sonuçları

Geliştirilen yöntemlerin değerlendirilmesi için dört farklı test yapılmıştır. Bu bölümde, bağımsız model katman testleri, art arda ekleme yöntemi, PCA uygulanan yöntem sonuçları ve son olarak skor seviyesinde kaynaşım sonuçları gösterilmektedir.

6.2.1 Öznitelik analiz sonuçları

Analizlerde sırasıyla GN-Triplet, AlexNet ve VGG19 CNN mimarilerinin Pool5, FC6, Pool5 katmanlarındaki bağımsız testleri göz önünde bulundurulmaktadır. Çizelge 6.3'te TU-Berlin veri kümesi sonuçları gösterilmektedir. Sonuçlara göre AlexNet FC6, VGG19 Pool5 ve GN-Triplet'in Pool5 katman başarımlarının daha yüksek olduğu gözlenmektedir.

6.2.2 Art arda ekleme ve boyut indirgeme yöntem sonuçları

Kombine bir biçimde kullanıldığında en iyi performans gösteren iki CNN mimarisi kullanılmıştır. TU-Berlin ve Sketchy veri setleri için elde edilen sonuçlar sırasıyla Çizelge 6.4 ve Çizelge 6.5'te gösterilmiştir. Tüm testler Şekil 4.2 ve Şekil 4.5'de verilen CNN-DVM iletimini izlemektedir. TU-Berlin veri seti için (Qian et al., [15]), sunulan çalışmayla karşılaştırıldığında, katman FC6, %67.26 doğrulukla AlexNet için en iyi performans gösteren FC tabakasıdır.

Bu sonuç, geçmiş çalışmalarda HOG-DVM (%56), topluluk uyumu (ensemble matching) (%61.5), MKL-DVM (%65.8), AlexNet-DVM (%67.1) ve LeNet'in (%55.2) yöntemlerinden daha iyi performans sergilemektedir (Qian et al., [15]). Bulgularda ayrıca önceki FC katmanları, VGG19 mimarisinin tam bağlı katmanlarından daha iyi performans gösterir. Boyutları daha büyük olmasına rağmen, VGG19'un katmanı olan Pool5, FC6'ya (%60.30) kıyasla %64.14 ile daha yüksek doğruluk oranı elde etmektedir. Bu sonuç, çizimler için özel olarak tasarlanmış bir CNN mimarisi olan Qian et al., [15] dışındaki mevcut tüm yöntemlerden belirgin biçimde daha iyi performans gösterir. TU-Berlin veri kümesindeki deneylerimize dayanarak sırasıyla Alexnet ve GN-Triplet mimarilerinin FC6 ve Pool5 katmanlarını birleştirdiklerinde en iyi sonuç elde edilmiştir. %72.5'lik tanıma doğruluğu ile aynı veri kümesindeki insan doğruluğunun (%73.1) yakınında bir sonuca varılmıştır.

Öznitelik seviyeli kaynaşım yönteminde, Sketchy veri seti için, AlexNet FC6 (%80.89) ve GN-Triplet Pool5 (%95.16) katmanından en iyi sonuçlar elde edildi. Buna ek olarak, Sketchy veri kümesindeki VGG19 tabakası sonucu Alexnet ve GN-Triplet'ten daha düşük doğruluk (%77.76) göstermektedir. TU-Berlin veri kümesindeki en iyi sonuç (%72.5), Alexnet-FC6 ve GN-Triplet-Pool5 CNN özelliklerini birleştirerek elde edildi.

Sonuçlar, öznitelik seviyesi kaynaşım ve PCA yönteminin, bağımsız CNN modelini kullanmaktan daha etkili olduğu gösterilmiştir.

Çizelge 6.4 Öznitelik seviyesinde kaynaşım yönteminin TU-Berlin veri kümesindeki tanıma sonuçları

Model Adı	Model Katman	Yöntem	Doğruluk (%)
Alexnet	FC8	-	56.20
Alexnet	FC7	-	59.23
Alexnet	FC6	-	67.26
VGG-19	FC6	-	60.30
VGG-19	Pool5	-	64.14
GN-Triplet	Pool5	-	68.16
Alexnet	FC6-FC7	Öznitelik Kaynaşım	66.50
Alexnet-VGG-19	FC6-FC8	Öznitelik Kaynaşım	67.15
Alexnet-VGG-19	FC6-FC6	Öznitelik Kaynaşım	68.25
Alexnet-VGG-19	FC6-Pool5	Öznitelik Kaynaşım	69.175
VGG19- GN-Triplet	Pool5-Pool5	Öznitelik Kaynaşım	69.71
Alexnet- GN-Triplet	FC6-Pool5	Öznitelik Kaynaşım	70.785
Alexnet- GN-Triplet	**FC6-Pool5**	**Öznitelik Kaynaşım + PCA**	**72.5**

Çizelge 6.5 Öznitelik seviyesinde kaynaşım yönteminin Sketchy veri kümesindeki tanıma sonuçları

Model Adı	Model Katman	Yöntem	Doğruluk (%)
Alexnet	FC6	-	80.89
VGG-19	Pool5	-	77.76
GN-Triplet	Pool5	-	95.16
Alexnet-VGG-19	FC6-Pool5	Öznitelik Kaynaşım	82.18
VGG19- GN-Triplet	Pool5-Pool5	Öznitelik Kaynaşım	95.47
Alexnet- GN-Triplet	FC6-Pool5	Öznitelik Kaynaşım	96.84
Alexnet- GN-Triplet	**FC6-Pool5**	**Öznitelik Kaynaşım + PCA**	**97.91**

6.2.3 Skor seviyesinde kaynaşım sonuçları

Çizelge 6.6' da, Şekil 4.5'de gösterilen sistem uygulanarak, Sketchy ve TU-Berlin veri setleri üzerinde skor kaynaşım yöntemlerinin (max, ortalama ve min) sonuçları gösterilmiştir. Sonuçlar en iyi sonuç veren modeller üzerinde denenmiştir.

Skor kaynaşım yöntemlerinden en iyi performans üç skor vektöründen maksimum olanın seçilmesiyle elde edilmiştir.

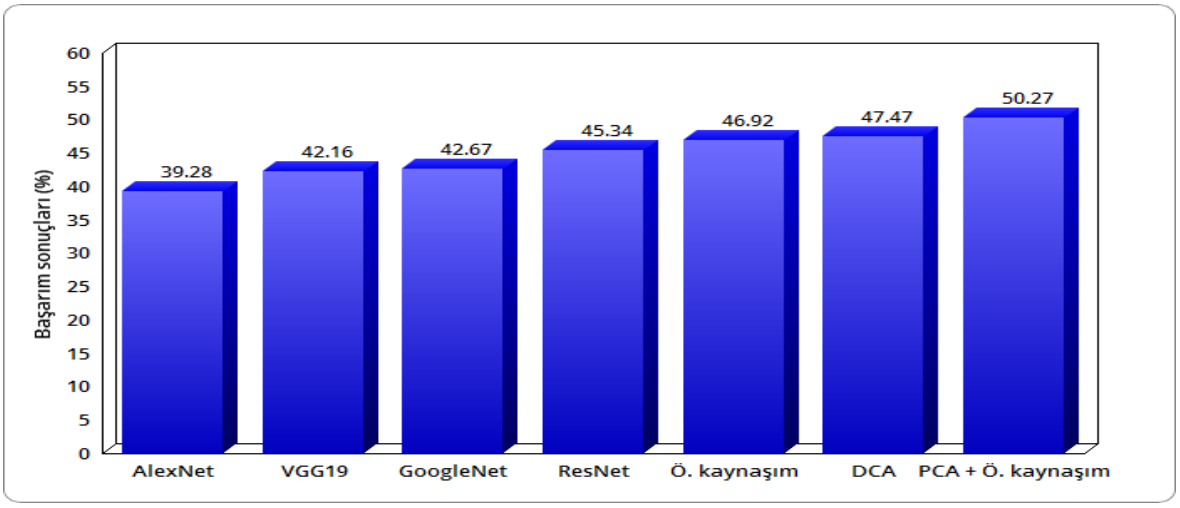
Sketchy veri setinde maksimum kaynaşım yapılarak %98.45 doğruluk oranında sonuç alınırken, TU-Berlin veri setinde ise %70.56 başarım alınmıştır.

Çizelge 6.6 Skor kaynaşım yöntemlerinin Sketchy ve TU-Berlin veri kümelerindeki tanıma sonuçları

CNN Model Adı ve Katman	Uygulanan Yöntem	Doğruluk (%) RTF Çekirdek DVM	Veri Kümesi
Alexnet-VGG19-GN-Triplet FC6-Pool5-Pool5	**PCA + Skor Kaynaşım (max)**	**70.56**	**TU-Berlin**
Alexnet-VGG19-GN-Triplet FC6-Pool5-Pool5	PCA + Skor Kaynaşım (ortalama)	70.10	TU-Berlin
Alexnet-VGG19-GN-Triplet FC6-Pool5-Pool5	PCA + Skor Kaynaşım (min)	70.06	TU-Berlin
Alexnet-VGG19-GN-Triplet FC6-Pool5-Pool5	**PCA + Skor Kaynaşım (max)**	**98.45**	**Sketchy**
Alexnet-VGG19-GN-Triplet FC6-Pool5-Pool5	PCA + Skor Kaynaşım (ortalama)	98.34	Sketchy
Alexnet-VGG19-GN-Triplet FC6-Pool5-Pool5	PCA + Skor Kaynaşım (min)	98.09	Sketchy

Elde edilen sonuçlara göre üç modelden alınan skor veri kaynaşım yöntemi Sketchy veri setinde daha başarılı sonuç vermiştir.

Akıllı telefon üzerinde çizim tanıma uygulaması için, en iyi sınıflandırıcıyı kullanarak en iyi özniteliklerin elde edileceği CNN mimarilerine karar verildi. Ardından, sınıflayıcı tasarımı oluşturulmuştur (Şekil 4.5). Çizim veri seti (TU-Berlin) için AlexNet-FC6 ve GN-Triplet Pool5 öznitelik kombinasyonu kullanılmıştır.

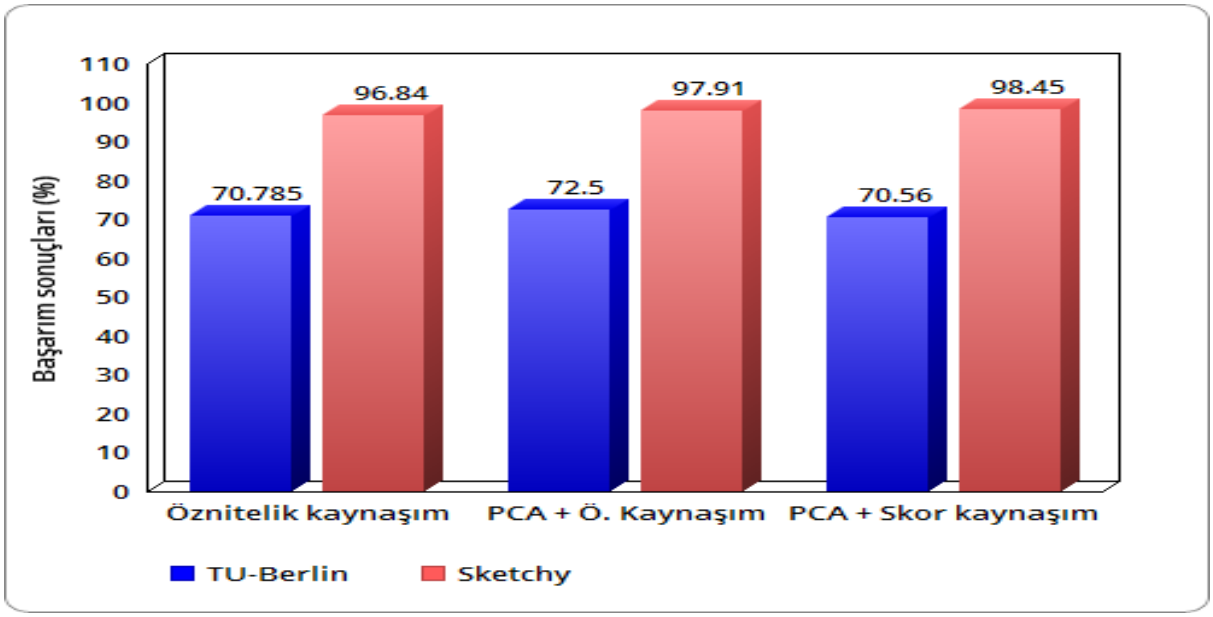


Şekil 6.4 Trecvid 2013 SIN genel sonuçlar

Sonuçları özetlemek gerekirse:

Şekil 6.4'de Trecvid 2013 SIN veri setinden elde edilen başarım sonuçları özet olarak verilmiştir. Sonuçlara göre boyut indirge yöntemi olan PCA ve art arda ekleme ile en yüksek başarım elde edilmiştir.

Şekil 6.5'de Sketchy ve TU-Berlin veri setleri üzerindeki sonuçlar verilmiştir. Uygulanan öznitelik kaynaşımı en yüksek başarım gösteren AlexNet ve GN-Triplet öznitelikleri ile elde edilmiştir.



Şekil 6.5 TU-Berlin ve Sketchy genel sonuçlar

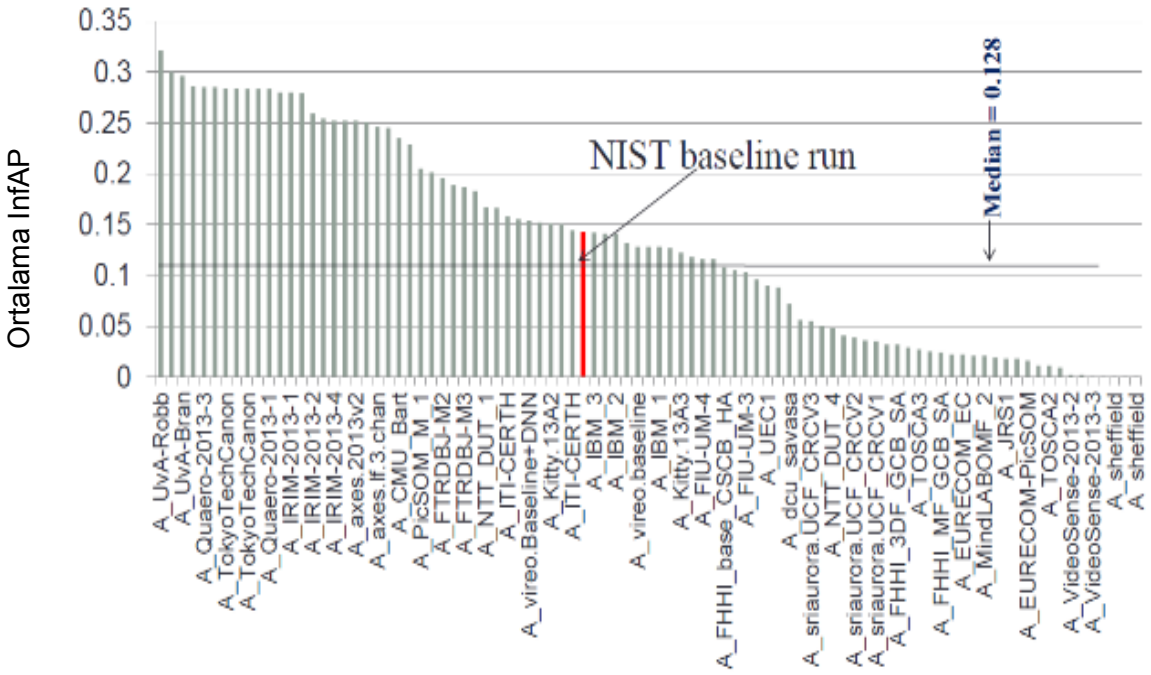
TU-Berlin veri kümesindeki sonuçlarımız, PCA ve AlexNet FC6 ve GN-Triplet Pool5 katmanlarını kullanan önerilen özellik kaynaşım şemasının, %72.5 doğrulukla bağımsız CNN modellerine kıyasla daha iyi tanıma doğruluğunu sağladığını göstermektedir. Bu sonuç, çizim için özel olarak tasarlanmış bir CNN mimarisi olan Qian et al., [15] dışındaki tüm mevcut yöntemlerden belirgin biçimde daha iyi performans gösterir. Elde edilen tanıma doğruluğu, aynı veri kümesindeki insan doğruluğunun yakınında (%73.1) verir.

Çizelge 6.7'de TU-Berlin veri kümesi üzerinde, son çalışmalarda elde edilen sonuçlar gösterilmektedir. Şekil 4.2'de gösterilen yöntemde TU-Berlin veri kümesinde %72.5 oranında başarım sağlanmıştır. Böylelikle insan tanıma başarımının %73.1 olduğu yerde, bu yöntemin oldukça başarılı olduğu gözlenmiştir.

Çizelge 6.7 TU-Berlin veri kümesinde önerilen yöntem ile geçmiş çalışmaların kıyaslanması

Yöntem	Doğruluk (%)
HOG - SVM	56
Ensemble	61.5
MKL - SVM	65.8
FV - SP	68.9
AlexNet - SVM	67.1
AlexNet - Sketch	68.6
LeNet	55.2
Öznitelik Kaynaşım + PCA (Önerdiğimiz)	**72.5**
İnsan Tanıma	73.1

Şekil 6.6'da Trecvid 2013 SIN görevinde 38 kavram için alınmış olan sonuçlar gösterilmektedir. En iyi sonuç Uva-Robb tarafından %32.1 ile birinci sırayı almıştır. Sonuçlar infAP (Ortalama Tahmin Çıkarımı, Inferred Average Precision) metriği ile değerlendirilmiştir (Yilmaz et al., [67]). Bu metriğin hesaplanabilmesi için tüm veri kümesine ihtiyaç duyulmaktadır. Bu çalışmada, video kavram tanıma için önerilen yöntem sonuçları, veri kümesindeki tüm verilere erişemediğimizden doğruluk (accuracy) metriği ile değerlendirilmiştir.



Şekil 6.6 Trecvid 2013 SIN sonuçları

7 SONUÇLAR VE GELECEK ÇALIŞMA PLANI

Bu tez çalışmasında, derin evrişimsel ağ mimarileri ile öznitelik düzeyli kaynaşım ve öznitelik seçimine dayalı bir video kavram sezim yöntemi önerilmiştir. ImageNet veri kümesi üzerinde ön-eğitimli derin evrişimsel sinir ağ modelleri olan AlexNet, VGG19, GoogleNet, ResNet101 ile elde edilen öznitelikler kaynaştırılmış ve elde edilen yeni öznitelik vektörlerinin boyutu PCA yöntemi ile indirgenmiştir. İndirgenen özniteliklerden DVM yöntemi ile öğrenilen 38 TRECVID 2013 SIN kavramı %50.27 doğruluk oranı ile sınıflandırılmıştır. Deneysel analizler göstermektedir ki, farklı ön-eğitimli evrişimsel sinir ağ modellerinin art arda bağlama işleci ile öznitelik düzeyli kaynaşımı başarımı %1.58 oranında artırmaktadır. DCA öznitelik kaynaşım tekniği ile boyut indirgenme sağlanmış ve %47.47 doğruluk oranında sonuca ulaşılmıştır. DCA yönteminde, 4 farklı model yerine en başarılı sonuç veren 2 farklı CNN modelinden elde edilen vektörlerin kullanılması ve boyut ingenmesi ile eğitim maliyeti yaklaşık olarak 1.6 kat oranında azaltılmıştır . TRECVID 2013 SIN veri kümesi üzerinde elde edilen sonuçlar, öznitelik düzeyli kaynaşımın veya öznitelik seçiminin video kavram sınıflandırma performansını artırdığını göstermektedir. DCA yönteminden etkin öznitelikler elde edilmesine rağmen, art arda ekleme ve sonrasında uygulanan PCA yöntemi ile elde edilen özniteliklerin daha etkin olduğu gözlenmiştir. DCA yöntemi ile vektör boyutları indirgenmiş fakat değer (öznitelik) kaybı yaşanmıştır.

Çalışmada geliştirilen yöntemlerin etkinliğini ölçmek amaçlı, görsel tanıma sistemlerinden biri olan çizim tanıma sisteminde, derin evrişimsel sinir ağları mimarilerinin öznitelik / skor seviyesinde kaynaşım ve öznitelik seçimine dayanan bir çizim tanıma sistemleri önerilmiştir. ImageNet veri setinde önceden eğitilmiş CNN modelleri olan görüntü sınıflandırma görevlerinde kanıtlanmış başarısı nedeniyle AlexNet, GN-Triplet ile elde edilen öznitelikler, öznitelik seviyesinde kaynaştırılmış ve elde edilen yeni öznitelik vektörlerinin boyutu PCA yöntemi ile azaltılmıştır. Ayrıca, 3 farklı öznitelik vektörlerin (AlexNet, VGG19, GN-Triplet) boyutları indirgendikten sonra DVM skor vektörleri üzerinde skor kaynaşım yöntemi uygulanarak Sketchy veri seti üzerindeki başarım artırılmıştır. Ayrıca akıllı telefonlar için istemci-sunucu mimarisine dayalı önerilen şemayı kullanan bir çizim tanıma uygulaması geliştirilmiştir.

Çizim veri seti için, en iyi sonucu, birleştirilen öznitelik vektör boyutunu ve öznitelik seçimi azaltmak için PCA uygulayarak skor kaynaşımından (%98.45) elde edilmiştir. Sonuçlar, veri kaynaşımı olarak öznitelik ve skor seçim teknikleri kullanıldığında TU-Berlin ve Sketchy veri kümeleri için sınıflandırma doğruluğunun sırasıyla %72.5 (öznitelik seviyesinde kaynaşım) ve %98.45 (skor kaynaşım) doğrulukla yaklaşık %3 ve %4 oranında arttığını göstermektedir. Elde edilen sonuçlara göre veri kaynaşım yöntemleri ile öznitelik etkinliğin arttığı gözlenmiştir. Ayrıca CNN-DVM iletim yönteminin kullanılması anlamsal kavram sınıflandırmada daha yüksek performans gösterdiği bulgusuna varılmıştır. Bununla birlikte önerdiğimiz yöntem Sketchy veri kümesi üzerinde, TU-Berlin veri kümesine göre daha yüksek başarım sağlamıştır. Bunun nedenleri arasında GN-Triplet modelinin Sketchy veri kümesinde bir resmin çizimi ile birlikte eğitilmesinin olduğu değerlendirilmektedir.

Videolardaki genel kavram sezimi, genel olarak zorlu bir problemdir ve sezim sağlamlığı, doğruluğu ve hızı iyileştirmek için birçok araştırma yapılmıştır. Çeşitli yaklaşımlar arasında CNN mimarileri ve kaynaşım yöntemleri birçok araştırmacıdan ilgi uyandırmaya devam etmektedir. Bu tez çalışmasında CNN öznitelikleri ile öznitelik ve skor seviyesinde kaynaşım yöntemleri ile başarımlar artırılmıştır. Ayrıca öznitelik seçimine dayalı PCA ile eğitim maliyeti azaltılmıştır.

Gelecek çalışma planı olarak, video kavram tanısı sistemi için farklı CNN özniteliklerine ek olarak ses özniteliklerinin de probleme uygulanması hedeflenmektedir.

KAYNAKLAR LİSTESİ

[1] Krizhevsky, A., Sutskever, I. and Hinton, G.E., Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, s.1097–1105, 2012.

[2] Simonyan, K. and Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition, Computing Research Repository (CoRR), arXiv 1409.1556, 2014.

[3] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, and Rabinovich, A., Going deeper with convolutions, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), s.1-9, 2015.

[4] He, K., Zhang, X., Ren, S., and Sun, J., Deep residual learning for image recognition, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), s.770-778, 2016.

[5] Long, J., Shelhamer, E., Darrell, T., Fully convolutional networks for semantic segmentation, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), s.3431-3440, 2015.

[6] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, Jonathan Ross Girshick, R., Guadarrama, R., and Darrell T., Caffe: Convolutional architecture for fast feature embedding. arXiv preprintarXiv:1408.5093, 2014.

[7] Ergun, H. and Sert, M., Fusing Deep Convolutional Networks for Large Scale Visual Concept Classification, IEEE International Conference on Multimedia Big Data (BigMM2016), 2016.

[8] He, K., Zhang, X., Ren, S., and Sun, J., Spatial pyramid pooling in deep convolutional networks for visual recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.37, no.9, s.1904– 1916, 2015.

[9] Haghighat, M., Abdel M., Alhalabi W., Discriminant Correlation Analysis : Real-Time Feature Level Fusion for Multimodal Biometric Recognition, IEEE Transactions on Information Forensics and Security, vol.11, no.9, s.1984-1996, Eylül 2016.

[10] Feichtenhofer, C., Pinz, A., Zisserman, A., Convolutional two stream network fusion for video action recognition, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), s.1933-1941, 2016.

[11] Girshick, R., Donahue, J., Darrell, T., Malik, J., Rich featurehierarchies for accurate object detection and semantic segmentation, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), s.580-587, 2014.

[12] Ergun, H., and Sert, M., Efficient bag of words based concept extraction for visual object retrieval, In Flexible Query Answering Systems 2015, s.389–402. Springer, 2016.

[13] Ergun, H., Akyuz, Y. C., Sert, M., Liu, J., Early and Late Level Fusion of Deep Convolutional Neural Networks for Visual Concept Recognition,

International Journal of Semantic Computing, vol.10, no.03, s.379-397, 2016.

[14] Eitz, M., Hays, J. and Alexa, M., How Do Humans Sketch Objects? ACM Trans. Graph., vol.31 no.4, s.1–10, 2012.

[15] Qian, Y., Yongxin, Y., Yi-Zhe, S., Xiang, T. and Hospedales, T.M., Sketch-a-Net that Beats Humans, Proceedings of the British Machine Vision Conference (BMVC), 7-10 Eylül, Swansea-UK, s.1–12, 2015.

[16] Eitz, M., Hildebrand, K., Boubekeur, T. and Alexa, M., Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors, The Institute of Electrical and Electronics Engineers (IEEE) Trans. on Visualization and Computer Graph, vol.17, no.11, s.1624–1636, 2011.

[17] Schneider, R.G. and Tuytelaars, T., Sketch Classification and Classification-driven Analysis Using Fisher Vectors, ACM Trans. Graph., vol.33 no.6, s.1–9, 2014.

[18] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., Large-Scale Video Classification with Convolutional Neural Networks, Proceedings of the 2014 (IEEE) Conference on Computer Vision and Pattern Recognition, s.1725–1732, 2014.

[19] Ergun, H. and Sert, M., Fusing Deep Convolutional Networks for Large Scale Visual Concept Classification, IEEE International Conference on Multimedia Big Data (BigMM2016), 2016.

[20] Ergun, H., Akyuz, Y. C., Sert, M., Liu, J., Early and Late Level Fusion of Deep Convolutional Neural Networks for Visual Concept Recognition, International Journal of Semantic Computing, vol.10, no.03, s. 379-397, 2016.

[21] Sivic, J., and Zisserman, A., Video google: a text retrieval approach to object matching in videos, Computer Vision, 2003. Proceedings. Ninth IEEE International Conference, vol.2, s.1470–1477, 2003.

[22] Sangkloy, P., Burnell, N., Ham, C., Hays, J., The sketchy database: learning to retrieve badly drawn bunnies, ACM Transactions on Graphics, vol.35, no.4, 2016.

[23] Aihkisalo, T., and Paaso, T., Latencies of service invocation and processing of the REST and SOAP web service interfaces, IEEE 8th World Congress on Services, s.100–107, 24-29 Haziran, Honolulu, HI-USA, 2012.

[24] Wagh, K., and Thool R., A comparative study of SOAP Vs REST web services provisioning techniques for mobile host, Journal of Information Engineering and Applications, vol.2, no.5, s.12–16, 2012.

[25] Boyaci, E., Sert, M., Feature–level fusion of deep convolutional neural networks for sketch recognition on smartphones, In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE2017), 8-10 Ocak, Las Vegas, Nevada- USA, s.485-486, 2017.

[26] Farrugia, P.J., Borg, J.C., Camilleri, K.P., Spiteri, C., and Bartolo A., A cameraphone-based approach for the generation of 3D models from paper

sketches, Eurographics Workshop on Sketch-Based Interfaces and Modeling, s.33-42, 2004.

[27] Tseng, K.Y., Lin, Y.L., Chen, Y.H., and Hsu, W.H., Sketch-based image retrieval on mobile devices using compact hash bits, In Proceedings of the 20th ACM International Conference on Multimedia, 29 Ekim-02 Kasım, Nara-Japan, s.913–916, 2012.

[28] Safadi, B., Derbas N., Quenot, G., Descriptor optimization for multimedia indexing and retrieval. Multimedia Tools and Applications, vol.74 no.4 s.1267-1290, 2015.

[29] Jongejan, J., Rowley, H., Kawashima, T., Kim, J., Nick Fox-Gieg, et. al at Google Creative Lab and Data Arts Team, Quick, Draw!, https://quickdraw.withgoogle.com/, 2017.

[30] Guo, J., Gould, S., Deep CNN ensemble with data augmentation for object detection, Computing Research Repository (CoRR), arXiv:1506.07224, 2015.

[31] Snoek, C., Worring, M., and Smeulders, A., Early versus late fusion in semantic video analysis, In Proceedings of the 13th Annual ACM International Conference on Multimedia, s.399–402, 2005.

[32] Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale A., Ferguson, D., Real-time pedestrian detection with deep network cascades, 2015.

[33] Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S., CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) IEEE Computer Society, s.512-519, Washington-USA, 2014.

[34] Guo, J., Wang, C., Roman-Rangel, E., Chao, H. , Rui, Y., Building Hierarchical Representations for Oracle Character and Sketch Recognition, IEEE Transactions on Image Processing (TIP), Ocak 29, 2016.

[35] Dalal, N., Triggs, B., Histograms of oriented gradients for human detection, in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol.1, s.886–893, Haziran, 2005.

[36] Oliva, A. and Torralba, A., Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. J. Comput. Vis., vol.42, no.3, s. 145–175, Mayıs, 2001.

[37] Nowak, E., Jurie, F., and Triggs, B., Sampling strategies for bag-of features image classification, in Computer Vision—ECCV, s.490–503, Springer-Verlag, New York-USA, 2006.

[38] Xiao, C., Wang, C., Zhang, L., PPTLens: Create Digital Objects with Sketch Images, ACM Conference on Multimedia, Temmuz 29, 2015.

[39] Srinivas, S., Ravi Sarvadevabhatla, K., Mopuri, K., R., Prabhu, N., Kruthiventi, S., Babu, R. V., A Taxonomy of Deep Convolutional Neural Nets for Computer Vision, Frontiers in Robotics and AI 2(36), Ocak, 2016.

[40] Chang C. and Lin C., LiBSVM : a Library for Support Vector Machines, ACM Transactions on Intelligent Systems and Technology, vol.2, no.3, s.27:1-27:27, Nisan, 2011.

[41] Zha, S., Luisier, F., Andrews, W., Srivastava, N., and Salakhutdinov, R., Exploiting image-trained cnn architectures for unconstrained video classification, arXiv preprint arXiv: 1503.04144, 2015.

[42] Van De Sande, K., Gevers, T., Snoek, C., Evaluating color descriptors for object and scene recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.32, no.9, s.1582–1596, 2010.

[43] Li, Y., Hospedales, T.M., Song, Y.Z. and Gong, S., Free-hand sketch recognition by multikernel feature learning, Computer Vision and Image Understanding, vol.137, s.1-11, 2015.

[44] Boyaci, E., Sert, M., Video Classification Based on ConvNet Collaboration and Feature Selection, IEEE 25th Signal Processing and Communications Applications Conference (SIU 2017), Antalya, Turkey, s.Tbd.

[45] Cortes, C., Vapnik, V., Support-vector networks, vol.20, no.3, s.273–297, 1995.

[46] MATLAB, version (R2016b). The MathWorks Inc., Natick, Massachusetts, 2016.

[47] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D., Handwritten digit recognition with a back-propagation network, Advances in Neural Information Processing Systems 2, s.396–404, Morgan-Kaufmann, 1990.

[48] Ovtcharov, K., Ruwarse, O., Kim, J., et al., Accelerating Deep Convolutional Networks Using Specialized Hardware, Microsoft Research, Şubat, 2015.

[49] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L., Speeded-Up Robust Features (SURF), Journal of Computer Vision and Image Understanding, vol.110, no.3, s.346-359, 2008.

[50] Lowe, D. G., Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, vol.60, no.2, s.91-110, 2004.

[51] Niblack, W., Barber, R., Equitz, W., Fickner, M., Glasman, E., Petkovic, D., and Yanker, P., The qbic project: Querying images by content using color, texture and shape. SPIE Conference on Geometric Methods in Computer Vision, vol.1908, s.173–187, 1993.

[52] Schiele, B., and Crowley, J., Recognition without correspondence using multidimensional receptive field histograms. International Journal of Computer Vision, vol.36, no.1, 2000.

[53] Stricker, M.A., and Orengo, M., Similarity of color images. Proc. SPIE Storage and Retrieval for Image and Video Databases, vol.2420, s.381–392, 1995.

[54] Vailaya, A., Jain, A., and Zhang, H.J., on image classification: City vs.

landscapes, Pattern Recognition, vol.31, no.12, s.1921–1935, 1998.

[55] Liu, S., Yi, H., Chia, L.T., and Rajan, D., Adaptive hierarchical multi-class svm classifier for texture-based image classification, Proc. IEEE International Conference on Multimedia and Expo, s.4, 2005.

[56] Manjunath, B.S., and Ma, W.Y., Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.18, no.8, s.837–842, 1996.

[57] Yanagawa, A., Hsu, W., and Chang, S.F., Brief descriptions of visual features for baseline Trecvid concept detectors. Columbia University ADVENT Tech. Report 219-2006-5, Temmuz, 2006.

[58] Chang, S.F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A.C., and Luo, J., Large-scale multimodal semantic concept detection for consumer video, ACM Internaltional Workshop on MIR, s.255–264, 2007.

[59] Hadjidemetriou, E., Grossberg, M., and Nayar, B., Multiresolution histograms and their use for recognition. IEEE Trasactions on Pattern Analysis and Machine Intelligence, vol.26, no.7, s.831–847, 2004.

[60] Ni, B., Yan, S., and Kassim A., Contextualizing histogram. Proc. IEEE International Conference on Computer Vision and Pattern Recognition, Miami, Florida, 2009.

[61] Chang, S.F., He, J.F., Jiang, Y.G., Yanagawa, A., Zavesky, E., Khoury, E., and Ngo, C.W., Columbia university/vireo-cityu/irit trecvid2008 high-level feature extraction and interactive video search, NIST TRECVID workshop, Gaithersburg, MD, 2008.

[62] Dong, X., and Chang, S.F., Visual event recognition in news video using kernel methods with multi-level temporal alignment. Proc. IEEE International Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 2007.

[63] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D., Backpropagation applied to handwritten zip code recognition. Neural computation, vol.1, no.4, s.541–551, 1989.

[64] Basanth Kumar, H.B., A Review on Shot Boundary Detection, Oriental Journal Of Computer Science & Technology, vol.7, no.1, s.39-44, Nisan, 2014.

[65] Wei Jiang, Advanced Techniques for Semantic Concept Detection in General Videos, Ph.D thesis, Columbia University, 2010.

[66] Safadi, B., Mulhem P., Quenot, G., Chevallet, J., LIG-MRIM at NTCIR-12 Lifelog Semantic Access Task, Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 2016.

[67] Yilmaz, E., Kanoulas, E., Aslam, J. A., A simple and efficient sampling method for estimating AP and NDCG. In SIGIR ’08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, s. 603–610, New York, USA, 2008.