BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK KİPLİ VİDEO KAVRAM SINIFLANDIRMASI

BERKAY SELBES

YÜKSEK LİSANS TEZİ

2018

MULTIMODAL VIDEO CONCEPT CLASSIFICATION

ÇOK KİPLİ VİDEO KAVRAM SINIFLANDIRMASI

BERKAY SELBES

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

2018

“Çok Kipli Video Kavram Sınıflandırması“ başlıklı bu çalışma, jürimiz tarafından 25/01/2018 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI’nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan :Prof. Dr. Mehmet Reşit Tolun

Üye (Danışman) :Yrd. Doç. Dr. Mustafa Sert

Üye :Doç. Dr. Sinan Kalkan

ONAY

.../02/2018

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Yüksek lisans eğitimim boyunca benden desteğini hiç esirgemeyen, benden hiçbir zaman ümidini kesmeyen, sürekli motivasyonumu yüksek tutarak zorluklara göğüs germemi sağlayan, engin tecrübe ve deneyimlerinden faydalandığım yol göstericim ve danışmanım Yrd. Doç. Dr. MUSTAFA SERT'e değerli katkılarından ve yardımlarından dolayı,

Bana iyi insan olmayı öğreten, bu günlere gelmem için beni yetiştiren, maddi ve manevi hiçbir desteğini esirgemeyen, hayatımda ne karar verirsem verim her zaman arkamda duran annem Songül Selbes'e ve babam İlhan Selbes'e bana gösterdikleri sabırdan dolayı,

Bana zor zamanlarımda destek olan, varlığı ile beni mutlu eden, her şeyden çok sevdiğim, değerli kardeşim Melis Selbes'e desteklerinden dolayı

TEŞEKKÜR EDERİM

ÖZ

ÇOK KİPLİ VİDEO KAVRAM SINIFLANDIRMASI

Berkay Selbes

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Çokluortam verileri, İnternet kullanımının artmasıyla, sürekli üretilmekte ve paylaşılmaktadır. Bunun bir sonucu olarak, çokluortam verilerinin büyüklüğü hızla artmakta ve bu verilerin içeriklerini analiz eden otomatik yöntemlere ihtiyaç duyulmaktadır. Video verisi, çokluortam verilerinin önemli bir bileşenidir. Video içerik analizi, video verisi içeriğindeki zamansal veya konumsal olayların ve kavramların otomatik yöntemlerle belirlenmesi olarak tanımlanabilen önemli bir araştırma konusudur. Video içerik analizi, video içeriğinin karmaşık yapısı nedeniyle zor bir görevdir ve içerdiği bilgilerin otomatik olarak elde edilebilmesi için etkin yöntemlere ihtiyaç duyulmaktadır. Video verisinin artan büyüklüğü bu görevi zorlaştırmaktadır. Bu tez çalışmasında, video verilerinin çok kipli analizi için, görsel ve işitsel kiplerin füzyonuna dayalı bir yöntem önerilmektedir ve büyük veri platformunda uygulaması gerçekleştirilmektedir. Önerilen yöntem, Evrişimsel Sinir Ağı (ESA) öznitelikleri ile Mel-frekansı Kepstrum Katsayıları (MFCC) özniteliğinin temsillerinin füzyonuna dayanmaktadır. Büyük veri platformlarından Apache Spark kullanılarak önerilen yöntem gerçeklenmektedir. Önerilen yöntemin başarısı TRECVID 2012 SIN veri kümesi üzerinde değerlendirilmektedir. Sonuçlar göstermektedir ki, çok kipli yaklaşım tek kipli yaklaşımın başarısını geliştirmekte ve büyük veri platformu, çok kipli video içerik analizi yönteminin işlem zamanını önemli oranda düşürmektedir.

ANAHTAR SÖZCÜKLER: Çok Kipli Video Kavram Sınıflandırması, Evrişimsel Sinir Ağları (ESA), Mel-frekansı Kepstrum Katsayıları (MFCC), Destek Vektör Makineleri (DVM), Apache Spark, Büyük Veri, Derin Öğrenme.

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

ABSTRACT

MULTIMODAL VIDEO CONCEPT CLASSIFICATION

Berkay Selbes

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

The multimedia data has been continuously produced and shared out at a high rate as a result of the internet usage escalation. Thus, the size of multimedia data has rapidly increased, and hence, automated methods are needed to analyze the contents of the data produced. Video data is an important component of multimedia data. Video content analysis is an important research topic for several applications, such as audio-video based surveillance, content-based search and retrieval and can be defined as the automatic determination of temporal or spatial events/concepts in content of video data. Video content analysis is a difficult task due to the complex nature of the video content and requires efficient algorithms for extraction of high-level information included in the content. The increasing size of video data makes this task more difficult. In this thesis, a method based on the fusion of audio-visual modalities for multimodal content analysis of video data is proposed and implemented on a big data platform. The proposed method is based on the fusion of representations of Mel-frequency Cepstral Coefficient (MFCC) features with Convolutional Neural Network (CNN) features. The proposed method is implemented on Apache Spark big data platform. The success of the proposed method is evaluated on the TRECVID 2012 SIN data set. Our results show that the multi-modal method improves the accuracy of the single-model approach and also the big data platform significantly reduces the computation time of the multi-modal video content analysis method.

KEYWORDS: Multimodal Video Concept Classification, Convolutional Neural Network (CNN), Mel Frequency Cepstral Coefficient (MFCC), Support Vector Machine (SVM), Apache Spark, Big Data, Deep Learning.

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

İÇİNDEKİLER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

KISALTMALAR

DVM	Destek Vektör Makineleri
EC2	Elastic Compute Cloud
ESA	Evrişimsel Sinir Ağları
HOG	Histogram of Oriented Gradients
MFCC	Mel-Frequency Cepstral Coefficients
SIFT	Scale Invariant Feature Transform
TBA	Temel Bileşen Analizi
YSA	Yapay Sinir Ağları

1 GİRİŞ

Teknoloji devriminin hızla ilerlemesiyle birlikte akıllı sistemlerin hayatımızdaki yeri büyük önem kazanmıştır. Akıllı sistemler güvenlik, sağlık, eğlence, gibi alanlarda yaşam standartlarımız yükseltmek için kullanılmaktadır. Akıllı sistemlerde çokluortam verileri büyük önem taşımaktadır. Çokluortam metin, ses, durağan resim, animasyon, video ve diğer etkileşimli içeriklerin tek başına veya birlikte bulunduğu veri türüdür. İnsanların gördüğünü, duyduğunu ve okuduğunu anlamlandırabilmesi gibi çokluortam verilerini anlamlandırabilen sistemleri geliştirmek üzere çalışmalar yapılmaktadır. Çokluortam verilerinden video verisini kullanarak, insanlara hizmet eden, içerik tanımlama [1], yüz tanıma [2] ve plaka tanıma [3] gibi çeşitli sistemler bulunmaktadır. Böyle sistemleri geliştirebilmek için ilk adımlardan birisi, video verilerinin içeriklerinden anlamsal bilgilerin elde edilmesidir. Video içerik analizi, video verisi içeriğindeki zamansal ve/veya konumsal olayların/kavramların otomatik yöntemlerle belirlenmesi olarak tanımlanabilen önemli bir araştırma konusudur.

Güvenlik alanında video verisini kullanan sistemlerin birçok örneğine rastlanmaktadır. Devletler suç önleme ve suç soruşturma gibi nedenlerden önemli bölgelere kameralar yerleştirmektedirler. Kameralardan gelen video verisini, insan gücü veya bilgisayar sistemleri kullanarak, kırmızı ışıkta geçme [4], hız sınırına uymama [5] gibi kural ihlallerini denetlemektedirler. Artan nüfus ve büyüyen şehirler ile birlikte kamera sistemleri çoğalmakta ve denetimler insan gücü ile çözülememektedir. Bu sebeple, bilgisayarların hızlı işlem kapasitesini kullanarak, denetimleri insanlar yerine bilgisayarlara yaptırılmak istenmektedir. Bilgisayarlar, kameralardan gelen video verisini Bilgisayarlı Görü, Görüntü İşleme ve Makine Öğrenme tekniklerini kullanarak bu probleme çözüm getirebilirler.

Güvenlik alanında olduğu gibi eğlence alanında da çokluortam verilerini kullanarak insanlara hizmet eden sistemlere rastlanmaktadır. Akıllı telefonların yaygınlaşması, çoğu kişinin İnternet'e her an bağlanabilmesi ve sosyal medya kullanımının artırması ile birlikte her gün imge, video ve ses veriler kaydedilerek

paylaşılmaktadır. Örneğin YouTube[1] web sitesine her dakikada ~100 saatlik video yüklenmektedir [6]. Sosyal medya şirketleri, reklamları ilgili kullanıcılara ulaştırmak, telif hakkı gereksinimlerini yerine getirmek ve kanuna aykırı, vahşet içeren, terörü öven paylaşımları kaldırmak gibi sorumlulukları vardır. Bu sorumluluklardan dolayı paylaşılan çokluortam verilerinin içeriği hakkında bilgi edinmek isterler. Büyük ölçekteki verilerin barındırdığı bilgileri insan gücü ile çıkarmak neredeyse imkansıza yakındır. Bazı araştırmalara göre çokluortam verileri, İnternet trafiğinin %60'ını, cep telefonu trafiğinin %70'ini, yapısal olmaya verilerin %70'ini oluşturmaktadır [7]. Bu büyüklükte bir verinin işlenmesi, depolanması geleneksel bilgisayar sistemleri için de zordur. Bilim adamları, mühendisler bu problemlerin üstesinden gelmek için birden fazla bilgisayarın birlikte, aynı amaç için çalışabildikleri, dağıtık bilgisayar sistemleri üzerinde çalışmaktadır. Bu nedenle çokluortam verileri büyük veri platformlarının önemli kaynaklarından sayılmaktadır.

Çokluortam verilerinden anlamsal bilgiyi çıkarmak uzun süredir üzerinde çalışmaların yürütüldüğü önemli bir araştırma konusudur. Her çokluortam verisi için bir birinden ayrı veya birlikte kullanıldığı uygulamalar vardır. Çokluortam verilerinden anlamsal veriyi çıkarmak için ses verisini kullanan uygulamalara örnek vermek gerekirse; ses algılama, ses tanıma ve çokluortam olay sezimi gibi uygulamalar karşımıza çıkmaktadır. Sadece ses verisini kullanan sistemlerde olduğu gibi sadece görsel çokluortam verisini kullanan uygulamalar da mevcuttur. Bu uygulamalara görsel nesne sınıflandırma, hareket algılama gibi sistemler örnek olarak verilebilir. Video verisi görsel ve işitsel verileri barındırabildiği için içerik analizi yapılırken her iki kipinde önemli katkıları vardır. Video verisinin içerik analizinin bir araştırma konusu da video verisinin içerdiği kavramları, nesneleri otomatik olarak algılamadır. Video verilerinden anlamsal bilgi çıkarımı, çözülmesi gereken zorlu problemleri barındırır. Bu zorluklara örnek olarak, bir imge içerisinde belirlenen kavram farklı duruş pozisyonlarında, farklı mesafelerde, başka bir nesnenin arkasında var olması gösterilebilir. Bu gibi nedenlerden kontrolsüz

1 https://www.youtube.com/

ortamlardan toplanan video verisinin görsel kipinin içerdiği bilgiyi çıkarmak zor bir işlemdir. Aynı şekilde ses verisi için belirlenen kavramın bulunduğu ses verisinde, kavrama özgü ses dışında gürültüler veya ses alıcısının kavrama göre farklı pozisyonlarda olması gibi veri içerisindeki bilgiyi çıkarmayı zorlaştıracak zayıflıkları vardır. Bu gibi sorunların bazılarını aşmak için ses ve görsel verinin birlikte kullanıldığı yaklaşımlar vardır. Bu türdeki, birden fazla veri kaynağının birlikte kullanıldığı yaklaşımlar çok kipli yaklaşımlar olarak adlandırılmaktadır.

Video verisi görsel ve işitsel kipleri içerisinde barındırabilen çok kipli çokluortam verisi olarak bilinmektedir. Video verisinin içerdiği anlamsal bilginin analizi için görsel ve işitsel kipler birbirinden ayrı veya birlikte kullanılabilir. Görsel kip bir nesnenin görünüşü ve zamanda akan görsel hareketi gibi bilgiler barındırır. İşitsel kip ise arka plan sesleri, konuşmalar ve nesneye özgü sesler barındırabilir. Görsel ve işitsel kipler birbirleri için tamamlayıcı olarak bilinirler [8]. Bir çok kavramı içerebilen video verisi için, başarılı video içerik analizi sistemleri genellikle video verisinin özniteliklerine dayanmaktadır. Görsel ve işitsel öznitelikler çok sayıda kavramı içerebilen video verisinin barındırdığı bilgiyi temsil ederler. Çok kipli yaklaşımlarda, içeriği oluşturan her kipten ayrı öznitelikler çıkarılabilmektedir. İşitsel kip için öznitelik çıkarma tekniklerine Mel-frekansı Kepstrum Katsayıları (MFCC) [9] örnek olarak verilebilir. MFCC, ses sinyalinin kısa zamanlı güç spektrumunu temsil eder. MFCC sadece ses verisini kullanan konuşma tanıma gibi uygulamalarda gösterdiği başarılı performans ile popülerliğini sürdürmektedir. Video verisinin görsel kipi nesnenin görünüşü ve zamanda akan görsel hareketi hakkında bilgi içerdiği için bu iki özellik içinde ayrı ayrı öznitelik çıkarma teknikleri vardır. Nesnenin görünüşü hakkında bilgi edinmek için videoyu oluşturan görüntü kareleri üzerinde işlem yapılır. Videoyu oluşturan görüntü karesi üzerinde işlem yapıldığı için göreceli olarak düşük hesaplama maliyeti vardır. Görüntü karesinin içerdiği belirlenen nesne hakkındaki öznitelikleri çıkarmak için Scale İnvariant Feature Transform (SIFT) [10], Speeded-Up Robust Features [11], Oriented FAST and Rotated and BRIEF [12] vb. yaygın olarak kullanılmaktadır. Aynı zamanda SIFT metodunun farklı amaçlar için farklı biçimleri kullanılmaktadır. Örneğin

görüntü karesindeki nesnelerin renk bilgilerini modellemek için color-SIFT kullanılmaktadır. Histogram of Oriented Gradients (HOG) [13] metodu da sıkça görüntü karelerinden öznitelik çıkarmak için kullanılmaktadır. Son zamanlarda popülerliğini artıran Evrişimsel Sinir Ağı (ESA) [14][15] mimarileri de öznitelik çıkarmak için kullanılmaktadır. ESA bir Yapay Sinir Ağı olup temelleri 1980'lere dayanmaktadır. ESA derin öğrenme mimarilerinin özel bir biçimidir. ESA mimarilerinin popülerliği, Alexnet [14], GoogleNet [15] gibi mimarilerin ImageNet [16] veri kümesinde gösterdiği başarı ile birlikte artmıştır. Aynı zamanda, ESA mimarileri ImageNet gibi büyük veri kümeleriyle bir kere eğitildikten sonra taşınabilir olması popülerliğinin artıran diğer bir etkendir. ESA mimarilerinin taşınabilir olması bu mimarilerin öznitelik çıkarımı gibi farklı amaçlar için kullanılmasına olanak sağlar.

Video içerik analizi için sadece işitsel veriyi kullanan, sadece görsel veriyi kullanan veya işitsel ve görsel verilerin birlikte kullanıldığı çok kipli yaklaşımlar mevcuttur. Çok kipli yaklaşımlar, çıkartılan işitsel ve görsel özniteliklerin füzyonuna dayalı yöntemlerdir. Çeşitli füzyon yöntemlerinden sonra işitsel ve görsel öznitelikler birlikte anlamlı şekilde kullanılabilir. Böylece video verisinin işitsel ve görsel kiplerinin birbirlerine olan tamamlayıcı etkilerinden faydalanılabilir. Füzyon işlemi öznitelik matrislerinin veya vektörleri için çeşitli matematiksel işlemler kullanarak, uç uca eklenerek, vektörler arasındaki ilişkiyi inceleyerek veya sınıflandırıcıların kararları arasındaki ilişkiyi inceleyerek vb. yapılabilir [17]. Füzyon işlemleri sonucunda bir kipin eksik kaldığı noktalarda diğer kipin özellikleri bu noktalarda tamamlayıcı olabilir ve sistemlerin performanslarına artı yönde etki edebilir.

Video verisinin kiplerinden öznitelikler çıkartıldıktan sonra problem belirlenen kavramlar için sınıflandırma problemine dönüşür. Sınıflandırma problemini çözmek için Makine Öğrenme teknikleri [18] kullanılabilir. Bu sınıflandırıcılar özniteliklerin veri uzayında gösterdiği özelliklere göre farklı başarılar sergileyebilirler. Destek Vektör Makineleri (DVM) [19] basitlikleri ve başarıları sayesinde video içerik analizi için hareket algılama [20] , nesne tanıma [21] vb. uygulamalarda sıkça kullanılmaktadır. DVM veri uzayında iki sınıfı birbirinden ayıracak en optimal hiper

düzlemi bulma motivasyonu ile sınıflandırma işlemini gerçekleştirir. DVM doğrusal olarak ayrılabilen veri dağılımı gösteren iki sınıfı birbirinden ayırmak için tasarlanmıştır. Fakat uygulanan Bire Karşı Bir gibi stratejiler ile çoklu sınıflandırma problemlerine de uygulanabilmektedir. Birbirinden doğrusal olarak ayrılamayan veri dağılımları için çekirdek fonksiyonu stratejisi kullanarak, veri uzayını bulunduğu boyuttan üst boyutlara taşıyarak, ayırmaya çalışır. Bulunan veriler ile DVM eğitildikten sonra oluşan model ile görülmemiş veriler sorgulanabilir. Video verisinin görsel ve işitsel kiplerden elde edilen öznitelikler, füzyon işlemi olsun veya olmasın, veri uzayında bir dağılım sergilerler. DVM, video içerik analizi için elde edilen öznitelikleri belirlenen kavramlara göre sınıflandırmak için kullanılır.

Video verisinin kiplerine ayrılması, kiplerinden öznitelikler çıkartılması ve sınıflandırıcıların eğitilmesi gibi nedenlerle, video içerik analizinin hesaplama maliyeti geleneksel veri türlerine kıyasla görece yüksektir. Günlük hayatımızdaki problemleri çözmek için sürekli ve çok sayıda kaynaktan gelen video verisinin içeriği analiz edilmek istenebilir. Büyük ölçekte gelen video verisi için video içerik analizi sistemleri uygulanabilir olmalıdır. Bu sebeple, video içerik analizi sistemleri sürekli gelen veriyi veya yüksek boyutlara ulaşan yığın halindeki video verisinin analizini gerçekleştirebilmek için büyük veri platformlarına ihtiyaç duyarlar. Büyük veri platformları dağıtık, paralel, küme hesaplama yaparak işlem kapasitesini ölçeklenebilir şekilde artırırlar. Büyük veri platformlarına örnek olarak Apache Spark [22] gösterilebilir. Apache Spark hızlı ve genel amaçlı küme hesaplama sistemidir [23]. Apache Spark birden fazla düğümü bulunan bilgisayar kümelerinde paralel şekilde çalışacak uygulamanın basitçe yazılmasını amaçlamaktadır.

Bu tez kapsamında, video kavram sınıflandırma problemi için çok kipli bir yaklaşım önerilmektedir. Önerilen yöntem videonun işitsel kipinden elde edilen MFCC öznitelik matrisinin istatistiksel temsilleri ile videonun görsel kipinden elde edilen ESA özniteliklerinin füzyonuna dayanmaktadır. Aynı zamanda, önerilen sistem Apache Spark büyük veri platformu üzerinde gerçekleştirilmiş ve hesaplama karmaşıklığına olan etkisi incelenmiştir. Bu tezde sunulan çalışmalar literatüre aşağıdaki açılardan katkıda bulunmaktadır:

Multimodal video concept classification based on convolutional neural network and audio feature combination [24].

Multimodal vehicle type classification using convolutional neural network and statistical representations of MFCC [25].

1.1 Tezin Organizasyonu

Bu tezin organizasyonu şu şekilde düzenlenmiştir; Bölüm 2’de Literatür taraması verilmiştir. Tez çalışması ile ilgili genel kavram ve tanımlar Bölüm 3'de anlatılmaktadır. Video kavram sınıflandırma problemi için önerilen yöntem Bölüm 4’de tanıtılmaktadır. Bölüm 5’de yapılan deneysel çalışma ve elde edilen sonuçlar değerlendirilmektedir. Bölüm 6’da deneysel sonuçlar ve tartışma sunulmaktadır.

2 LİTERATÜR TARAMASI

Literatürdeki video içerik analizi yöntemleri kullanılan veri türüne göre üç başlık altında toplanabilmektedir. Bu başlıklardan birincisi görüntü ve akan görüntü verilerini kullanan görsel tabanlı yaklaşımlar. İkincisi ses verisini kullanan işitsel tabanlı yaklaşımlar. Üçüncüsü ise hem ses hem görsel verileri kullanan çok kipli yaklaşımlardır. Dördüncü başlık ise video içerik analizi için büyük veri teknolojilerini kullanan çalışmalar olarak belirlenebilir.

Video içerik analizi aslında çok genel bir tanım olup altında bir çok görevi barındırır. Bu görevlere örnek olarak video sınıflandırma, video tanıma, nesne tanıma, nesne sınıflandırma, hareket tanıma, olay tanıma gibi konuları örnek verilebilir. Aynı zamanda, hazırlanan veri kümesine göre video içerik analizini farklı başlıklarla sınıflandırılabilir. Tez çalışması ile ilgili olduğunu düşündüğümüz çalışmalar aşağıda dört başlık altında sunulmaktadır.

2.1 Görsel Tabanlı Yaklaşımlar

Görsel tabanlı yaklaşımlarda video verisi görsel kipi üzerinde işlemler yapılır. Görsel kipdeki görüntü kareleri bir birinden ayrı işleme sokulursa statik görünüş öznitelikleri elde edilebilir. Görüntü kareleri beraber ve zamansal düzlemde kullanıldığı zaman ise hareket öznitelikleri elde edilir. Görsel tabanlı yaklaşımlar genellikle bu iki öznitelik üzerine yoğunlaşmışlardır. Video içerik analizinde görsel tabanlı yaklaşımlar genel olarak video verisinden görsel özniteliklerin çıkarılmasına dayanan yöntemlerdir. İlk yaklaşımlar global özniteliklere dayanmaktaydı [26] [27]. Videonun anahtar çerçevelerinden elde edilen global özniteliklerin (şekil, desen, renk histogram, vb.) Makine Öğrenme algoritmaları ile birlikte kullanıldığı bu yaklaşımlar, anahtar çerçeveyi genel olarak tanımladığından, nesne tanıma gibi problemlerdeki başarımları yerel özniteliklere kıyasla düşük olabilmektedir. Bu nedenle yerel öznitelik tabanlı yöntemler nesne tanıma gibi problemlerde tercih edilebilmektedir. Bu yöntemlerde ilk olarak SIFT [10], HOG [13] gibi yöntemler kullanılarak video verisinin görsel kipinden öznitelikler çıkartılır. Bu öznitelikler K-

ortalama (K-means) gibi teknikler kullanılarak görsel sözcüklerden oluşan bir sözlük oluşturulur. Daha sonra kelime kümesi gibi yaklaşımlar ile temsiller oluşturulur ve sınıflandırıcılar eğitilir. Bu video içerik analizi için genel olarak uygulanan yöntemdir.

Video görsel kipinden ayrıştırılan görsel çerçevelerden, statik görünüm özelliği üzerinde yapılan çalışmalar, imge veri kümeleri üzerinde yapılan çalışmalar ile benzerlik göstermektedir. Bu sebeple imge veri kümelerindeki çalışmalar, video içerik analizi için de katkı sağlarlar.

Csurka vd. [21], çalışmasında nesne kategorilendirme üzerine çalışmışlardır. Bu yaklaşımda ilk aşamada çeşitli detektörler ve SIFT tanımlayıcısı kullanarak imgelerden öznitelik elde etmektedir. Daha sonra vektör niceleme (vector quantization) algoritması kullanarak, öznitelikleri kümelemiş ve görsel sözcükler elde etmişlerdir. Bundan sonra ise kelime kümesi tabanlı yaklaşımla oluşturulan kümeleri temsillerini elde edilmektedir. En son aşamada ise çok sınıflı sınıflandırıcı olarak Naive bayes ve DVM sınıflandırıcı kullanılmaktadır.

J.Zhang vd. [28], çalışmasında metin ve nesne sınıflandırma üzerine çalışmışlardır. Çalışmasında Harris- Laplacian ve Laplacian detektörleri ile SIFT ve SPIN [29] tanımlayıcılarına kombine ederek görsel öznitelikler elde edilmektedir. Sınıflandırıcı olarak DVM seçilmiştir. DVM için Earth Mover's mesafesi ve X^2 çekirdek fonksiyonları kullanılmaktadır.

Jingen vd. [30], çalışmasında video verisinden gerçekçi hareketleri algılama için bir yaklaşım önermektedir. Bu yaklaşımda statik öznitelik için Harris-Laplacian, Hessian-laplacian ve MSER [31] yerel dedektörlerini kullanmış ve her özniteliği (x,y) lokasyonu gamma scalası ve SIFT ile tanımlamıştır. Hareket özniteliği için ise, 2 boyutlu gaus filtresini, Dollar'ın [32] çalışmasında sunduğu yaklaşımı kullanmaktadır. Hareket özniteliği Temel Bileşen Analizi uygulayarak boyut indirgemektedir. Daha sonra hareket öznitelikleri için hareket istatistiklerini kullanarak bir budama işlemi gerçekleştirmiştir. Statik öznitelik için ise hareket

özniteliğinin tanımladığı ilgili bölgeler için bir budama işlemi gerçekleştirilmektedir. Daha sonrada statik öznitelikler için PageRank algoritması kullanılarak en ayırt edici bilgi içeren öznitelikler seçilmiştir. Bu özniteliklerden görsel sözcükler elde etmek amaçlı bilgi teorisi metrikleri kullanmaktadır. Daha sonrada AdaBoost algoritması kullanılarak sınıflandırıcı eğitmişlerdir. Bu yaklaşımlarını ise KTH [33] veri kümesinde ve Youtube'dan kendileri oluşturdukları 11 kategoriden oluşan bir veri kümesinde denenmektedir. Elde edilen sonuçları K-ortalama ile elde edilen sonuçlar ile karşılaştırılmaktadır. Bu sonuçları ortalama doğruluk olarak belirlemişlerdir. Kelime boyutu arttıkça iki yöntem içinde performansında artış gözlemlenmiştir. Önerilen yöntem K-ortalama yöntemine göre daha başarılı bir performans sergilemektedir.

Ergün ve Sert [34], çalışmasında video sahne sınıflandırma üzerine çalışmaktadır. Video sahne sınıflandırma uygulaması için video verisinin görsel kipinden SIFT özniteliği çıkartmaktadır. Algılayıcı olarak Difference of Gaussian algılayıcısı tercih etmişlerdir. Elde edilen öznitelikler uzamsal piramit gösterimi ile temsilleri yaratılmıştır. Sınıflandırıcı olarak DVM kullanmaktadır. Çalışmalarında DVM parametreleri, uzamsal piramit seviyesinin, örnekleme parametrelerinin ve sözcük sayısının sınıflandırma performansına etkisini gözlemlemişlerdir. Sözcük sayısının belirli bir noktadan sonra performansa katkısı olmadığı, DVM X^2 çekirdek fonksiyonu en optimal sonucu elde ettiği gözlemlenmektedir.

ESA mimarilerinin, arttan hesaplama gücü ve ESA mimarilerini eğitebilecek büyüklükte veri kümelerinin var olması sayesinde, son zamanlarda başarısını ve popülerliğini artırmıştır. Video içerik analizi ve bilgisayarlı görü uygulamalarında sıkça kullanılmaya başlanmıştır.

Krizhevsky vd. [14], çalışmasında imge sınıflandırma görevi için 3 tam bağlı, 5 evrişim katmanı bulunan bir ESA mimarisi eğitmiştir. Aşırı eğitimden kaçınmak için ise tam bağlı katmanlarda dropout dedikleri bir düzenleme metodu geliştirmişlerdir. 2010 yılındaki ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) [16] veri kümesi üzerinde eğittikleri ESA mimarisini test edilmiştir. En iyi 1 ve en iyi

2 hata oranlarında sırasıyla %37.5 ve %17'lik bir başarı ile daha önce yapılan çalışmalardan daha iyi başarı göstermiştir.

Girshick vd. [35], çalışmasında nesne tanıma görevi üzerinde durmuştur. Bu görev için bölgesel olarak çalışabilmesi için ESA mimarisini geliştirmişlerdir. Bu mimariye bölgesel Evrişimsel Sinir Ağı ile isimlendirilmektedir. Önerdikleri yaklaşım 3 aşamadan oluşmaktadır. İlk olarak imgelerdeki bölgelerin belirlemesinde Uijlings [36] çalışmasında uygulamış olduğu gibi seçici arama metodu kullanarak, kategoriden bağımsız bölgeler belirlenmiştir. İkinci aşamada ise AlexNet [14] kullanarak belirlenen her bölgeden 4096 boyutunda öznitelik çıkmışlardır. Üçüncü aşamada ise sınıfa özel doğrusal DVM kullanmaktadır. Bu çalışmasını 2012 yılındaki VOC [37] veri kümesinde test etmişlerdir. Ortalama hassasiyet (mean avarge precision) ölçütü olarak %53.3'lük bir başarı ile daha önceki bu veri kümesi üzerindeki çalışmaları geride bırakmaktadır.

Szegedy vd. [15], çalışmasında imge sınıflandırma ve nesne tanıma üzerinde durmuştur. Bu amaçla parametresi bulunan 22 katmanı ve 5 birleştirme katmanından oluşan adına GoogLeNet dedikleri karmaşık ESA mimarisi eğitilmektedir. 2014 yılındaki ILSVRC yarışmasında birincilik elde ederek daha önceki çalışmalardan daha iyi başarı göstermiştir.

Karpathy vd. [38], çalışmasında video sınıflandırma üzerinde durmuştur. ESA mimarilerinin statik görünüm özelliği için gösterdiği başarılardan faydalanarak, video verisi için zaman düzleminde yerel uzay zamansal bilgileri elde etmek amaçlı çalışmaktadır. Bu çalışmasını UCF101 [39] veri kümesinde test etmiştir. Bu veri kümesi için belirlenen taban başarının üstünde yüksek bir başarı kaydetmektedir.

Razavian vd. [40], çalışmasında imge tanıma, sahne tanıma, ince taneli tanıma (fine grained recognition), özellik tanıma (attribute detection) ve imge getirme uygulamaları üzerinde çalışmaktadır. ESA mimarilerinin bir çok tanıma görevinde başarı sağlayacağına düşünerek bu çalışmayı gerçekleştirmişlerdir. Bu çalışmaları

yaparken overfeat [41] isimli ESA mimarisini kullanarak öznitelik çıkarmıştır. Çıkartılan öznitelik L2 normalizasyonu işlemi sonrasında DVM eğitimi için kullanılmaktadır. Aynı zamanda veri kümesindeki örnekleri kırpıp ve çevirip tekrar veri kümesine ekleyerek veri kümesindeki örnek sayısını artırmaktadır. Bu çalışmasında bir çok veri kümesi kullanmış bir çok başarılı sonuç elde etmiştir.

2.2 İşitsel Tabanlı Yaklaşımlar

Ses bilgisi videonun önemli bileşenlerinden birisidir. Bu nedenle, video içerik analizinde ses kipi de yaygın olarak kullanılmaktadır. Xu vd. [42], işitsel tabanlı yöntem kullanarak, spor videolarını analiz etmişlerdir. Spor videoları olarak futbol, tenis ve basketbol videolarını seçmişlerdir. Bu videolar içerisinden ıslık, seyirci, yorumcu, heyecanlı seyirci ve heyecanlı yorumcu kategorilerini ayırmaktadır. Bu işlemi gerçekleştirmek için ses verisinden MFCC öznitelikleri çıkarılmış ve bu öznitelikler ile gizli markov modelleri eğitmişlerdir. Daha önceki çalışmasında eğittikleri DVM ile karşılaştırmışlar ve gizli markov modelleri, DVM'den daha başarılı sonuç elde etmiştir.

Lee vd. [43], tüketici videoları için işitsel tabanlı kavram sınıflandırma üzerine çalışmaktadır. Çalışmalarında ses verisinden MFCC öznitelikleri çıkartılmaktadır. Elde edilen öznitelikleri tek gauss modeli, gauss karışım modeli ve gauss bileşenlerinin olasılıksal gizli anlam analiz yöntemleri ile temsillerini yaratmaktadır. Elde edilen temsiller ile DVM eğitmişlerdir. Youtube'dan elde ettikleri 1873 videodan oluşan veri kümesinde, belirledikleri 25 kategori üzerinde önerdikleri yöntemi gerçekleştirmişlerdir.

Lee vd. [44], video verisinde çevresel ses tanıma üzerine çalışmaktadır. Bu çalışmada ses verisinden MFCC öznitelikleri çıkarılmıştır. Çıkarılan öznitelikler gauss karışım modelli ile beraber gizli markov modelinin eğitiminde kullanılmaktadır. Youtube'dan elde ettikleri videolar ile hayvan, bebek, bot ve sevinç kategorilerini sınıflandırmaktadır.

Okuyucu vd. [45], çalışmasında çevresel sesleri tanıma üzerine çalışmıştır.

Çevresel ses tanıma uygulaması için ses verisinin MPEG-7 ailesi, ZCR, MFCC özniteliklerini çıkarmışlar ve kombinasyonlarını kullanmaktadır. Elde edilen öznitelikler ve kombinasyonları ile gizli markov modelleri ve DVM eğitmekte ve değerlendirilmektedir. Çevresel sesler olarak acil durum sireni, araba kornası, silah patlaması, araba, motosiklet, helikopter, rüzgar, su, yağmur, alkış, kalabalık, kahkaha kavramlarını seçmişlerdir. MPEG-7 ailesinden Audio Spectrum Flatness (ASF), Audio Spectrum Centroid (ASC), Audio Spectrum Spread (ASS), Audio Harmonicity (AH) öznitelikleri ile eğitilen DVM en iyi performansı sergilemektedir.

2.3 Çok Kipli Yaklaşımlar

Oneata vd. [20], çalışmasında video verisinden hareket ve olay tanıma için çok kipli bir yaklaşım önerilmiştir. Çalışmasında farklı veri kümeleri üzerinde bir çok değerlendirmede bulunmaktadır. Bu değerlendirmeleri 3 başlık altında toplamıştır. Bu başlıklardan birincisi, 5 veri kümesi üzerinde kısa hareketlerin sınıflandırılmasıdır. İkincisi filmlerde bazı hareketlerin yerinin belirlenmesidir. Sonuncusu ise karmaşık olayların geniş ölçekli tanınmasıdır. Ses ve görsel verilerin birlikte kullanıldığı yaklaşım bu değerlendirmelerden karmaşık olayların geniş skalalı tanınması değerlendirmesinde uygulamıştır. Uygulanan yöntemde, görsel verinin statik görünüm özelliğinden, SIFT öznitelikleri çıkartılmıştır. Hareket özelliğinden ise hareket sınır histogramı (Motion Boundary Histogram) [46] tekniği kullanılmaktadır. Bu iki görsel özniteliğin Fisher vektör [47] metodu kullanarak temsillerini elde etmiştir. Ses verisinden ise MFCC öznitelikleri çıkarılmıştır. Daha sonra bu üç öznitelik uç uca eklenerek, füzyon işlemine tabi tutulmuştur. Sınıflandırıcı olarak ise DVM eğitilmiştir. Veri kümesi olarak 2011 yılında kullanılan TRECVID MED veri kümesini tercih edilmektedir. Elde edilen sonuçlarda, 3 özniteliği birlikte kullanıldığı yöntem, hareket sınır histogramı özniteliğinin tek başına kullanıldığı, SIFT öznitelığinin tek başına kullanıldığı ve SIFT özniteliği ile hareket sınır histogramı özniteliğinin uç uca eklenerek uygulanan yöntemlerden daha başarılı sonuç elde etmiştir.

Jiang vd. [48], çalışmasında video içerik analizi için olay tanıma görevinde, çok

kipli bir yöntem önermektedir. Çalışmasında video verisinin görsel kipinden statik görünüm özelliği için SIFT öznitelikleri çıkartılmış, bu öznitelikler K-ortalama ile görsel kelimeler elde edilmiştir. Hareket bilgisi için ise uzamsal-zamansal ilgi noktalarını (spatial-temporal interest points), Laptev [49] metodu ile gerçekleştirmiştir. Elde edilen hareket öznitelikleri HOG ve Histogram of Optical Flow metotları ile tanımlanmış ve elde edilen tanımlamalar uç uca eklenmiştir. İşitsel kip için ise ses verisinden MFCC öznitelikleri çıkarılmaktadır. Daha sonra ise bu üç öznitelik kelime kümesi tabanlı yaklaşım ile temsil edilmiştir. Her öznitelik için ayrı DVM eğitilmiştir. DVM çekirdek fonksiyonu X^2 seçilmiştir. DVM tahmin olasılıkları ortalama değer ile füzyon işlemi uygulanmaktadır. Çalışmasını 2010 yılındaki TRECVID MED veri kümesi üzerinde gerçekleştirmiştir. Üç özniteliği birlikte kullanıldığı yöntem, her özniteliği ayrı ve STIP ve SIFT füzyonuna dayalı yöntemden başarılı sonuç elde etmiştir.

2.4 Büyük Veri Teknolojileri Kullanan Çalışmalar

Tan vd. [50], çalışmasında video verisinde yüz tanıma ve nesne tanıma ve takip etme uygulamalarını Hadoop [51] küme ortamında gerçekleştirmişlerdir. Bu uygulamaların bilgisayar küme boyutuna ve video sayısına göre sonuçlar almaktadır. Kümeyi oluşturan toplam bilgisayar sayısı altı olarak seçmiştir. Video işleme algoritmaları için OpenCV [52] ve FFmpeg [53] kütüphanelerini tercih edilmiştir. Küme boyutu arttıkça işlemler için harcanan sürenin azaldığını gözlemlemiştir.

Yang vd. [54], çalışmasında video verisinde hareket tanıma uygulamasını Apache Spark büyük veri platformunda gerçekleştirmiştir. Hareket tanıma uygulaması için STIP detektörü kullanarak çıkarılan öznitelik, HOG tanımlayıcı ile tanımlamıştır. Sonraki işlemde ise kelime kümesi tabanlı vektörler elde edilmiştir. Elde edilen vektörler DVM eğitiminde kullanılmıştır. Apache Spark kümesinde işlemlerin zamana göre kıyaslamasını bilgisayar çekirdeklerine göre belirlenmiştir. Bilgisayar çekirdeklerini 6, 12, 18, 24 olarak belirlenmiştir. Hareket algılama için gerekli işlemler olan öznitelik çıkarımı, görsel kelime üretimi, kelime kümesi işlemleri için

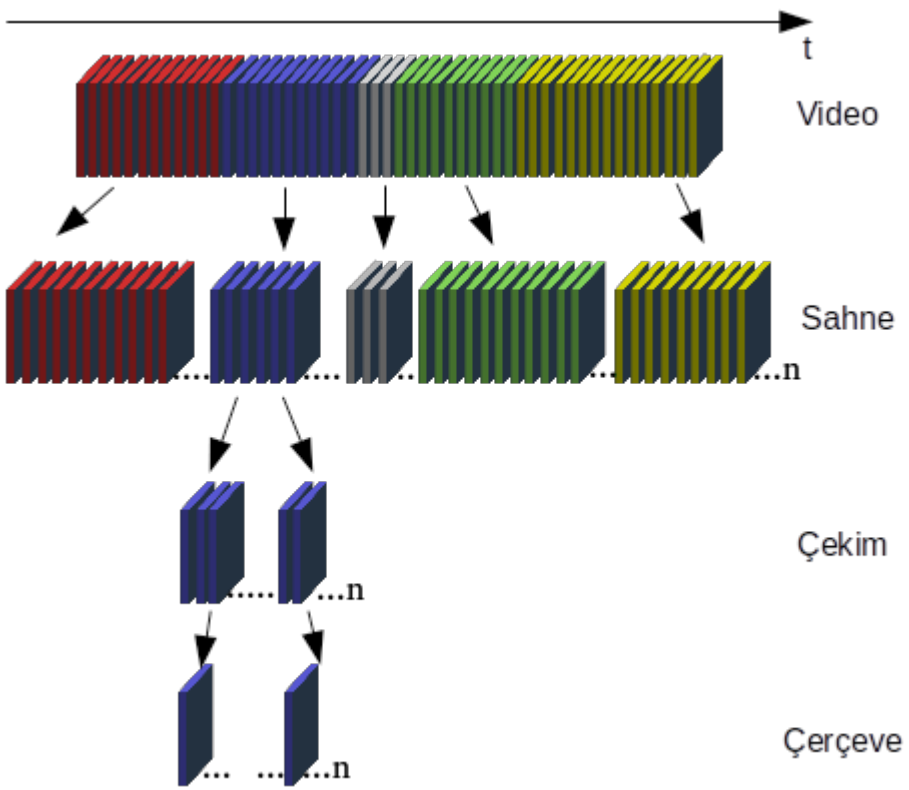
bilgisayar çekirdeklerini zamana göre performanslarını test edilmiştir. Veri kümesi olarak KTH [33] ve Holywood2 [55] veri kümeleri kullanılmıştır. Video kütüphaneleri olarak OpenCV ve FFmpeg kullanılmıştır. DVM kütüphanesi için libSVM kullanılmıştır. Çekirdek sayısı arttıkça performansın artığı gözlemlenmiştir.

Wang vd. [56], çalışmasında video verisi için hareket tanıma uygulamasını, Apache Spark büyük veri teknolojisini kullanarak, gerçekleştirmiştir. Hareket tanıma uygulaması için video verisinin görsel kipinden yörünge (Trajectory) tabanlı öznitelik çıkarmıştır. Çıkarılan öznitelik Gaus karışım modeli ile modellenmiş daha sonra ise Fisher vektör yöntemi ile temsilleri yaratılmıştır. Bu çalışmasını Spark küme ortamında gerçekleştirmek için 9 düğümden oluşan bilgisayar kümesi oluşturulmuştur. Sonuçlarını, tek düğüm ve küme boyutunda uygulanan işlemlerin zamana karşı performanslarını test ederek alınmıştır. Aynı zamanda küme boyutunda uyguladığı işlemler için, öznitelik çıkarımı sonrası elde edilen verilerin Hadoop dağıtık dosyalama sistemine kaydederek veya kaydedilmeden doğrudan Fisher vektör işlemi uygulanmış hali ile performans testleri gerçekleştirilmiştir. Küme boyutunda uygulanan işlemler tek düğümde gerçekleşen işlemler arasında bariz bir performans farkı belirlenmektedir. Hadoop dağıtık dosyalama sistemine kaydetmeyerek doğrudan Fisher vektör işlemi uygulanan yöntem, kaydederek uygulanan yöntemden daha iyi performans sergilemiştir.

3 TEMEL TANIM VE KAVRAMLAR

3.1 Sayısal Video

Sayısal video bir dizi imge veya çerçevenin sabit hızda oynatılmasıyla oluşmaktadır [57]. Sayısal videolar mantıksal olarak ayrılmış bölümlerden meydana gelmektedir. Sayısal video içeriğinin hiyerarşik olarak düzenlenmesi Şekil 3.1'de sunulmuştur. Sayısal videoyu oluşturan bağımsız imgelere çerçeve denmektedir. Mantıksal olarak birbiri ile ilişkili çerçeveler birleşerek çekimleri (shot)



Şekil 3.1 Video içeriğinin hiyerarşik temsili

oluşturur. Çekimleri oluşturan çerçeveler bazı ortak özelliklere sahiptir. Bu özellikler çekimler aynı sahneyi tasvir ediyor olmalıdır, tek bir kamera operasyonunu belirtmelidir, görüntü içerisinde arka plan değişmemelidir. Bulunduğu çekimi en iyi temsil eden çerçeveye anahtar çerçeve denir. İçerik olarak birbiri ile ilişkili çekimler birleşerek sahneleri oluştururlar. Sayısal video sahnelerin

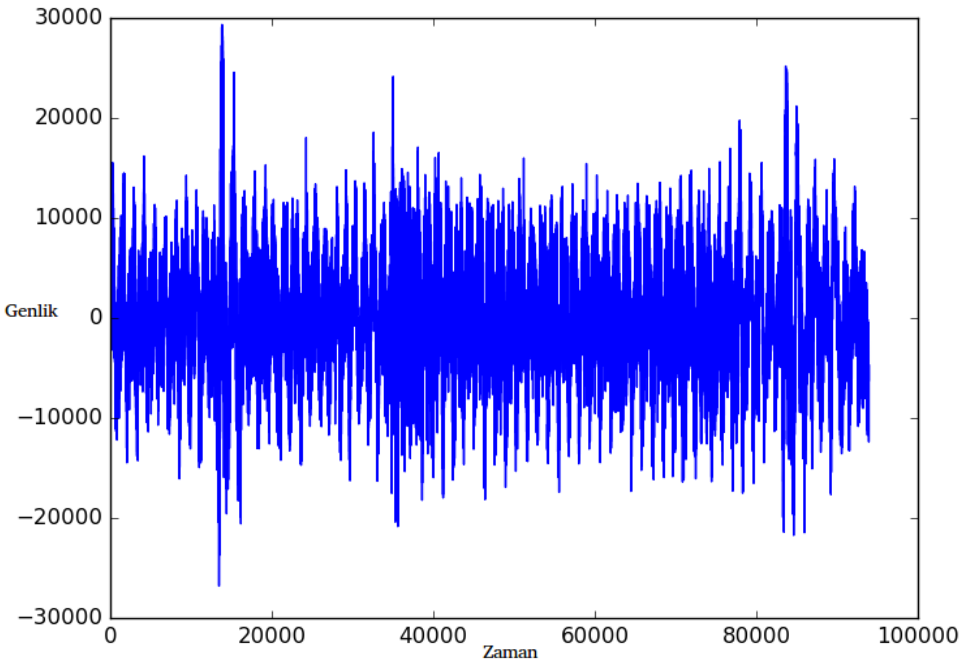
birleşiminden oluşmaktadır. Anahtar çerçeveler video çekimini en iyi temsil eden çerçeve olduklarından, içerdikleri bilgi çekimler hakkında yorum yapmamızı sağlar. Video içerik analizi sistemlerinin performanslarını ölçmek için çevrim içi yayınlanan video veri kümelerinden faydalanılır. Video veri kümeleri hazırlanırken anahtar çerçeveler ve bulunduğu çekimlerin etiketlenmesi önemlidir. Videoyu oluşturan farklı çekimler farklı kategorileri barındırabilirler. Sayısal video içeriğinin hiyerarşik olarak hangi bölümlerden oluştuğunu bilmek, video içerik analizi üzerine çalışanlar için önemlidir.

Videonun çerçeveleri belirlenen kategorinin görünüşü hakkında bilgi barındırır. Çerçevelerde kategoriler farklı duruş pozisyonlarında, farklı büyüklüklerde, farklı kategoriler ile birlikte ve kategoriye özgü farklı hallerde bulunabilir. Video içerik analizi sistemlerinin performansının artırmak ve kategoriyi en iyi temsil eden yapıyı çıkarabilmek için kategorilerin bulunabileceği her halinden oluşan çerçevelere ihtiyaç duyulur. Böylece Görüntü İşleme, Bilgisayarlı Görü ve Makine Öğrenme teknikleri ile kategoriyi diğer kategorilerden ayırt edebilecek sistemler oluşturulabilir.

3.2 Ses

Ses, bir iletim ortamında (gaz, sıvı, veya katı) duyulabilir bir basınç dalgası olarak yayılım gösteren bir titreşim olarak tanımlanabilir. [57]. Sürekli ses dalga formu mikrofonlar aracılığı ile sürekli bir elektrik sinyaline dönüştürülür. Bilgisayarlar ile bir sesi işlemek ve iletmek için, sürekli elektrik sinyalini sayısal ses sinyaline dönüştürülmelidir. Video verisini kaydeden kameralarda mikrofon bulunuyorsa video verisinin görsel kipi ile eş zamanlı şekilde ses verisi videoda bulunur. Bu şekilde ortamdaki seslerden de bilgi edinilebilir. Video verisinin içeriğindeki ses sinyalinden içerik analizine katkıda bulunacak bilgiler çıkarmak için, videoda belirlenen çerçeveye karşılık gelen belirli süredeki zaman diliminde ses sinyali elde edilebilir. Bu ses sinyali videonun kaydedildiği ortamdaki belirlenen kavrama özgü sesleri barındırabilir. Örneğin Şekil 3.2'de TRECVID veri kümesinden *Alvaromelo-SundayRide712_512kb.mp4* video klibinden elde edilen ses bileşeninin motosiklet

kavramı için zaman-genlik temsili gösterilmektedir. Kavrama özgü ses sinyaliyle birlikte ortamdaki gürültüler ve başka kavramlara özgü sesleri barındırabilir. Aynı zamanda ses alıcılarının kavrama olan pozisyonlarındaki farklılıklar, farklı kalitede ses alıcılarından toplanan ses sinyalleri gibi nedenlerden kavrama özgü ses sinyalleri farklı özellikler sergileyebilirler. Bu gibi nedenlerden video içerik analizi



Şekil 3.2 Video klibinden elde edilen ses bileşeninin zaman-genlik grafiği

sistemlerinin performansını artırmak için ses sinyalini farklı ortamlardan, alıcılardan toplanmalıdır. Böylece kavrama özgü sesi en iyi şekilde karakterize eden yapı ses işleme teknikleri ile ifade edilebilir.

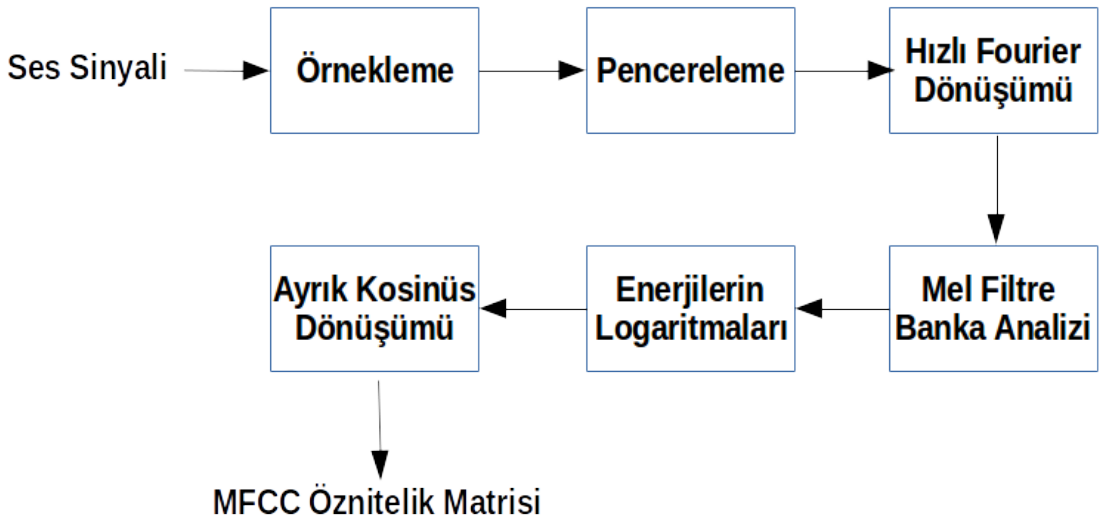
3.3 Öznitelik Çıkarımı

Başarılı video içerik analizi sistemleri genellikle video verisinin işitsel ve görsel öznitelikleri çıkarılarak yapılmaktadır. Öznitelikler analiz edilmek istenen görsel veya işitsel kipten elde edilen ve bu kipleri en iyi şekilde karakterize eden değerler kümesi olarak ifade edilir. Çıkarılan öznitelikler, bilgi edinilmek istenen kavramı olabileceği en iyi şekilde karakterize etmelidir. Özniteliklerin kalitesi doğrudan

video içerik analizi sistemlerinin performansına etki etmektedir. Açıkça ifade edilmediği sürece video içerik analizi sistemlerinde tanımak istenen her kavram için aynı öznitelik çıkarımı teknikleri kullanılmaktadır. İşitsel ve görsel kiplerin birbirinden farklı öznitelik çıkarımı teknikleri mevcuttur. Bu tez kapsamında videonun işitsel kipi için MFCC öznitelik çıkarımı tekniği kullanılacaktır. Videonun görsel kipi için ise ESA mimarilerinden öznitelikler çıkartılacaktır.

3.4 Mel-frekansı Kepstrum Katsayıları

MFCC ses tanıma, müzik algılama gibi uygulamalarda gösterdiği başarılardan dolayı ses sinyali için sıkça kullanılan bir öznitelik çıkarımı tekniğidir. MFCC insanların duyma özelliği incelenirken ortaya çıkmıştır. Bir bakıma insan kulağının algılama şeklinin modellemesidir. Sesin kısa süreli güç spektrumunu temsil etmektedir. MFCC özniteliklerini çıkarmak için ses sinyallerine bir dizi işlem uygulanır. Bu işlemler Şekil 3.3'de verilmiştir. Sürekli olan ses sinyalini işleyebilmek için ilk olarak ses sinyali örneklenir. Ses sinyali sürekli değişir. Fakat ses sinyalini kısa zamanlı bir skalada incelediğimizde istatistiksel olarak uzun zamanlı bir skaladaki ses sinyalinden çok daha az değişmektedir. Bu sebeple 20-40 milisaniyelik çerçeveler ile ses sinyali pencereleme işlemi uygulanır. Pencereleme işlemi için Hamming pencereleme fonksiyonu yaygın olarak kullanılmaktadır. Eğer çerçeve boyutu çok kısa seçilirse yeterli sayıda örnek toplanamaz. Eğer çerçeve çok uzun seçilirse öznitelikler sesin uzun sürede çok fazla değişmesinden etkilenir. Bundan sonraki aşamada her çerçevenin güç spektrumu hesaplanır. Güç spektrumu, hızlı fourier dönüşümü yapılarak elde edilir. Elde edilen spektrumuna, insan duyma sistemine benzer bir Mel filtre bankası analizi uygulanır. Mel filtre bankası üçgensel filtrelerden oluşmaktadır. Bu üçgensel filtreler güç spektrumu ile çarpılır ve her bir filtrenin altındaki enerji hesaplanır. Elde edilen enerjileri logaritma işlemi uygulanır ve logaritma işleminin sonuçlarına ayrık kosinüs dönüşümü uygulanır. Ayrık kosinüs dönüşümü sonucunda MFCC öznitelikleri elde edilmiş olur.



Şekil 3.3 MFCC öznitelik çıkarım aşamaları

3.5 Temel Bileşen Analizi

TBA'nın temel fikri, çok sayıda birbiriyle ilişkili değişkenden oluşan bir veri kümesinin boyutunu indirgerken aynı zamanda veri kümesindeki varyasyonu olabildiğince korumaktır. Bunu yapmak için veri kümesini temel bileşenlerden oluşan yeni bir veri kümesine dönüştürür. Temel bileşenlerden oluşan veri kümesi ilintisiz ve düzenlidirler. İlk bir kaç temel bileşen tüm orijinal değişkenlerin varyasyonlarının çoğunu korur [58].

Temel Bileşen Analizinde amaçlanan $X=(x_1,x_2,x_3,\ldots,x_p)$ gibi *p* adet rastgele değişkenden oluşan X vektörünü, en az bilgi kaybı olacak şekilde, bu vektörü temsil edecek daha az sayıda değişkene indirmektir. X vektörü (C) kovaryans matrisi ve (e_i,λ_i) özvektör ve özdeğer çiftlerine sahip olsun. Öz değerler sıfırdan büyük olması koşulu ile sıralanır $\lambda_1>\lambda_2>\lambda_3\ldots>\lambda_p$. Bu sıralanışa göre doğrusal kombinasyonlar kurulur ve 2.4 de gösterilen y_i temel bileşenler oluşturulur [58]. Toplam varyansın en büyük kısmını açıklayan temel bileşen birinci temel bileşen , ikinci büyük kısmını açıklayan temel bileşene ikici temel bileşen denir [58].

$$y_1 = e_{11}x_1 + e_{21}x_2 + \cdots + e_{p1}x_p$$

$$y_2 = e_{12}x_1 + e_{22}x_2 + \cdots + e_{p1}x_p \quad (2.4)$$

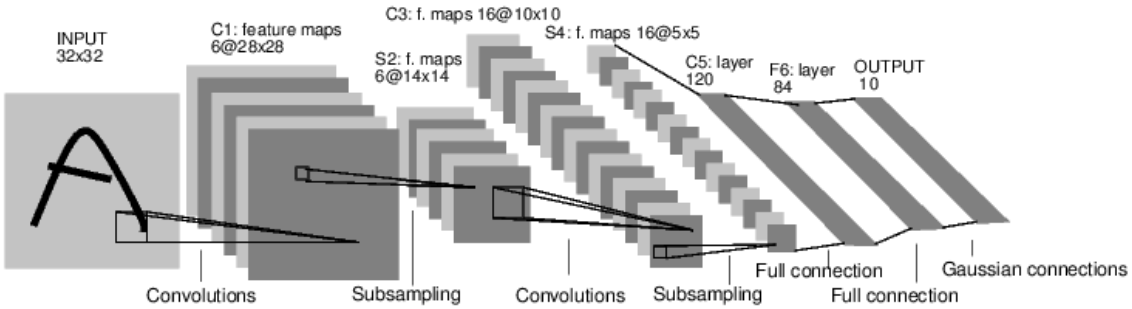
$$\vdots$$

$$y_p = e_{1p}x_1 + e_{2p}x_2 + \cdots + e_{pp}x_p$$

3.6 Evrişimsel Sinir Ağları (Convolutional Neural Networks)

Günümüzde, derin öğrenme tabanlı sistemler gösterdikleri başarılar ile popülerliğini artırmaktadır. ESA derin öğrenme başlığının altında yer alıp bir Yapay Sinir Ağı mimarisidir. ESA biyolojik olarak hayvanlar aleminin görsel korteksinden [59] esinlenerek yapılmıştır. ESA'nın genel yapısı LeCun [60] tarafından LeNet-5 isimli mimari ile sunulmuştur. LeNet mimarisi Şekil 3.4'de sunulmuştur. LeNet-5 mimarisi karakter tanıma uygulamaları için geliştirilmiştir.

ESA mimarileri evrişim katmanlarını takip eden alt örnekleme katmanları ve tercihe göre tam bağlı katmanlardan oluşur. Evrişimsel katmanlar öğrenebilen filtreler kümesidir. Her filtre evrişim işleminden sorumludur. Filtrenin boyutu gibi özelliklerine göre evrişim katmanı öznitelikler üretir. Örneğin evrişim katmanının girdisi *m* x *m* x *r* boyutlarında bir imge olsun. *m* imgenin genişliği ve boyu. *r* ise kanal sayısıdır. Örnek olarak kırmızı, yeşil, mavi kanallar için *r* = 3 tür. Evrişim katmanını *n* x *n* x *q* boyutlarında *k* adet filtresi olsun. *n* filtrenin genişliği ve boyudur. *q* ise kanal sayısıdır. *n* değeri *m* değerinden küçüktür. *q* ise *r* ye eşit veya küçük olabilir. Filtreler girdi olan imge ile evrişim işlemi uygulanır. Bu evrişim işlemi sonu *k* adet, $m-n+1$ boyutlarında öznitelik üretilir [61]. Her öznitelik ortalama veya en büyük değer gibi alt örnekleme işleminin uygulandığı birleştirme (pool) katmanı da denilen katmana eşlemlenir. Birleştirme katmanı genellikle boyut indirgeme, gereksiz bilgilerden kurtulma gibi işlevler görür. Aynı zamanda Makine Öğrenme tekniklerinde sıkça karşılaşılan bir problem olan aşırı eğitim (overfitting) problemine kontrol sağlar. Bir çok evrişim ve birleştirme katmanlarından sonra tam bağlı katmanlar yer alır. Tam bağlı katmanlarda bulunan nöronlar bir önceki



Şekil 3.4 LeNet-5 Evrişimsel Sinir Ağı mimarisi temsili gösterimi

katmanda bulunan bütün nöronlara bağlanır ve her bağlantının kendi ağırlığı vardır.

Evrişim işleminin Makine Öğrenme sistemlerine yardımcı olabilecek üç önemli işlevi vardır. Bunlardan birincisi seyrek etkileşimler (sparse interactions), ikincisi parametre paylaşımı (parameter sharing) ve eşdeğer gösterimlerdir (equivariant representations) [62]. Geleneksel Yapay Sinir Ağları matris çarpımı işlemini kullanır. Her girdi ünitesi ile her çıktı ünitesi arasındaki ilişkiyi temsil eden parametrelerden oluşan matrisleri çarparak bu işlemi yaparlar. Bunun anlamı her çıktı ünitesi her girdi ünitesi ile ilişkilidir. Bunun aksine ESA seyrek etkileşimler kullanır. Bu amaca evrişim işleminde kullanılacak filtreleri girdinin boyutlarından küçük seçilerek ulaşırlar. Örnek olarak bir imge işlerken, girdi imgesi binlerce veya milyonlarca piksele sahip olabilir, ancak sadece on veya yüzlerce piksel kaplayan filtreler ile kenarlar gibi küçük ve anlamlı özellikleri tespit edebiliriz. Bunun anlamı daha az parametreye ihtiyacı olmasıdır. Bu özellikler modelin bellek karmaşıklığını azaltır ve istatistiksel olarak etkinliğini artırır. Parametre paylaşımı ise parametrelerin modelde birden fazla fonksiyonda kullanılmasıdır. Geleneksel yapay sinir ağlarında her elementin ağırlık matrisi, çıktı katmanını hesaplarken sadece bir kez kullanılır. Girdi ile çarpılır ve bir daha ziyaret edilmez. Parametre paylaşımında ise ağı oluşturan ağırlıklar bir birlerine bağlıdır. Bunun anlamı ağırlığın değeri girdiye uygulandığında başka bir yerdeki ağırlık değerine bağlıdır.

Parametre paylaşımı evrişim operasyonu için kullanılır. Birbiri ile ilişkisiz farklı yerlerdeki parametreleri öğrenmek yerine bir küme içindeki parametreler öğrenilir. Evrişim operasyonu ve parametrelerin paylaşılması eşdeğer gösterimler sonucunu ortaya çıkarır. Eşdeğer fonksiyon demek eğer girdi değişirse çıktıda aynı şekilde değişir demektir. Spesifik olarak $f(x)$ fonksiyonu $g(x)$ fonksiyonu ile eşdeğer ise $f(g(x))=g(f(x))$ özelliğini gösterir [62].

Büyük veri kümeleri ve süper bilgisayarlar ile eğitilen ESA mimarileri öznitelik çıkarımı için kullanılabilmektedir. ESA mimarilerinden öznitelik çıkarımını anlamak için teknoloji transferi olarak da bilinen transfer öğrenmesini bilmek gerekmektedir. Makine Öğrenme yöntemleri belirli kabuller çerçevesinde iyi işler çıkarmaktadır. Bu kabullerden birisi eğitim verisi ve test verisinin aynı öznitelik uzayında ve aynı dağılımda olması gerekmektedir. Eğer dağılım değişirse istatistiksel olarak model yeni dağılıma göre eğitim kümesinden tekrar yapılandırmak gerekir. Gerçek hayattaki problemler için tekrardan eğitim veri kümesinin hazırlanması ve yeni model eğitilmesi pahalı ve zor bir işlemdir. Bazı problemlerde bu sebeplerden dolayı transfer öğrenmesi kullanılır [63]. Transfer öğrenmesinin makine öğrenimi üstündeki temel motivasyonu daha önceden öğrenilen bilgileri koruyup yeni gereksinimler için de ömür boyu kullanılmasıdır. Torrey ve Shavlik [64] transfer öğrenmesini, daha önce öğrenilmiş olan ilgili bir görevdeki bilgiyi kullanarak yeni bir sorunun öğrenmesindeki gelişme olarak tanımlamışlardır.

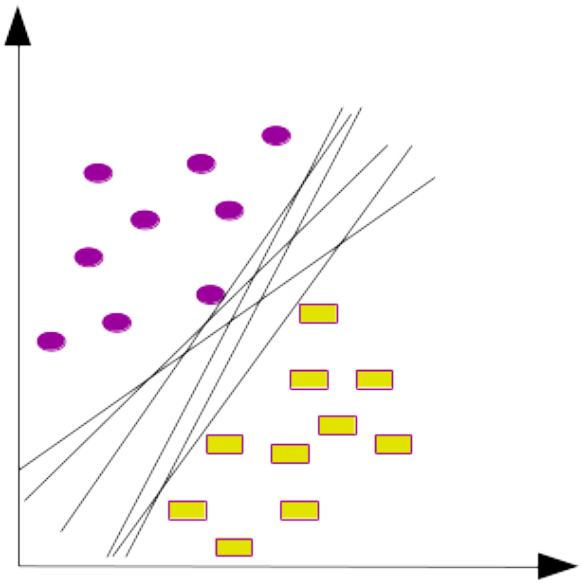
Bu çalışmada, Imagenet gibi büyük veri kümeleri ile süper bilgisayarlarda eğitilen AlexNet ve GoogLeNet ESA mimarileri transfer edilerek öznitelik çıkarımı için kullanılmıştır. Böylece belirli bir probleme yönelik hazırlanmış ESA mimarilerini kendi problemimizin çözümü için kullanılmıştır. Aynı zamanda problemimizin çözümü için çok kipli bir yaklaşım uyguladığımızdan ve ses özniteliklerinin görsel öznitelikler ile birleşimine ihtiyaç duyulmaktadır. Bu sebeple ESA mimarilerinin sınıflandırma için kullanılma stratejilerinden öznitelik çıkarma stratejisi kullanılmıştır.

3.7 Destek Vektör Makineleri

Destek Vektör Makineleri Makine Öğrenme disiplinin altında yer alan bir öğrenme algoritmasıdır. Makine Öğrenme algoritmalarını anlamak için öğrenme terimine hakim olmak gerekmektedir. Mitchell bir öğrenme problemini, bir bilgisayar programının bazı görevler için belirlenmiş sınıfları (*T*) ve performans kriteri (*P*) gibi değerler göz önünde bulundurularak, geçmiş tecrübelerinden (*E*) öğrenmesi ancak *T* görevinin *P* ile ölçülen performansının *E* tecrübesi ile artırması, olarak tanımlamaktadır [18]. Makine öğrenme algoritmaları genellikle belirli bir görev için geçmiş deneyimleri kullanarak bu görevin başarımını artırmaya çalışırlar.

Destek Vektör Makineleri Vapnik ve arkadaşları [65] [19] tarafından 1992 yılında tanıtılmıştır. Destek Vektör Makineleri bir sınıflandırma metodudur. Sınıflandırma işlemi için iki sınıfı birbirinden ayıran bir hiper düzlem oluşturarak yapar. Bu hiper düzlemi istatistiksel öğrenme teorisini kullanarak hesaplar.

Destek Vektör Makinelerindeki ana fikir doğrusal olarak birbirinden ayrılabilen veri kümesini ayıran en optimal hiper düzlemi bulmaktır. Doğrusal olarak ayrılmayan problemlerde ise veri kümesindeki örüntüleri yeni bir uzaya geçirerek bu uzayda bir hiper düzlem aramaktır. Şekil 3.5'de verildiği gibi doğrusal olarak birbirinden ayrılabilen iki sınıfın veri kümesinden oluşan bir düzlemde bu iki sınıfı birbirinden ayırmak için bir hiper düzlem yeterlidir. Bir kere bu iki sınıfı ayıran hiper düzlemin denklemi elde etiğimizde hiper düzleme göre verilerinin pozisyonlarından hangi sınıfa ait olduğu bilinebilir. Bu bilgi ile bu sınıfları birbirinden ayırt edilebilir. Fakat bu iki sınıfı bir birinden ayıran birden fazla hiper düzlem vardır. Destek Vektör Makineleri bu hiper düzlemlerin arasında en optimal hiper düzlemi verir. Bu durumda en optimal hiper düzlem, hiper düzleme en yakın noktaya uzaklığı en büyük olandır. Şekil 3.6'da görüldüğü gibi iki sınıfı bir birinden ayıran hiper düzeleme en yakın veri noktasına olan mesafesine margin denmektedir. Destek Vektör Makineleri margin mesafesinin en büyük olanı arar. Hiper düzleme en yakın veri noktalarına destek vektörleri denmektedir. Denklemi 3.5'de İki sınıfı birbirinden ayıran hiper düzlemin fonksiyonu verilmiştir. Bu denkleme göre *W* ağırlık



Şekil 3.5 Ayırıcı hiper düzlemler

vektörüdür. *b* ise sabit bir sayıyı ifade etmektedir. *X* sınıfı bilinmeyen bir noktayı ifade eder. Birbirinden doğrusal olarak ayrılabilen verilerin karar fonksiyonu verilmiştir (3.6). Bu fonksiyona göre $f(x) \geqslant 0$ için bir sınıfa $f(x) < 0$ için ise diğer sınıfa ait olacaktır.

$$w.x+b=0 \tag{3.5}$$

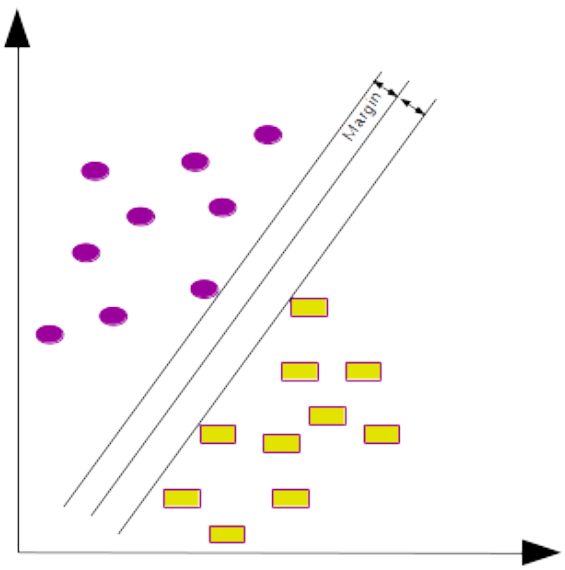
$$f(x)=w.x+b=\sum_{j=1}^{N} w_j.x_j+b \tag{3.6}$$

Destek Vektör Makineleri doğrusal olarak ayrılan veri kümeleri için en optimal hiper düzlemi bulabilirler. Fakat doğrusal olarak ayrılamayan veri kümeleri için çekirdek fonksiyonlar kullanırlar. Çekirdek fonksiyon ile güncellenmiş karar fonksiyonu 3.7'de verilmiştir.

$$f(x)=w.x+b=\sum_{j=1}^{N} w_j.x_j.K(x_i,x)+b \tag{3.7}$$

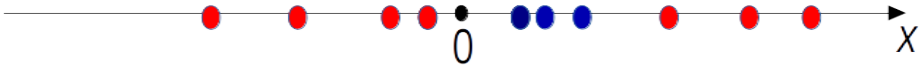
Çekirdek fonksiyonların görevleri doğrusal olarak ayrılamayan veri kümelerini bulunduğu uzaydan doğrusal olarak ayrılabilen daha büyük boyutlu uzaylara taşımaktır. Örnek olarak Şekil 3.7'de verilen bir boyutlu düzlemde birbirinden

doğrusal olarak ayrılamayan veri kümesi verilmektedir. Bu boyutta bu veri kümesinin dağılımından dolayı doğrusal olarak ayırabilen herhangi bir hiper düzlem bulunmamaktadır. Bu sebepten dolayı çekirdek fonksiyonları kullanarak Şekil 3.8'de görülen iki boyutlu düzleme veri kümesi taşınmaktadır. Şekilde görülen hiper düzlem ile bir boyutlu düzlemde bir birinden ayrılamayan veri kümesi iki boyutlu düzlemde bir birinden ayrılabilir. Destek Vektör Makineleri için radyal tabanlı fonksiyon, polinom ve doğrusal çekirdek fonksiyonları yaygın olarak kullanılır.



Şekil 3.6 Optimal hiper düzlem

Destek Vektör Makineleri iki sınıftan oluşan veri kümelerinde sınıflandırma yapabilirler. İkiden fazla sınıfı bulunan veri kümelerinde tek bir Destek Vektör Makineleri sınıflandırma yapamaz. İkiden fazla sınıfı bulunan veri kümelerini sınıflandırmak için Destek Vektör Makineleri kullanılabilmesi için bazı yöntemlere ihtiyaç vardır. Bu yöntemlerin temeli iki sınıftan fazla sınıfı olan veri kümeleri için bu sınıflar gruplayarak birden fazla Destek Vektör Makineleri eğitmektir. Böylece Destek Vektör Makineleri ikiden fazla sınıfı bulunan veri kümelerine uygulanabilir duruma gelir. Bu yöntemlerden biri Bire Karşı Bir yöntemidir. Bire Karşı Bir yönteminde s adet sınıftan oluşan bir veri kümesinde bu veri kümesi sınıflarının



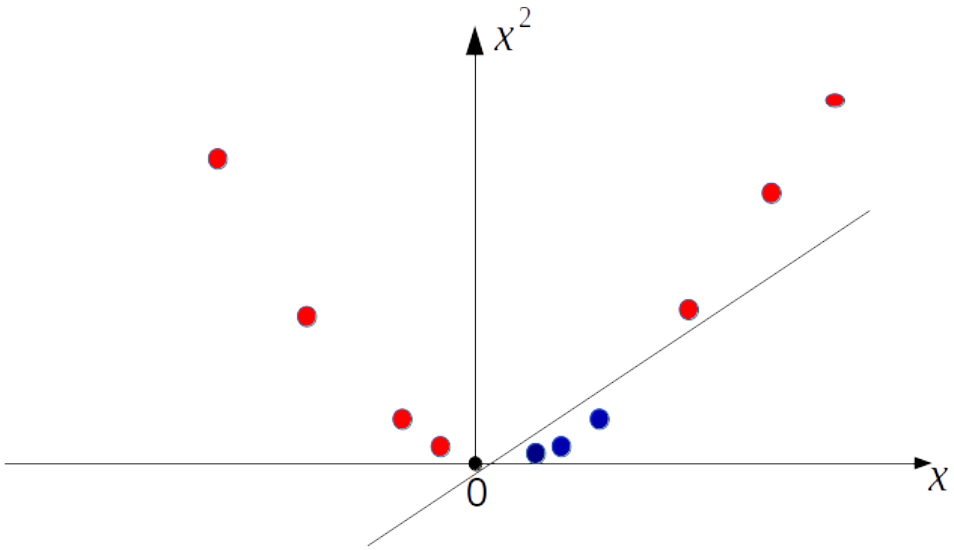
Şekil 3.7 Doğrusal olarak ayrılamayan veri kümesi

her ikili alt kümesi için bir Destek Vektör Makinesi eğitilir. s adet sınıf için d adet Destek Vektör Makinesi eğitilir (3.8).

$$d=\frac{s*(s-1)}{2} \tag{3.8}$$

Bir diğer yöntem ise bire karşı hepsi yöntemidir. Bire karşı hepsi yönteminde s adet sınıftan oluşan bir veri kümesi için bir sınıfa karşı diğer bütün sınıflar eğitilir ve her sınıf için bu tekrarlanır. s adet sınıfı bulunan bir veri kümesi için s adet Destek Vektör Makinesi eğitilir. Bire karşı hepsi yönteminde daha az Destek Vektör Makinesi eğitileceği için Bire Karşı Bir yöntemine göre daha az bir hesaplama ve bellek karmaşıklığı vardır.

Bu tez kapsamında video verisinin MFCC istatistiksel gösterimi ve ESA öznitelikleri füzyon işleminden sonra oluşan veri kümesi ile Destek Vektör Makineleri eğitilmiş ve performansı incelenmiştir. Destek Vektör Makinelerinin temelleri 1992'lerde atılmasında rağmen günümüzde bir çok araştırmada kullanılmakta ve başarılarını



Şekil 3.8 Çekirdek fonksiyon ile düzenlenmiş veri kümesi

devam ettirmektedir. Destek Vektör Makinelerinin sınıflandırma problemlerindeki en optimal hiper düzlem motivasyonu ve bir çok Makine Öğrenme araştırmasında kullanılması bu çalışma için tercih sebebi olmuştur.

3.8 Büyük Veri

Günümüzde büyük veri üzerinde çalışmaların yürütüldüğü, popülerliği artmış bir konudur. Büyük veri akademinin dışında endüstriyel ve teknolojik şirketlerinde araştırmalarını sürdürdüğü bir konudur. Bir çok farklı alan büyük veriye kendi bakış açısından yaklaştığı için herkesi tatmin edebilecek ortak bir tanımı yoktur. Shahrivari [66] çalışmasında büyük veri teriminin genel kullanımının, klasik çözümlerle örneğin ilişkisel veri tabanı sistemleri ile işlenemeyecek ve yönetilemeyecek büyüklükteki veri kümeleri olarak, dile getirmiştir. Jason Bloomberg ise büyük verinin tanımını, büyük ölçüdeki yapısal veya yapısal olmayan veri o kadar büyük ki geleneksel veri tabanları ve yazılım teknikleri ile işlem yapmak çok zor, şeklinde ifade etmiştir [67]. Fakat bu büyük veri tanımlamalarında büyük verinin sadece veri büyüklüğü veya hacmi (Volume) olan depolama alanlarında kapladığı alan karakteristiğine değinilmiştir. Büyük veri sahip olduğu karakteristiklerden çeşitlilik (Variety) ve hız (Velocity) gibi aşılması gereken zorluklar vardır [68]. Bu sebeple büyük veriyi tanımlarından büyük verinin diğer karakteristiklerinin de bulunduğu tanımlar vardır. Bu tanımlardan biri büyük veri, bilgiyi elde etmek, depolamak, dağıtmak, yönetmek ve analiz edebilmek için ileri düzey teknik ve teknolojiler gerektiren , geniş büyüklükte, yüksek hızda, karmaşık ve değişken veridir [69]. Büyük veri bir çok çalışmada ve önde gelen teknoloji şirketlerin raporlarında farklı tanımları mevcuttur [70] [71] [72] [73].

Büyük veriyi tanımlayan karakteristikler mevcuttur. Bu karakteristikler ilk olarak Laney'in çalışmalarına bakılarak 3V olarak adlandırılan büyüklük, çeşitlilik ve hız olarak kabul görmüştür [74]. Daha sonraki çalışmalarda büyük verinin karakteristiklerine iki kabul görmüş karakteristik eklenmiştir. Bu karakteristikler değer (Value) ve doğruluk (Veracity) olarak adlandırılmaktadır [75]. Büyüklük veya hacim depolama alanlarında kapladığı alan olan Terabyte, Zetabyte, Petabyte,

seviyesinde büyüklüğünü ifade eder. Örneğin Lofar teleskobu saatte 5 Petabyte'lık veri üretir. Bu veri üzerinde bir doğrulama işlemi yürütülür ve sadece doğrulanan veri depolanır [75]. Büyük verinin bir diğer karakteristiği olan hız, verinin büyük ölçekte, sürekli ve yüksek hızda üretilmesidir. Bu karakteristiği sergileyen veri bir dizi algılayıcıdan veya bir çok olaydan toplanırken gerçek zamanlı, gerçek zamana yakın veya yığın halinde veya akan veri olarak işlenmeye ihtiyaç olabilir [75]. Çeşitlilik ise verinin karmaşıklığı ile alakalıdır. Gelen verilerin farklı format, farklı kaynak, farklı yapıda olması ile ilişkilidir. Bu karakteristiği olan verilerin işlenmesi ve saklanması büyük verinin karakteristiğini oluşturan başka bir etkendir. Büyük veri 5V'sinde bulunan değer ise, toplanan veri üzerinde uygulanacak işlemler ve analizlere bağlı olup, bu işlemlerden sonra artı değer katması ile ilgilidir. Doğruluk ise verinin istatistiksel olarak güvenilirliği olarak açıklanabilecek verinin istikrarı ve verinin güvenilir bir alt yapıda olması, verinin kökeni, işleme metotları gibi bir dizi etkenden oluşmaktadır ve veri güvenilirliği ile alakalıdır. Doğruluk verinin güvenilir, doğrulanmış ve izinsiz erişimlere, değiştirmelere karşı korumalı olmasını garantilemektir [75].

Çokluortam verileri için büyük veri ise, sosyal medyanın yaygınlaşması ve kullanıcıların sürekli içerik üretmesi ile beraber önemini artırmıştır. Örneğin Youtube'a her dakikada ~100 saatlik video yüklenmektedir. Facebook'a ise günlük olarak ~350 milyon fotoğraf yüklenmektedir [6]. Çokluortam büyük verisinin analiz edilmesi, depolanması ve yönetilmesi geleneksel ve sıklıkla kullanılan bilgisayar sistemleri ve yazılımlar ile mümkün değildir. Çokluortam verilerinde de olduğu gibi büyük veri sahip olduğu karakteristiklerden dolayı büyük veri ile çalışmak için klasik yöntemlerden farklı dağıtık sistemlere ve bu sistemler üzerinde işlem yapabilecek yazılımlara ihtiyaç vardır. Dağıtık sistemler birden çok bilgisayarın tek bir bilgisayar gibi ölçeklendirilmesi ve kullanılmasını ifade eder. Büyük veri ile çalışmak için kullanılabilecek önde gelen iki teknoloji Apache Hadoop [51] ve Apache Spark'dır [76]. Büyük veri dünyasında dağıtık sistemler için kullanılabilecek Apache Spark'a [76] ait olan RDD ve MapReduce [77] örnek olabilecek programlama modelleridir.

Bu çalışmada video içerik analizi için önerilen yöntem büyük veri teknolojileri kullanılarak gerçekleştirilecektir. Büyük veri teknolojisinin önerilen yöntemin işleme katmanlarına olan etkisi incelenecektir.

3.9 Apache Spark

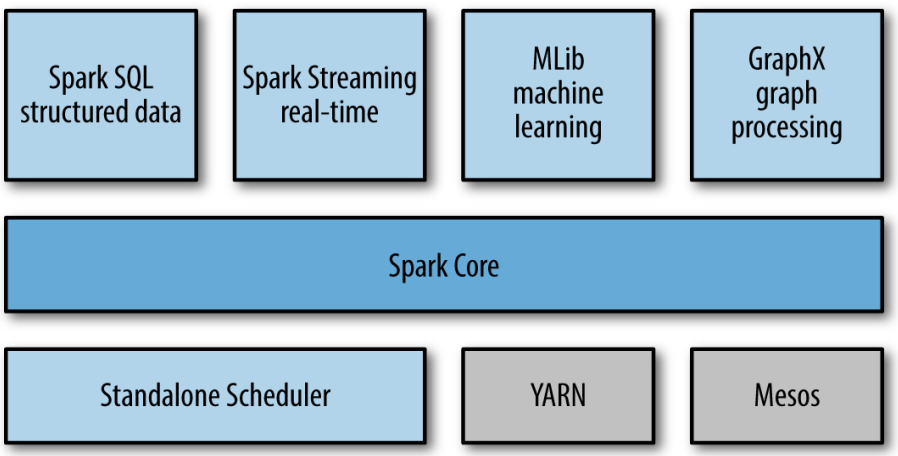
Apache Spark, 2009 yılında Berkley Üniversitesi AMPLab laboratuvarında Matei Zaharia tarafından geliştirilen dağıtık sistemlerde veri analizi gerçekleştiren teknolojidir. İnternet sitesinde yer alan tanıma göre Apache Spark, hızlı ve genel amaçlı küme hesaplama sistemidir [23].

Apache Spark Scala, Python ve Java programlama dilleri ile kullanabilmekte ve bu özelliği ile bilişim dünyasında hitap etiği kişi sayısı artırmaktadır. Apache Spark teknolojisinin popülerliğini artıran bir diğer etken ise, verinin bellek içi olarak kullanıldığı durumlarda Apache Hadoop'tan 100 kat daha hızlı ve verinin disk üzerinden kullanıldığı durumlarda ise 10 kat daha hızlı olduğudur [23]. Apache Spark veriyi bellek içinde sakladığından, çok sayıda iterasyon içeren algoritmalar için okuma yazma işlemi yapılmayacağından, Apache Hadoop'a göre daha avantajlıdır.

Apache Spark bir dizi bileşenden oluşur. Bu bileşenler Şekil 3.9'da verilmiştir. SparkCore bellek yönetimi, depolama sistemleri ile etkileşim gibi temel fonksiyonellikleri barındırır. Aynı zamanda Apache Spark'ın temel programlama soyutlaması olan Resilient Distributed Datasets (RDD) yapısını barındırır. RDD bir küme makinede bölünmüş, bir parçası kayıp olduğunda tekrar yapılandırılabilen ve sadece okuma yapılabilen nesne koleksiyonudur [76]. RDD'ler yerel dosyalar gibi herhangi dış bir dosya sisteminde bulunan verilerden veya kullanıcı tarafından Spark programında yaratılabilir. Yaratılan RDD'ler dağıtık sistemde yer alan bilgisayarlarda konumlandırılan çalıştırıcı düğümler ile işlenir. RDD'lere yaratıldıktan sonra Transformations ve Actions olmak üzere iki işlem uygulanabilir. Transformations işlemleri RDD'ye uygulandıktan sonra başka bir RDD geri döndürür. Actions işlemlerinde ise Transformation sonrası yaratılan RDD'ler

toplanarak sonuç geri döndürürler. Sonuçlar ana düğüme (driver program) geri döndürülebileceği gibi her hangi bir dış depolama sistemine de yazılabilir.

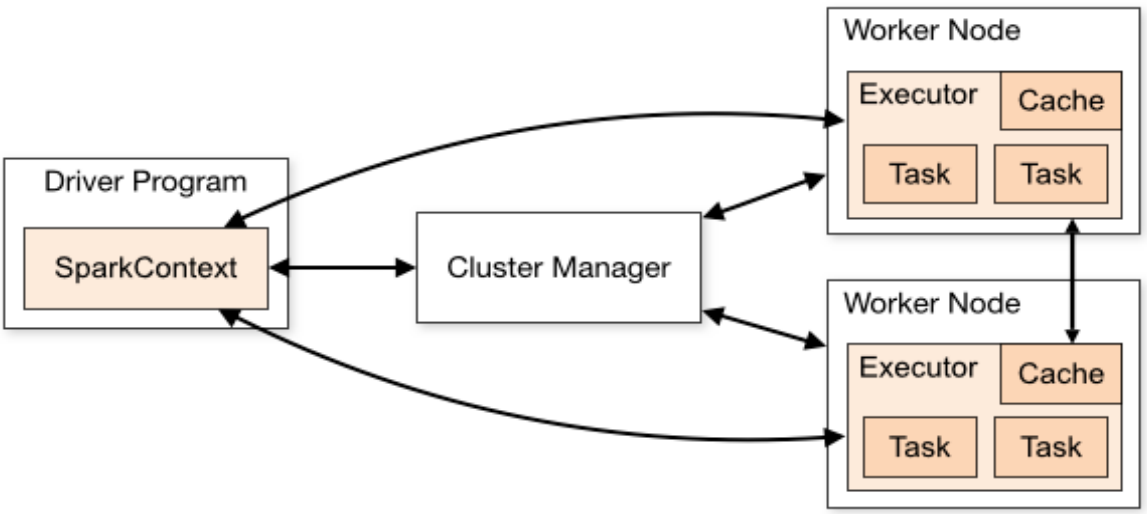
Apache Spark dağıtık sistem kaynak yöneticisine ihtiyaç duyar. Spark kendi kaynak yöneticisine Standalone Scheduler ismi verilir. Aynı zamanda YARN [78] ve Mesos [79] gibi kaynak yöneticileri ile birlikte çalışabilmektedir. Spark SQL yapısal veriler için, Spark Streamin akan veriler için, Mlib Makine Öğrenme uygulamaları için, GraphX ise Graph veri yapıları için özelleşmiş Spark bileşenleridir. Apache Spark yazılan kodların test edebilmek, denemeler yapabilmek için herhangi küme



Şekil 3.9 Apache Spark bileşenleri [97]

bilgisayara ihtiyaç duymadan yerel kipde çalışabilir. Fakat Spark genel kullanım amacı küme kipinde çalışması içindir. Küme modunda çalışma prensibi Şekil 3.10'da verilmiştir. Spark uygulamaları küme üzerinde bağımsız bir dizi işlem olarak çalışır. Bu işlemler ana programda (driver program) bulunan, SparkContext adı verilen nesne ile kontrol edilir. Küme kipinde kullanılırken, SparkContext nesnesi kaynakları uygulamaya göre paylaştıran kaynak yöneticisine bağlanır. Bu bağlanma gerçekleştikten sonra Spark küme üzerinde çalıştırıcıları (executor) elde eder. Bu çalıştırıcılar uygulamada bulunan hesaplamaları ve veri depolamaları işlemeden sorumludurlar. Sonraki işlem ise uygulamada bulunan kodlar çalıştırıcılara yollanarak uygulamanın gerçekleşmesi sağlanır.

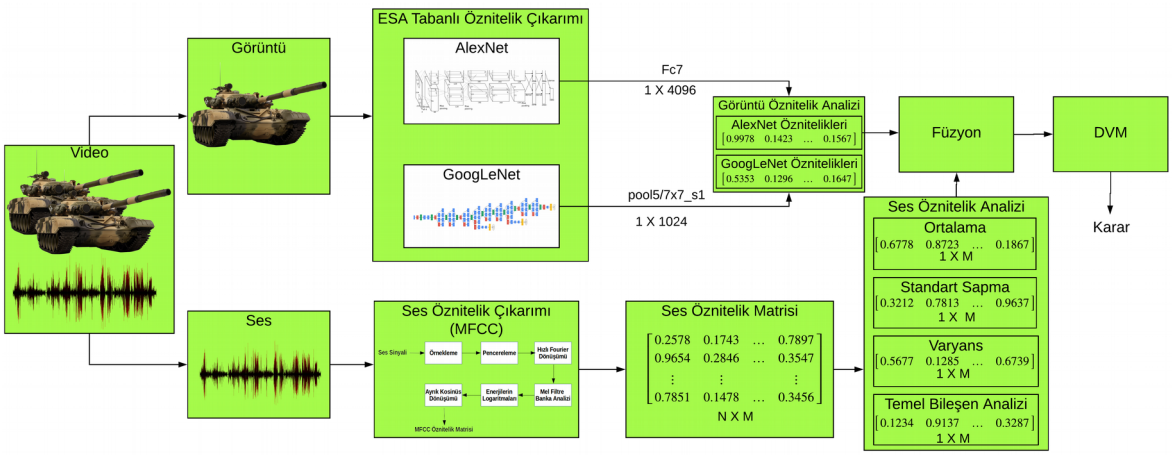
Apache Spark'ın diğer teknolojilere göre daha hızlı olması, popüler olarak kullanılması ve başarı sonuçlar elde etmesi nedeniyle bu çalışmada tercih edilmiştir.



Şekil 3.10 Spark küme kipi genel bakış [23]

4 ÇOK KİPLİ VİDEO KAVRAM SINIFLANDIRMASI

Bu tezde, video kavram sınıflandırma problemi için çok kipli bir yöntem önerilmektedir (Şekil 4.1). Önerilen yöntem dört ana aşamadan oluşmaktadır: Görsel-işitsel kiplerin ayrıştırılması, öznitelik çıkarımı, veri füzyonu ve sınıflandırma. Video verisinin ses ve görüntü kiplerinin tamamlayıcı etkileri baz



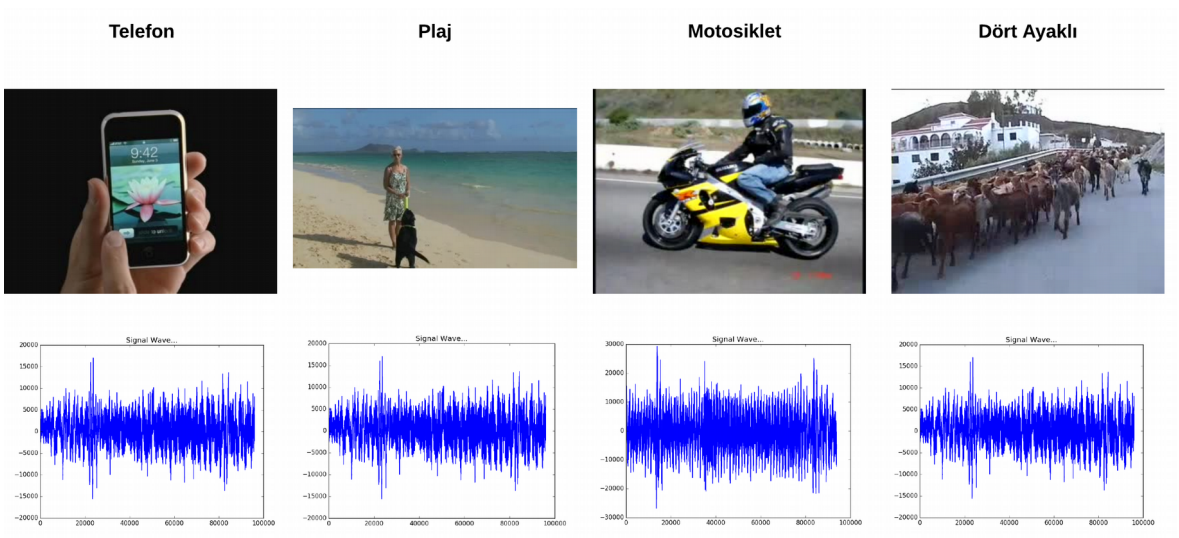
Şekil 4.1 Önerilen çok kipli sınıflandırma yöntemi

alınarak bu çalışmada, çok kipli bir yöntem tercih edilmiştir. Son yıllarda, ESA mimarilerinin bilgisayarla görü alanındaki başarımlarından dolayı ESA özniteliklerinin kullanılmasına karar verilmiştir. Ses kipi için, MFCC öznitelik çıkarımının genel ses tanıma ve konuşma tanıma uygulamalarında gösterdiği performans ve video içerik analizi için sıkça kullanılmasından dolayı tercih edilmiştir. Bu çalışmaya özgü olarak, ses özniteliklerinin farklı istatistiksel temsilleri de görsel özniteliklerle birlikte kullanılmıştır. Böylece, zamansal ve büyük boyutlu ses özniteliklerinin görsel öznitelikleri baskılamasının önüne geçilmiş ve aynı zamanda ses öznitelik boyutu indirgendiğinden bellek ve hesaplama karmaşıklığının azaltılması amaçlanmıştır. Çok kipli bir yaklaşım önerildiğinden veri birleştirme işlemi uygulanmıştır. Füzyon işlemi için sınıflandırıcıların kararlarını etkilememek, ham verinin korunması ve ham veride ilgili kavramı temsil eden özniteliğin deformasyonuna engel olmak amacı ile öznitelik düzeyinde füzyon tercih edilmiştir. Sınıflandırıcı olarak ise, iki boyutlu düzlemde en optimal hiper

düzlemi bulma motivasyonu ve bilgisayarla görü problemlerindeki başarımlarından [20] [21] [34] [80] [81] [82] dolayı DVM tercih edilmiştir.

4.1 İşitsel ve Görsel Kiplerin Ayrılması

Uygulanan yöntemde ilk aşamada video verisi görsel ve işitsel kiplerine ayrılır. Bu işlem için görsel kipten belirlenen kavramı temsil eden anahtar çerçeve çıkarılır. Belirlenen kavram veri kümesi içerisinde bir çok videoda ve videoyu oluşturan faklı çerçevelerde bulunabilir. Anahtar çerçeveler ve bu çerçevenin video zamanına karşılık gelen ses sinyalinin temsili Şekil 4.2'de verilmiştir. İşitsel kipi için anahtar çerçevenin video zamanında belirdiği an göz önünde bulundurularak, bu ana denk gelecek şekilde 1 saniye süresinde ses verisi elde edilir. Anahtar çerçevenin belirdiği an, belirlenen kavramı temsil eden en iyi ses verisini seçmek amacı ile ses verisinin 1 saniyelik süresinin tam orta noktasına denk gelecek şekilde ayarlanmıştır. Bu işlem sonrasında belirlenen kavramı temsil eden anahtar çerçeve ve bu çerçeveye denk gelecek 1 saniye süresindeki ses verisi elde edilerek, video verisi işitsel ve görsel kiplerine ayrılmış bulunmaktadır. Elde edilen görüntü çerçeveleri jpeg formatında kaydedilmektedir. Elde edilen ses verisi wav formatında kaydedilmektedir. İşitsel ve görsel kiplere ayrılma işlemini işitsel ve görsel öznitelik çıkarımı işlemi takip etmektedir.



Şekil 4.2 Anahtar çerçeve ve bir saniyelik ses sinyali temsili

4.2 Öznitelik Çıkarımı

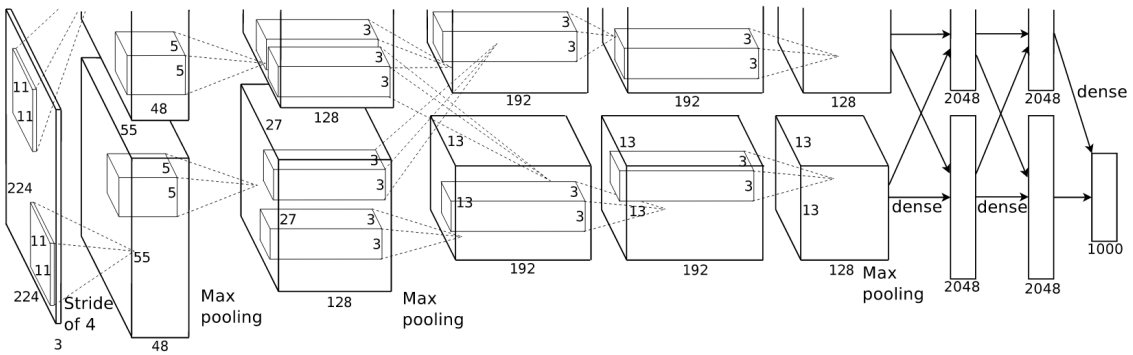
4.2.1 Görsel öznitelik çıkarımı

Uygulanan yöntemde video verisinin görsel kipinden elde edilen, görüntü çerçeveleri için ESA mimarileri kullanılarak öznitelik çıkarımı yapılmıştır. Birbirleri ile karşılaştırmak amacı ile biri diğerine göre daha karmaşık iki adet ESA mimarisi tercih edilmiştir. ESA mimarileri için gösterdikleri başarılı performansı ve popüler olarak çeşitli uygulamalarda kullanıldığı için AlexNet ve GoogLeNet tercih edilmiştir. GoogLenet AlexNet'e göre katmanlarının fazla olması gibi nedenlerden daha karmaşık bir ESA mimarisidir. ESA mimarilerin öznitelik çıkarımı bir çok katmandan yapılabilmektedir. Yapılan çalışmalarda [40] [83] [80] video kavram sınıflandırma, obje tanıma vb. problemler için son tam bağlı katmandan elde edilen öznitelikler daha başarılı sonuç vermektedir. Görsel öznitelik çıkarımı için video verisinin görsel kipinden elde edilen anahtar çerçeveler ESA mimarilerine verilmekte ve ILSVRC için hazırlanmış olan softmax katmanından önce gelen son tam bağlı katmanındaki veriler çekilerek öznitelik çıkarımı işlemi yapılmaktadır. Bu katmanlardan çıkarılan öznitelik ESA mimarilerinin özelliklerine göre AlexNet için 1x4096 GoogLeNet için 1x1024 k boyutunda $\vec{G}$ görsel öznitelik vektörüdür. Her iki mimariden elde edilen görsel öznitelik ile DVM eğitilerek ve değerlendirilerek en iyi performansı sergileyen DVM seçilerek görsel öznitelik analizi yapılmaktadır. En iyisi belirlenen görsel öznitelik ses öznitelik temsilleri ile birlikte kullanılması amacı ile füzyon işlemine yönlendirilir.

AlexNet nesne sınıflandırma görevi için Krizhevsky [14] ve arkadaşları tarafından hazırlanan bir Evrişimsel Sinir Ağıdır. AlexNet 1.2 milyon imge ve 1000 farklı kategoriden oluşan ILSVRC-2010 veri kümesi ile eğitilmiştir [14]. Katıldığı yarışmada kendinden daha önceki yöntemlerin başarısına göre büyük bir sıçrayış yaparak Evrişimsel Sinir Ağlarının popülerliğini artırmıştır. AlexNet'in genel mimarisi Şekil 4.3'de verilmiştir. AlexNet 60 milyon parametresi bulunan 650000 nörondan oluşan ve 5 evrişimsel katmanı, bu katmanları bazılarını takip eden en büyük birleştirme (max-pooling) katmanı ve 3 adet tam bağlı katmanı bulunan

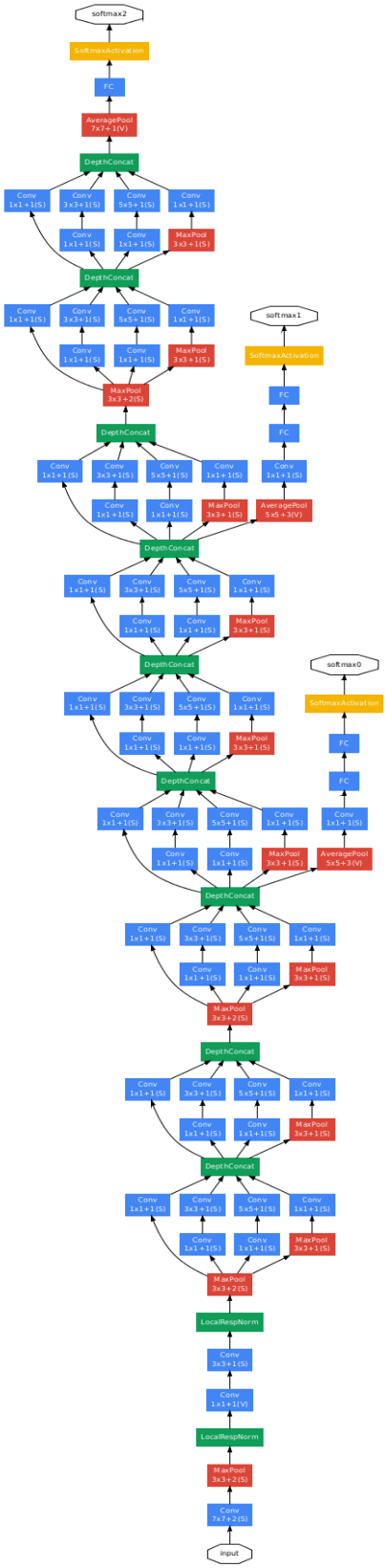
toplamda 8 katmanlı bir Evrişimsel Sinir Ağı mimarisidir [14].

AlexNet nesne tanıma sistemleri için büyük bir sıçrayış göstererek ESA popülerliğini artırmış ve bizim çalışmamızda kullanmamız için motivasyonumuzu oluşturmuştur. Aynı zamanda AlexNet kendisinden sonra sunulmuş ESA mimarilerine göre daha anlaşılabilir ve sade bir yapısı vardır. AlexNet video içerik analizi için ESA ile çalışacaklar için iyi bir başlangıç noktasıdır.



Şekil 4.3 AlexNet mimarisi [14]

Bu tez kapsamında kullanılan diğer bir Evrişimsel Sinir Ağı mimarisi Szegedy ve arkadaşları tarafından hazırlanan GoogLeNet'tir [15]. GoogLeNet, AlexNet [14] gibi ILSVRC-2014 yarışmasına katılıp, en iyi performansı sergileyerek kendini tanıtmıştır. GoogLeNet mimarisinin genel yapısı Şekil 4.4'de verilmiştir. ESA mimarilerinin performansını geliştirmek için iki direkt yöntem vardır. Bu yöntemler ESA mimarilerinin derinliğini ve genişliğini artırarak olur. ESA'nın derinliğini artırmak demek bir biri ile anlamlı daha çok katman eklemektir. Genişliğini artırmak ise katmanlarda bulunan elementlerin sayısını artırmak demektir [84]. Her iki yönteminde sisteme katıkları, hesaplama karmaşıklığı ve aşırı eğitim gibi problemleri vardır. GoogLeNet mimarisi AlexNet mimarisine göre daha çok katmanı bulunan daha karmaşık bir Evrişimsel Sinir Ağı mimarisidir. GoogLeNetin daha fazla katmanı bulunmasına rağmen AlexNet'e göre 12 kat daha az parametresi bulunmaktadır. GoogLeNet parametreleri bulunan katmanlar sayılırsa 22, birleştirme (pooling) katmanlarıyla beraber 27 katmandan oluşmaktadır [15].



Şekil 4.4 GoogLeNet mimarisi [15]

GoogLeNet ESA mimarisi AlexNet'e göre daha fazla katmanı bulunmasından ve daha karmaşık bir yapısı olmasından dolayı AlexNet ile karşılaştırmak için seçilmiştir. GoogLeNet'i seçmemizin diğer bir sebebi ise ILSVRC-2014 yarışmasında gösterdiği üstün başarıdır.

4.2.2 İşitsel öznitelik çıkarımı

Video verisinin işitsel kipinden elde edilen t_a = 1 saniyelik ses verisi için MFCC öznitelik çıkarımı işlemi uygulanmıştır. MFCC öznitelikleri t_w =10ms'lik çerçeveler için t_h = 5ms'lik kaymalar ile hesaplanmıştır. Her analiz çerçevesinden m = 20 MFCC katsayısı çıkarılmıştır. MFCC öznitelik çıkarımı işlemi sonrası oluşan öznitelik matrisi n değeri 4.1 olmak koşulu ile $n \times m$ boyutunda bir A öznitelik matrisi elde edilir (4.2).

$$n = \frac{t_a}{t_h} \tag{4.1}$$

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{pmatrix} \tag{4.2}$$

Elde edilen MFCC öznitelik matrisi A boyutsallık lanetinden kurtulmak ve hesaplama ve bellek karmaşıklığı azaltmak amacı ile matris sütunlarına sırasıyla ortalama (4.3), varyans (4.4), standart sapma (4.5) işlemleri uygulanır. Tanımlanan operatörler A matrisinin sütunlarına uygulandığında ortalama, varyans ve standart sapma için sırasıyla 1xm boyutlarında V , $\grave{V}$, $\dot{V}$ vektörleri elde edilir(4.6) (4.7) (4.8).

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{4.3}$$

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{4.4}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{1}^{N} (x_i - \bar{x})^2} \tag{4.5}$$

$$\bar{V} = \langle v_1, v_2, \ldots, v_m \rangle \tag{4.6}$$

$$\grave{V} = \langle \grave{v}_1, \grave{v}_2, \ldots, \grave{v}_m \rangle \tag{4.7}$$

$$\dot{V} = \langle \dot{v}_1, \dot{v}_2, \ldots, \dot{v}_m \rangle \tag{4.8}$$

MFCC A öznitelik matrisinin sütunlarının bulundurduğu bilgiyi en iyi şekilde koruyarak boyut indirgemek amacı ile TBA işlemi uygulanmaktadır. TBA işleminden en iyi performansı almak için MFCC matrisine zero-mean ve unit-variance normalizasyon işlemi uygulanmıştır [85]. TBA işlemi için toplam varyansın en büyük kısmını açıklayan temel bileşen olan birinci temel bileşen korunarak boyut indirgenmektedir. TBA işlemi sonrası A öznitelik matrisinden $\check{V}$ vektörü elde edilir (4.9). Daha sonra elde edilen her istatistiksel gösterim için bir DVM eğitilip test edilerek en iyileme çalışması sonucu ses özniteliği analiz edilir. Görsel öznitelikle birlikte kullanılması için füzyon işlemine yönlendirilir.

$$\check{V} = \langle \check{v}_1, \check{v}_2, \ldots, \check{v}_m \rangle \tag{4.9}$$

4.3 Veri Füzyonu

Veri füzyonu, Veri Füzyonu Modeli ismi ile 1985'de JDL(Joint Directors of Laboratories) organizasyonunun Veri Füzyonu grubu tarafından geliştirilmiştir. Akademik alanda bir çok araştırmacıya göre veri füzyonu JDL tarafından sunulan tanımı kabul görmektedir. Bu tanıma göre veri füzyonu, çoklu seviyeli işlemlerde bir veya birden fazla kaynaktan gelen veri ve bilginin ilişkisi, ilintisi ve kombinasyonunun yeni bir pozisyona ulaşmak için tahminlerin saptanması, eksiksiz ve zamanında değerlendirilmesi olarak tanımlanabilir [86]. Veri füzyonu ile ilgili daha çok bilinen diğer bir tanım ise Hall ve Llinas [87] tarafından belirtilmiştir. Halll ve Llinasın tanımına göre veri füzyonu, tek bir algılayıcı kullanılarak elde edilen doğruluk veya spesifik çıkarımları geliştirmek amacı ile bir çok algılayıcıdan veya ilişkili veri tabanlarındaki ilgili verileri kombine etmektir.

Sınıflandırma tabanlı sistemler için farklı kaynaklardan gelen veriler arasındaki ilişkiler belirli özellikler göstermektedir. Bu özellikler Durrant ve Whyte [88] üç

başlık altında toplamıştır. Bu çalışmaya göre faklı kaynaklardan gelen veriler birbirleri ile tamamlayıcı özelliğe sahip olabilirler. Tamamlayıcı özellik, farklı kaynaklardan gelen verilerde bir kaynaktan gelen verinin diğer kaynaktan gelen verinin eksiğini kapatmaya yönelik bir bilgi barındırmasıdır. Farklı bir özellik ise farklı kaynaklardan gelen veriler bir biri ile ihtiyaç fazlası bir eğilim göstermeleridir. Farklı kaynaklardan gelen veriler hedef hakkında aynı bilgiyi barındırıyorsa bu kaynaklardaki verilerden biri ihtiyaç fazlası olduğu söylenebilir. Diğer bir özellik ise farklı kaynaklardan gelen verileri bir biri ile işbirlikçi özellik sergilemeleridir. İşbirlikçi özellik gösteren farklı kaynaklardan gelen veriler birbirleri ile füzyon işlemine uğradıklarında oluşan yeni veri eski halinden daha karmaşık bir hal alıyorsa bu veriler bir birleri ile iş birlikçi bir özellik sergiliyor denebilir.

Çokluortam verileri için sınıflandırma sistemlerinde genel olarak işlenmemiş veriden elde edilen öznitelik ile sınıflandırıcı eğitilir ve daha sonra görülmemiş veriler için bu sistemin kararı sorgulanır. Sınıflandırma sistemleri için füzyon işlemi ise bu aşamaların herhangi birinde uygulanabilir. Şekil 4.5'de görüldüğü gibi Dasarathy [17] çalışmasında sınıflandırma sistemleri için füzyon işlemi hangi aşamalarda olabileceğini kategorilendirmiştir. Bu çalışmaya göre sınıflandırma sistemleri için füzyon işlemine veri girer veri çıkar, veri girer öznitelik çıkar, öznitelik girer öznitelik çıkar, öznitelik girer karar çıkar, karar girer karar çıkar. Füzyon işlemine veri girer veri çıkar aşamasında algılayıcılardan alınan işlenmemiş veri doğrudan füzyon işlemine tabii tutulur. Füzyon işlemi sonucu ise yine işlenmemiş veridir. İkinci aşama olan veri girer öznitelik çıkar aşamasında ise kaynaklardan gelen işlenmemiş veri füzyon işlemi uygulanır. Füzyon işlemi sonucu kategoriyi karakterize eden öznitelik elde edilir. Öznitelik girer öznitelik çıkar aşamasında ise kaynaklardan gelen işlenmemiş veriler üzerinde öznitelik çıkarımı yöntemleri uygulanır ve bu yöntemler sonucu elde edilen birbirinden farklı özniteliklere füzyon işlemi uygulanır. Füzyon işleminin sonucu belirlenen kategoriyi karakterize eden bir öznitelikdir. Öznitelik girer karar çıkar aşamasında ise farklı kaynaktan toplanan veriye ait öznitelikler füzyon işlemine sokulur bu füzyon işlemi sonucu sınıflandırma sistemlerinin kategori hakkındaki kararıdır. Karar girer karar çıkar

Şekil 4.5 Füzyon işlemini girdi ve çıktısına göre kategorilendirilmesi [17]

aşamasında ise farklı kaynaklardan gelen veri hakında sınıflandırma sistemleri karar verir. Bu karar füzyon işlemini uygulanarak yeni bir karar çıktısı elde edilir. Bu aşamadaki füzyon işlemine karar füzyonu denmektedir.

Bu çalışmada MFCC ve ESA özniteliklerinin kendi alanlarında gösterdikleri başarımlarından dolayı öznitelik girer öznitelik çıkar aşamasında füzyon işlemi uygulanmaktadır. Basit ve etkili bir yöntem olması nedeniyle, füzyon yöntemi olarak uç uca ekleme yöntemi uygulanmıştır. V ve G sırasıyla işitsel ve görsel kiplerden elde edilen m ve k elemanlı öznitelik vektörleri olmak üzere, füzyon işleminin sonucunda $m + k$ boyutunda görsel ve işitsel öznitelikleri içeren $\vec{F}$ öznitelik vektörü elde edilir (4.10). Füzyon işlemi sonucu veri artırımı

sağlanmaktadır. Böylece ses ve görsel verilerin birbirleri ile olan tamamlayıcı ilişkilerinin sınıflandırma sistemlerine etkileri incelenmiştir.

$$\vec{V}=\langle v_1, v_2, \ldots, v_m \rangle, \vec{G}=\langle g_1, g_2, \ldots, g_k \rangle \Rightarrow \overrightarrow{F_{m+k}}=\langle V \| G \rangle \tag{4.10}$$

4.4 Sınıflandırıcı Tasarımı

Elde edilen $\vec{F}$ vektörü DVM eğitiminde kullanılır. DVM çekirdeği olarak yapılan çalışmalar ve deneyler sonucunda radyal tabanlı çekirdek fonksiyonu seçilmiştir. Çok sınıflı sınıflandırmalar için Bire Karşı Bir stratejisi kullanılmaktadır. Bire Karşı Bir stratejisi için sınıf sayısı s olması koşulu ile d adet DVM eğitilmektedir (4.11). DVM için gama ve C parametreleri ızgara arama algoritması ile aranmaktadır. Eğitilen bütün DVM için aynı stratejiler kullanılmıştır.

$$d=\frac{s*(s-1)}{2} \tag{4.11}$$

5 DENEYSEL ÇALIŞMA VE SONUÇLAR

Önerilen yöntemi değerlendirmek amacıyla belirlenen kavramları içeren veri kümeleri oluşturulmuştur. Önerilen yöntem 3 farklı senaryo için test edilmiştir. Bu senaryolar video kavram sınıflandırması, taşıt türü sınıflandırması ve Apache Spark çalışması olarak adlandırılmaktadır. Bu veri kümeleri üzerinde önerilen yöntem ve önerilen yöntemi oluşturan bileşenlerin performans kriterine olan etkileri analiz edilmiştir.

5.1 Veri Kümesi

5.1.1 Video kavram sınıflandırması için veri kümesi

Video veri kümesi olarak TRECVID (TREC Video Retrieval Evaluation) organizasyonun 2012 yılında anlamsal indeksleme (SIN) uygulaması için hazırlanmış olan IACC.1.A-C (Internet Archive Creative Commons) veri kümesi kullanılmıştır [89]. TRECVID organizasyonun hazırladığı görevler için özelleşmiş veri kümeleri uzun yıllardır araştırmacıların önerdikleri yöntemleri değerlendirmeleri için kullandıkları bir veri kümesidir. IACC.1.A-C veri kümesinde 500 kategoriden oluşan 8000 video bulunmaktadır. Videoların toplam süresi 600 saattir. Videolar toplam 160 GB boyutundadır. Veri kümesinde bulunan videolar MPEG-4/H.264 formatında en az 10 saniye en fazla 3.5 dakika süresindedir. Bu veri kümesi aktif öğrenme [90] tekniği kullanılarak etiketlenmiştir. IACC.1.A-C veri kümesinden 2012 yılındaki SIN görevi için belirlenmiş 30 kategoriden oluşan alt küme hazırlanmıştır. Kategori isimleri Çizelge 5.1'de verilmiştir. Hazırlanan veri kümesi 5103 videodan oluşmaktadır. 5103 video içerisinde etiketlenen 30740 adet belirlenen kategorilerin bulunduğu çerçeve vardır. Önerilen yöntemin ve bileşenlerinin test aşamasındaki doğruluk performanslarının analizinde adil bir yargılama yapılabilmesi için ve tüm örneklerin test edilebilmesi için K=3 çapraz doğrulama tekniği kullanılmıştır. Çapraz doğrulama tekniğinin ilk tekrarı için kategorilere göre örnek sayısı Çizelge 5.1'de verilmiştir. İlk tekrar için toplam 20505 adet örnek eğitim için 10235 adet örnek ise test için kullanılmıştır. Diğer

Çizelge 5.1 Video kavram sınıflandırması için veri kümesi

Kavram	Eğitim	Test	Toplam
Uçak	472	235	707
Sandalye	682	340	1022
El	1273	636	1909
Dört ayaklı	1304	651	1955
Bebek	327	163	490
Tezahürat	439	219	658
Otoban	601	300	901
Koşmak	318	158	476
Basketbol	101	50	151
Sınıf	256	127	383
Enstrümantal Müzik	1978	989	2967
Şarkı Söylemek	2721	1360	4081
Plaj	1240	620	1860
Bilgisayar	874	436	1310
Göl	286	142	428
Kayakçı	162	81	243
Bisiklete Binmek	238	118	356
Protesto	598	299	897
Motosiklet	234	116	350
Stadyum	496	247	743
Gemi	533	266	799
Bayrak	500	250	750
Haber Stüdyosu	1421	710	2131
Telefon	158	79	237
Köprü	318	158	476
Orman	607	303	910
Gece yarısı	1087	543	1630
Otobüs	90	45	135
George_Bush	242	120	362
Okyanus	949	474	1423
Toplam	20505	10235	30740

tekrarlar için ise bazı kategorilerin örnek sayıları tam üçe bölünemediğinden yaklaşık olarak bu örnek sayılarına eşittir.

5.1.2 Taşıt türü sınıflandırması için veri kümesi

Taşıt türü sınıflandırılması için TRECVID organizasyonunun IACC.1.A-C veri kümesinden faydalanılmıştır. Bu veri kümesinden *zırhlı taşıt, inşaat taşıtı, vinç taşıtı, acil durum taşıtı, askeri taşıt, motosiklet, kurtarma taşıtı* kavramları belirlenerek alt küme oluşturulmuştur. Oluşturulan alt küme videoların sürelerinin toplamı 12 saat 33 dakika 4 saniye süresindedir. Bu veri kümesi için K=3 kat

değerlendirme yöntemi kullanılmıştır. Oluşturulan alt kümenin etiketlenen örnek sayısına göre eğitim ve test sayıları Çizelge 5.2'de verilmiştir. Elde edilen veri kümesine göre ilk tekrarda 654 adet örnek eğitim için kullanılırken 330 örnek test için kullanılmıştır. Deneysel sonuçlar Bölüm 5.4'de sunulmaktadır.

Çizelge 5.2 Taşıt türü sınıflandırması için veri kümesi

Kavram	**Eğitim**	**Test**	**Toplam**
Zırhlı Taşıt	73	37	110
İnşaat Taşıtı	167	84	251
Vinç Taşıtı	30	16	46
Acil Durum Taşıtı	42	21	63
Askeri Taşıt	67	34	101
Motosiklet	233	117	350
Kurtarma Taşıtı	42	21	63
Toplam	654	330	984

5.2 Değerlendirme Yöntemleri

5.2.1 Çapraz doğrulama

Makine Öğrenme teknikleri kullanılarak yapılan çalışmalarda ilk adım veri kümesinin hazırlanmasıdır. Veri kümesi genellikle eğitim verisi, doğrulama verisi ve test verisi olarak üçe bölünür. Doğrulama verisi ile model parametreleri en iyilenir. Eğitim verisi ile Makine Öğrenme yöntemleri eğitilir. Test kümesi ile eğitilen Makine Öğrenme yönteminin performansı ölçülebilir. Örnek olarak toplanan tüm veri kümesinin % 70'i eğitim % 30'u test veri kümesi olarak bölünebilir. Bu oranlar, üzerinde çalışılan göreve göre değişebilir. Fakat eğitilen Makine Öğrenme yönteminin performansının adil değerlendirilebilmesi için bu ayrımın doğru yapılması lazımdır. Aynı zamanda Makine Öğrenme yöntemlerinin aşırı eğitim gibi test verilerinde performansının yüksek çıkması fakat hiç görülmemiş verilerdeki performansının düşük çıkması gibi bir problemi vardır. Aşırı eğitim, *H* hipotez uzayında $h \in H$ hipotezinin aşırı eğitilmiş olması için alternatif $\acute{h} \in H$ hipotezinin eğitim verisindeki hatasının *h* hipotezine göre daha fazla fakat geri kalan tüm veri

uzayındaki hatasının ise daha az olması anlamına gelir [18]. Eğitilen Makine Öğrenme tekniği performansının aşırı eğitim probleminden kaçınmak için çapraz doğrulama tekniği kullanılabilir.

Çapraz doğrulama toplanan bütün veri kümesinin eğitim ve test kümeleri olarak bölünmesinde uygulanabilecek istatistiksel bir yöntemdir. Çapraz doğrulama tekniğinde veri kümesi rastgele *k* adet alt kümeye bölünür. Bu alt kümelerden her defasında farklı bir tanesi test amacı ile seçilir. Geri kalan alt kümeler birleştirilerek seçilen Makine Öğrenme tekniği eğitilir. Her alt küme bir kere test için kullanılacak şekilde *k* kere bu işlem tekrar edilir. Her test sonrası elde edilen performans değerlerinin ortalaması alınarak sistemin performansı elde edilir. Böylece veri kümesindeki her veri test işleminde kullanımış olur. Örneğin k=3 seçilirse veri kümesi üç eşit alt kümeye bölünür. İlk tekrarda bu üç alt kümeden bir tanesi test iki tanesi eğitim için ayrılır. İkinci tekrarda ise test için kullanılmayan bir alt küme test için seçilir ve geri kalan iki alt küme eğitim için kullanılır. Üçüncü tekrarda hiç test için kullanılmamış alt küme test için seçilir ve geri kalan iki alt küme ise eğitim için kullanılır. Her tekrarda elde edilen performans değerlerinin ortalaması bize eğitilen Makine Öğrenme tekniğinin başarısı hakkında bilgi verir.

Bu çalışmada kullanılan veri kümesinin değerlendirilmesinin adil olması ve aşırı eğitim sorununu denetlemek için 3 kat çapraz doğrulama tekniği kullanılmıştır.

5.2.2 Performans kriteri

Makine Öğrenmenin temellerinden biri de öğrenmenin performansının hesaplanmasıdır. Böylece sınıflandırma için kullanılan yöntemleri performansları karşılaştırılabilir ve en uygun yöntem seçilebilir. Eğitilen Makine Öğrenme tekniğinin test sonuçları hata matrisi ile ifade edilir. Hata matrisi Makine Öğrenme tekniğinin verdiği kararları ve test için kullanılan örneklerin ait olduğu sınıfları tek bir tabloda gösterir. Hata matrisinin genel yapısı Çizelge 5.3'de verilmiştir. Hata matrisine göre, sütunlar eğitilen sistemin verdiği karar sınıfları, satırlar ise test veri kümesindeki gerçek sınıflardır. Hata matrisinde görülen TP (Doğru – Kabul) değeri

sistemin tahmin ettiği sınıf pozitif test veri kümesindeki gerçek sınıfında pozitif olduğu örneklerin sayısını ifade eder. FN (Yanlış – Red) değeri ise test veri kümesindeki gerçek sınıfın pozitif fakat eğitilen sistemin tahmininin negatif olduğu örneklerin sayısını ifade eder. FP (Yanlış – Kabul) değeri ise test veri kümesindeki gerçek sınıfın negatif fakat eğitilen sistemin pozitif olarak tahmin ettiği örneklerin sayısını verir. TN (Doğru – Red) değeri ise test veri kümesindeki gerçek sınıfın negatif ve eğitilen sisteminde negatif olduğu örneklerin sayısını vermektedir. Hata matrisi incelenerek tahmin edilen sınıflar ile test verisindeki gerçek sınıflar arasında yorum yapılabilir. Hata matrisinden sınıflandırıcıların performansı

Çizelge 5.3 Hata matrisi

Tahmin Edilen / Gerçek	**Pozitif Sınıf**	Negatif Sınıf
Pozitif Sınıf	**TP**	FN
Negatif Sınıf	FP	TN

hakkında yorum yapabilmemiz ve karşılaştırılabilmek amacı ile belirli ölçütler vardır. Bu ölçütlerden biri doğruluk ölçütüdür. Doğruluk tüm test veri kümesi içerisinde doğru sınıflandırılmış pozitif ve negatif örneklerin toplamının tüm veri kümesindeki örnek sayısına bölümüdür. Doğru sınıflandırılmış verilerin yüzdesini verir. Denklem 5.1'de hata matrisinden elde edilen doğruluk formülü verilmiştir.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Bu tez kapsamında uygulanan yöntemin değerlendirilmesinde doğruluk ölçütü kullanılmıştır.

5.3 Video Kavram Sınıflandırması Deneyleri

Uygulanan yöntem hazırlanan veri kümesinde değerlendirilmiştir. Veri kümesinde bulunan videolar uygulanan yönteme göre görsel kipinde etiketlenen çerçeveler ayrıştırılmıştır. Bu çerçevelere denk gelen 1 saniyelik ses verileri elde edilerek işitsel kip videodan ayrıştırılmıştır. Elde edilen görsel çerçeveler AlexNet ve GoogLeNet ESA mimarileri kullanılarak öznitelik çıkartılmıştır. Görsel tabanlı bir yaklaşım uygulansaydı nasıl sonuçlar elde edilirdi karşılaştırmak için ve önerilen yönteme katkılarını gözlemlemek için görsel özniteliklerden DVM eğitilmiş ve değerlendirilmiştir. Eğitilen DVM doğruluk performans kriteri AlexNet öznitelikleri için Alex-DVM, GoogLeNet öznitelikleri için GoogLeNet-DVM olarak Çizelge 5.4'de gösterilmiştir. Elde edilen ses verilerinden MFCC öznitelikleri çıkarılmış ve uygulanan yönteme göre istatistiksel gösterimleri elde edilmiştir. MFCC özniteliklerinin istatistiksel gösterimleri ile önerilen yönteme olan etkilerini gözlemlemek, birbirleri ile karşılaştırmak amacı ile DVM eğitilmiş ve değerlendirilmiştir. Eğitilen DVM değerlendirilmesinde özniteliklerin temsillerine göre standart sapma öznitelik temsili için MFCC-Std-DVM, ortalama öznitelik temsili için MFCC-Ort-DVM, varyans öznitelik temsili için MFCC-Var-DVM, TBA öznitelik temsili için MFCC-TBA-DVM olarak doğruluk performans kriterleri Çizelge 5.4'de gösterilmiştir. En iyi performansı gösteren işitsel öznitelik temsili ve en iyi performansı gösteren görsel öznitelik füzyon işlemi için seçilmiştir. Füzyon işlemi için seçilen görsel öznitelik ve işitsel öznitelik temsili uç uca eklenerek elde edilen veriler ile DVM eğitilmiştir. Eğitilen DVM doğruluk performans kriteri Çizelge 5.4'de Yöntem ismi ile gösterilmiştir. Eğitilen ve değerlendirilen bütün DVM K=3 kat çapraz doğrulama tekniği kullanılmıştır. Çizelge 5.4'de gösterilen performans

Çizelge 5.4 Video kavram sınıflandırma sonuçları

Yaklaşım	Metot	Doğruluk Ort.	Doğruluk K=1	Doğruluk K=2	Doğruluk K=3
Görsel Tabanlı	Alex-DVM	0,689	0,695	0,684	0,688
	GoogLeNet-DVM	**0,703**	0,697	0,709	0,704
İşitsel Tabanlı	MFCC-Std-DVM	0,271	0,264	0,279	0,27
	MFCC-Ort-DVM	0,239	0,233	0,238	0,247
	MFCC-Var-DVM	0,165	0,167	0,171	0,159
	MFCC-TBA-DVM	**0,376**	0,373	0,379	0,376
Çok Kipli	Önerilen Yöntem	**0,724**	0,728	0,724	0,721

kriterleri çapraz doğrulama sonucu elde edilen 3 doğruluk performans kriterinin ortalaması Doğruluk Ort olarak gösterilmiştir. Çapraz doğrulama sonucu elde edilen 3 doğruluk kriteri ise doğruluk K=1, doğruluk K=2, ve doğruluk K=3 olarak gösterilmiştir.

Görsel tabanlı bir yaklaşım için en iyi sonucu AlexNet özniteliklerine göre daha iyi performans gösteren GoogLeNet öznitelikleri sergilemiştir. İşitsel tabanlı yaklaşımlarda ise MFCC özniteliğinin TBA kullanılarak elde edilen temsili diğer metotlara göre daha başarılı bir performans sergilemiştir. Bu nedenlerden dolayı yöntem olarak belirtilen metotta, MFCC özniteliğinin TBA kullanılarak elde edilen temsili ile GoogLeNet özniteliklerinin füzyon işlemi sonucu elde edilen veriler kullanılmıştır. Elde edilen sonuçlardan çok kipli yaklaşım en yakın rakibi olan GoogLeNet özniteliklerini kullanan yaklaşıma göre yaklaşık %2'lik bir performans artışı ile en başarılı sonucu sergilemiştir. Analizlerde elde edilen hata matrisleri Çizelge 5.5-5.7 olarak sunulmuştur. Görsel tabanlı sınıflandırma için Çizelge

Çizelge 5.5 GoogLeNet-DVM hata matrisi

	Uçak	Sandalye	El	Dört Ayaklı	Bebek	Tezarühat	Otoban	Koşmak	Basketbol	Sınıf	Enstrümantal M.	Şarkı Söylemek	Plaj	Bilgisayar	Göl	Kayakçı	Bisiklete B.	Protesto	Motorsiklet	Stadyum	Gemi	Bayrak	Haber S.	Telefon	Köprü	Orman	Gece yarısı	Otobüs	George B.	Okyanus
Uçak	198	1	2	8	0	0	4	0	0	0	4	1	6	0	0	0	1	0	0	0	2	0	0	0	2	1	2	0	0	3
Sandalye	0	250	6	2	1	1	1	0	0	3	38	14	2	13	0	0	2	1	0	1	0	0	3	0	0	0	0	1	1	0
El	1	2	541	6	6	0	1	0	0	3	18	44	0	5	0	0	0	0	0	0	0	1	1	1	0	0	6	0	0	0
Dört Ayaklı	0	2	3	562	1	0	3	5	0	0	3	14	31	2	0	0	1	5	2	4	0	2	2	0	0	0	6	0	1	2
Bebek	0	1	18	4	131	0	0	0	0	2	0	4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2
Tezarühat	3	2	1	1	1	111	0	0	0	0	5	39	0	0	0	0	0	36	0	6	0	5	2	0	1	0	5	0	1	0
Otoban	2	0	1	3	0	0	253	1	0	0	1	0	3	0	0	0	8	3	2	1	0	0	0	0	11	7	2	2	0	0
Koşmak	1	0	0	9	0	1	7	61	15	0	2	1	4	0	0	0	0	4	1	43	0	1	0	0	0	8	0	0	0	0
Basketbol	1	0	0	1	0	3	0	4	28	0	2	3	0	0	0	0	0	0	0	7	0	0	0	0	0	0	1	0	0	0
Sınıf	1	36	4	1	1	1	0	0	0	60	4	3	0	7	0	1	0	1	0	1	0	0	4	0	0	0	2	0	0	0
Enstrümantal M.	1	16	22	7	0	2	0	1	0	0	723	191	3	0	0	0	0	2	0	0	0	0	7	0	0	0	13	0	1	0
Şarkı Söylemek	1	12	13	10	4	8	5	0	0	2	333	936	2	0	0	0	0	5	0	1	3	5	7	0	1	0	9	0	3	0
Plaj	5	0	2	16	1	0	3	2	0	0	2	2	420	2	25	2	0	0	0	1	25	0	0	0	6	13	2	0	0	91
Bilgisayar	0	27	7	0	0	0	0	0	0	6	1	2	0	374	0	0	0	0	0	0	1	1	9	7	0	0	1	0	0	0
Göl	1	0	1	0	0	0	0	0	0	0	0	0	27	0	43	0	0	0	0	0	18	0	0	0	1	7	1	0	0	43
Kayakcı	2	0	0	8	0	0	0	0	0	0	0	3	1	0	0	52	0	1	0	12	0	0	0	0	1	1	0	0	0	0
Bisiklete Binmek	0	2	5	3	0	1	13	1	0	0	3	4	2	0	0	0	69	4	2	2	1	0	0	0	1	3	2	0	0	0
Protesto	0	1	1	3	0	12	3	0	0	0	8	12	0	0	0	0	4	244	0	1	1	4	2	0	0	1	2	0	0	0
Motosiklet	0	0	0	2	0	0	8	2	0	0	0	1	0	0	0	0	4	4	92	0	0	0	0	0	0	3	0	0	0	0
Stadyum	5	0	0	11	0	3	2	36	10	1	4	9	1	2	0	0	0	8	1	140	4	0	3	0	4	1	2	0	0	0
Gemi	7	1	0	2	0	0	2	0	0	0	1	1	27	0	1	0	0	0	1	0	189	0	0	0	7	0	2	0	0	25
Bayrak	0	3	2	5	0	14	0	1	0	0	4	13	0	2	0	0	0	20	0	0	1	153	13	0	0	1	7	1	10	0
Haber Stüdyosu	0	16	1	4	1	1	0	0	0	0	2	15	0	15	0	0	0	1	0	1	0	2	647	0	0	0	0	0	4	0
Telefon	0	2	17	2	0	0	0	0	0	1	1	5	0	35	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0
Köprü	4	0	0	2	0	0	25	0	0	0	1	2	7	1	0	0	0	0	0	3	7	0	0	0	95	5	4	1	0	1
Orman	0	0	0	11	0	1	9	7	0	0	0	2	10	0	2	1	2	0	7	3	1	0	0	0	1	240	4	0	0	2
Gece yarısı	2	2	4	8	0	7	2	4	0	0	31	34	4	0	7	4	0	5	0	3	2	3	0	0	3	3	412	0	0	3
Otobüs	0	0	3	0	0	0	10	0	0	0	0	1	0	1	0	0	0	1	0	0	3	0	0	0	2	0	1	23	0	0
GeorgeBush	0	5	3	0	0	3	0	0	0	0	0	18	0	1	0	0	0	1	0	0	0	9	19	0	0	0	1	0	60	0
Okyanus	4	0	0	5	0	0	2	0	0	0	0	0	167	0	30	1	0	0	1	0	47	1	0	0	6	3	2	0	0	205

5.5’de GoogLeNet öznitelikleri ile eğitilen DVM hata matrisi verilmektedir. İşitsel tabanlı sınıflandırma için Çizelge 5.6’de MFCC özniteliklerinin TBA ile elde edilen temsili ile eğitilen DVM hata matrisi sunulmuştur. Çok kipli yaklaşım için görsel öznitelik olan GoogLeNet öznitelikleri ile MFCC özniteliklerinin TBA ile elde edilen temsillerinin füzyonu sonucu oluşan veri ile eğitilen DVM hata matrisi Çizelge 5.7’de verilmiştir.

Çizelgelerden görüleceği üzere, TRECVID veri kümesi zorlu video klipleri içermektedir. Bu videolar gerçek hayattan derlenmiş ve çok farklı özelliklerde kavramlar içermektedir. Kategoriler incelendiğinde, bazı kavramların birbirlerine çok benzer olduğu görülmektedir. Buna örnek olarak veri kümesindeki *okyanus*, *plaj*, *göl* ve *gemi* kavramları gösterilebilir. *Okyanus*, *plaj*, *göl* kategorileri video verisinde mekan gibi davranan kategoriler olup hepsinin su ile bağlantısı vardır. Gemi kategorisi de su ile bağlantılı olup bu kategoriler ile benzerlik göstermektedir. *Okyanus* kategorisi için görsel yöntem için Çizelge 5.5 incelendiğinde 205 doğru

Çizelge 5.6 MFCC-TBA-DVM hata matrisi

	Uçak	Sandalye	El	Dört Ayaklı	Bebek	Tezarühat	Otoban	Koşmak	Basketbol	Sınıf	Enstrümantal M.	Şarkı Söylemek	Plaj	Bilgisayar	Göl	Kayakçı	Bisiklete B.	Protesto	Motorsiklet	Stadyum	Gemi	Bayrak	Haber S.	Telefon	Köprü	Orman	Gece yarısı	Otobüs	George B.	Okyanus
Uçak	54	1	19	11	1	1	3	1	0	0	27	50	7	13	0	0	1	1	1	0	1	1	7	0	1	22	6	0	0	6
Sandalye	2	47	19	24	2	3	8	1	0	0	58	66	17	16	0	0	0	6	3	0	4	3	35	0	0	8	9	0	2	7
El	1	6	169	34	0	6	4	0	2	0	69	165	33	27	0	0	0	10	0	3	6	0	68	0	1	8	8	0	1	15
Dört Ayaklı	4	5	44	243	0	2	9	2	0	0	49	109	54	24	0	0	0	9	0	8	7	1	36	0	0	16	10	0	1	18
Bebek	0	1	14	12	31	1	0	0	0	0	18	40	9	7	0	0	0	1	0	3	2	0	8	0	0	4	5	0	0	7
Tezarühat	1	1	3	2	0	77	3	0	1	0	14	75	7	4	0	0	0	8	1	6	1	2	3	0	0	3	5	0	1	1
Otoban	2	0	11	20	0	0	137	0	0	1	18	61	12	5	0	0	4	2	1	0	3	0	9	0	0	4	6	0	0	4
Koşmak	0	0	4	7	0	2	12	27	4	0	20	45	8	1	0	0	0	2	1	4	0	0	9	0	0	9	2	0	0	1
Basketbol	0	0	0	0	0	0	0	0	6	0	15	26	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
Sınıf	0	4	4	5	1	1	2	0	0	21	10	35	6	11	0	0	0	1	0	0	3	0	13	0	0	5	4	0	0	1
Enstrümantal M.	6	7	33	30	1	9	5	1	0	0	401	351	51	16	0	0	0	8	1	6	1	4	22	0	0	3	25	0	0	8
Şarkı Söylemek	6	5	22	23	0	8	6	1	0	0	153	973	50	27	0	0	0	5	3	4	2	2	32	0	0	7	25	0	1	5
Plaj	8	0	25	46	2	2	4	0	0	0	59	144	180	11	1	0	0	4	0	1	8	1	32	0	0	23	9	0	0	60
Bilgisayar	3	11	30	8	1	0	6	0	1	0	29	82	32	130	0	0	0	2	0	1	9	1	79	1	0	4	2	0	0	4
Göl	2	2	3	19	0	1	4	0	0	0	6	23	18	4	7	0	0	1	0	0	6	0	4	0	0	14	6	0	0	22
Kayakcı	1	1	4	3	0	1	0	0	0	0	13	43	3	1	0	2	0	0	0	4	0	0	2	0	0	2	1	0	0	0
Bisiklete Binmek	1	0	6	8	0	1	10	0	0	0	13	22	6	2	0	0	32	3	0	0	0	0	1	0	0	2	7	0	0	4
Protesto	4	2	10	7	1	13	2	0	1	0	16	71	12	7	0	0	0	111	0	0	2	2	11	0	0	4	20	0	0	3
Motosiklet	1	0	0	1	0	0	3	0	0	0	8	75	0	0	0	0	0	1	18	1	0	1	4	0	0	2	1	0	0	0
Stadyum	5	2	2	0	0	1	1	23	2	0	40	73	6	1	0	0	0	5	1	47	0	0	18	0	0	7	10	0	1	2
Gemi	8	1	16	17	1	2	7	0	0	0	17	70	28	18	0	0	0	1	0	2	29	0	11	0	1	26	3	0	0	8
Bayrak	5	8	18	9	0	24	5	0	0	0	23	49	10	13	0	0	0	10	0	7	0	17	32	0	0	5	6	0	8	1
Haber Stüdyosu	4	11	13	12	0	0	0	0	0	0	29	41	23	14	0	0	0	5	0	2	2	2	540	0	0	9	3	0	0	0
Telefon	2	1	9	1	2	0	0	0	0	0	9	6	6	14	0	0	0	2	0	1	2	0	17	0	0	3	2	0	1	1
Köprü	1	0	8	7	0	0	16	0	0	0	9	27	6	4	0	0	0	3	0	1	6	0	2	0	40	16	3	0	0	9
Orman	1	6	18	24	0	0	8	1	1	0	19	49	16	8	0	0	0	7	6	1	9	4	8	0	1	102	4	0	0	10
Gece yarısı	1	1	27	14	4	6	6	0	0	3	37	124	27	15	0	0	1	7	0	7	4	0	9	0	1	8	239	0	0	2
Otobüs	0	0	3	6	1	1	3	0	0	0	4	14	3	1	0	0	0	1	0	0	1	0	1	0	0	2	3	0	0	1
GeorgeBush	0	4	14	2	0	5	3	1	0	0	11	12	7	5	0	0	0	0	0	0	0	2	19	0	0	1	4	0	30	0
Okyanus	12	0	30	28	3	1	4	0	0	2	24	97	81	10	1	0	0	3	0	0	5	0	19	0	1	26	14	0	0	113

karar verilirken, 167 *plaj*, 47 *gemi*, 30 *göl* kararı çıkmıştır. Çok kipli yöntem için Çizelge 5.7 incelendiğinde 230 doğru karar verilirken 153 plaj, 41 gemi, 30 göl kararı verilmiştir. İşitsel tabanlı yöntem için Çizelge 5.6 incelendiğinde 113 doğru karar verilmişken 1 *göl*, 5 *gemi* ve 81 *plaj* kararı çıkmamıştır. Birbirinden kavramsal olarak zor ayrılan kategorilerde çok kipli yaklaşım en yakın rakibi GoogLeNet özniteliklerine dayalı görsel yaklaşımda toplamda doğru karar sayısını artırarak başarılı performans sergilemiştir. Kavramsal olarak birbirinden zor ayrılan kategoriler için işitsel kipten gelen bilgi görsel kipin bulundurduğu bilgiye tamamlayıcı bir etki yaratmıştır.

Kavramsal olarak birbirinden zor ayrılan kategorilere örnek olarak *enstrümantal müzik* ve *şarkı söylemek* kategorileri verilebilir. Enstrümantal müzik şarkı söyleyen bir vokal olmadan, enstrümanlar ile yapılan müziktir. Şarkı söylemek ise genellikle enstrümanlar eşliğinde söylenirken, enstrümansız da söylenebilir. Bu şekildeki bazı kavramların sınıflandırılması insanoğlu için bile zorlu bir problemdir. Görsel

Çizelge 5.7 Önerilen yöntem hata matrisi

	Uçak	Sandalye	El	Dört Ayaklı	Bebek	Tezarühat	Otoban	Koşmak	Basketbol	Sınıf	Enstrümantal M.	Şarkı Söylemek	Plaj	Bilgisayar	Göl	Kayakçı	Bisiklete B.	Protesto	Motorsiklet	Stadyum	Gemi	Bayrak	Haber S.	Telefon	Köprü	Orman	Gece yarısı	Otobüs	George B.	Okyanus
Uçak	194	0	0	6	0	0	3	0	0	0	2	6	7	0	0	0	0	1	0	1	5	0	0	0	3	1	3	0	0	3
Sandalye	0	242	8	3	1	1	1	0	0	6	37	6	1	22	0	0	2	1	0	0	0	1	7	0	0	0	0	0	1	0
El	0	1	550	10	3	0	0	0	0	0	33	25	0	4	0	0	0	0	0	0	0	1	6	0	0	0	3	0	0	0
Dört Ayaklı	4	4	10	551	2	0	4	2	0	0	0	20	29	1	0	0	1	3	0	3	0	0	2	0	0	3	6	1	1	4
Bebek	0	5	23	2	123	0	0	1	0	0	0	7	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Tezarühat	4	0	1	2	1	123	0	0	0	0	10	29	1	0	0	0	0	26	0	7	1	4	1	0	1	0	5	0	3	0
Otoban	4	1	0	0	0	0	260	2	0	0	0	0	1	0	0	1	3	0	3	2	2	0	0	0	10	8	0	3	0	0
Koşmak	0	0	1	11	0	1	8	70	16	0	1	1	2	0	0	0	1	0	1	36	1	2	0	0	0	5	1	0	0	0
Basketbol	0	0	1	3	0	2	0	1	34	0	0	2	0	0	0	0	0	1	0	3	0	0	1	0	0	0	2	0	0	0
Sınıf	0	38	6	1	2	1	0	0	0	53	1	11	0	9	0	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0
Enstrümantal M.	0	12	32	2	0	3	0	0	0	0	755	168	0	4	0	0	0	1	1	1	0	1	1	0	0	0	7	0	1	0
Şarkı Söylemek	1	10	19	4	0	10	1	0	0	1	319	964	2	2	0	0	0	3	0	0	1	3	1	0	0	0	17	1	1	0
Plaj	13	0	1	17	0	0	2	0	0	0	1	3	403	0	31	0	0	0	1	0	26	1	2	0	7	21	2	0	0	89
Bilgisayar	0	19	5	0	0	0	0	0	0	7	1	5	0	386	0	0	0	0	0	0	0	0	7	4	0	0	1	0	1	0
Göl	1	0	1	0	0	0	0	0	0	0	0	0	36	0	49	0	0	0	0	0	9	0	0	0	1	4	0	0	0	41
Kayakcı	0	0	0	5	0	0	1	0	0	0	0	7	3	0	0	60	0	0	0	5	0	0	0	0	0	0	0	0	0	0
Bisiklete Binmek	0	3	2	2	0	1	14	2	0	0	2	2	0	0	0	0	79	6	0	1	0	0	0	0	1	1	1	1	0	0
Protesto	0	3	2	4	0	23	2	0	0	0	4	12	0	0	0	0	2	235	0	0	1	5	0	0	0	0	6	0	0	0
Motosiklet	0	0	0	3	0	0	9	1	0	0	0	2	1	0	0	1	2	3	86	0	0	0	0	0	0	7	1	0	0	0
Stadyum	1	0	3	11	0	7	1	37	10	0	4	12	2	2	0	0	0	5	0	143	3	0	2	0	1	0	3	0	0	0
Gemi	9	0	2	5	0	0	6	0	0	0	1	1	13	0	1	0	0	0	1	0	196	1	0	0	5	2	3	1	0	19
Bayrak	1	1	2	2	0	15	2	0	0	0	4	9	1	2	0	0	0	12	0	1	1	162	18	0	0	0	5	1	11	0
Haber Stüdyosu	1	17	0	1	0	1	1	0	0	0	5	17	0	6	0	0	0	2	0	1	0	5	648	0	0	1	0	0	4	0
Telefon	0	6	12	2	0	0	0	0	0	0	1	6	0	33	0	0	0	0	0	0	0	0	1	18	0	0	0	0	0	0
Köprü	3	0	2	4	0	0	15	0	0	0	0	2	9	0	0	0	0	2	0	0	9	0	1	0	93	7	7	1	0	3
Orman	0	1	1	12	0	0	10	6	0	0	0	1	8	0	9	1	3	0	3	1	0	0	0	0	1	238	5	0	0	3
Gece yarısı	1	1	8	5	0	5	5	0	0	0	32	25	3	2	13	1	1	8	0	6	0	4	0	0	1	1	416	0	0	5
Otobüs	0	2	0	2	0	0	5	0	0	1	0	2	0	2	0	0	1	1	0	0	3	1	0	0	2	0	0	23	0	0
GeorgeBush	0	4	1	0	0	5	0	0	0	0	1	11	0	1	0	0	0	1	0	0	0	6	19	0	0	0	2	0	69	0
Okyanus	6	0	0	3	0	0	2	0	0	0	1	1	153	0	30	2	0	0	0	0	41	1	0	0	2	1	1	0	0	230

yöntem için Çizelge 5.5 incelendiğinde *enstrümantal müzik* kategorisi için 723 doğru karar verilirken baskın hatalı karar olarak 191 *şarkı söylemek* kararı verilmiştir. Aynı şekilde *şarkı söylemek* kategorisi için 936 doğru karar verilirken baskın hatalı karar olarak 333 *enstrümantal müzik* kararı verilmiştir. İşitsel yöntem için Çizelge 5.6 incelendiğinde *enstrümantal müzik* kategorisi için 401 doğru karar verilirken baskın hatalı karar olarak 351 *şarkı söylemek* kararı verilmiştir. Aynı şekilde şarkı söylemek kategorisi için 973 doğru karar verilirken baskın hatalı karar olarak 153 *enstrümantal müzik* kararı verilmiştir. Çok kipli yöntem için Çizelge 5.7 incelendiğinde *enstrümantal müzik* kategorisi için 755 doğru karar verilirken baskın hatalı karar olarak 168 *şarkı söylemek* kararı verilmiştir. Aynı şekilde şarkı söylemek kategorisi için 964 doğru karar verilirken baskın hatalı karar olarak 319 *enstrümantal müzik* kararı verilmiştir. Bu iki kategori arasında yapılan doğru karar kıyaslamasında, *şarkı söylemek* kategorisi için işitsel yöntem en iyi performansı sergilemiştir. Bu kategoriler içinde işitsel kip ve görsel kip bir biri ile tamamlayıcı bir etki göstererek çok kipli yöntem için toplamda doğru karar sayısı artmaktadır.

Genellikle tüm çapraz doğrulamalardan elde edilen hata matrisleri incelendiğinde kendine özgü ses olmayan sandalye, köprü, orman, gibi kategoriler için çok kipli yöntem, görsel yönteme yakın bir doğru karar performansı sergilemiştir. Görünüş olarak diğer kategorilerden farklı uçak, sandalye, el, otobüs gibi kategoriler için görsel tabanlı yaklaşım ile çok kipli yaklaşım bir birine yakın performans sergilemiştir. Aynı zamanda diğer kategorilere göre daha ayırt edici kendine özgü sesi olan tezahürat, otoban, bebek, telefon gibi kategoriler için çok kipli yöntem daha başarılıdır. Yine eylemsel kategoriler olan tezahürat, koşmak, *şarkı söylemek*, bisiklete binmek kategorilerinde çok kipli yöntem daha başarılıdır.

Deneysel sonuçlardan elde edilen bilgilere göre kendine özgü sesi olan, eylemsel olan ve birbirinden kavramsal olarak zor ayrılan kategoriler için çok kipli yöntem başarılı sonuçlar elde etmektedir. Statik görünüm özelliği diğerlerinden ayrılabilen kategoriler için ise görsel yöntem ile bir birine yakın performanslar sergilemiştir. Toplamda bütün kategoriler için performans görsel yönteme göre %2'lik bir

performans artışı ile çok kipli önerilen yöntem en başarılı performansı sergilemektedir. Bunun ana nedeni video verisinin işitsel kipi ile görsel kipinin bir birlerine olan tamamlayıcı etkileridir. Görsel kip ile işitsel kipin veri boyutları ele alındığında 1x1024 boyutunda olan görsel öznitelikler 1x20 boyutunda olan işitsel özniteliklerin temsiline göre oldukça büyük boyuttadır. Bu boyut farkında bile %2'lik bir performans artışı sağlanmıştır.

5.4 Taşıt Türü Sınıflandırması Deneyleri

Önerilen yöntem, taşıt türü sınıflandırması için hazırlanan veri kümesi üzerinde değerlendirilmiştir [25]. Bu değerlendirme ile elde edilen sınıflandırma sonuçları Çizelge 5.8'de verilmiştir. Önerilen yöntemi taşıt türü sınıflandırması uygulaması için görsel yaklaşımın ve işitsel yaklaşımın çok kipli olan yaklaşıma olan etkileri değerlendirilmektedir. Bu değerlendirme için görsel yaklaşımı oluşturan AlexNet ve GoogLeNet mimarilerinin son tam bağlı katmanından elde edilen öznitelikler ile eğitilen DVM performansları sırasıyla Alex-DVM ve GoogLeNet-DVM ile gösterilmektedir. İşitsel kipli yaklaşım için elde edilen MFCC özniteliklerinin ortalama, standart sapma, varyans ve TBA ile oluşturulan temsilleri ile eğitilen DVM performansları sırasıyla MFCC-Ort-DVM, MFCC-Std-DVM, MFCC-Var-DVM ve MFCC-TBA-DVM ile gösterilmektedir. En son olarak önerilen yöntem olan işitsel özniteliklerden en iyi performansı sergileyen MFCC özniteliklerinin TBA ile temsili ile en iyi performansı sergileyen görsel öznitelik olan GoogLeNet özniteliklerinin uç uca eklenerek elde edilen çok kipli öznitelikler ile eğitilen DVM performansı önerilen yöntem olarak verilmektedir. Değerlendirmenin adil yapılması için K=3 kat çapraz doğrulama tekniği uygulandığından elde edilen 3 doğruluk

Çizelge 5.8 Taşıt türü sınıflandırma sonuçları

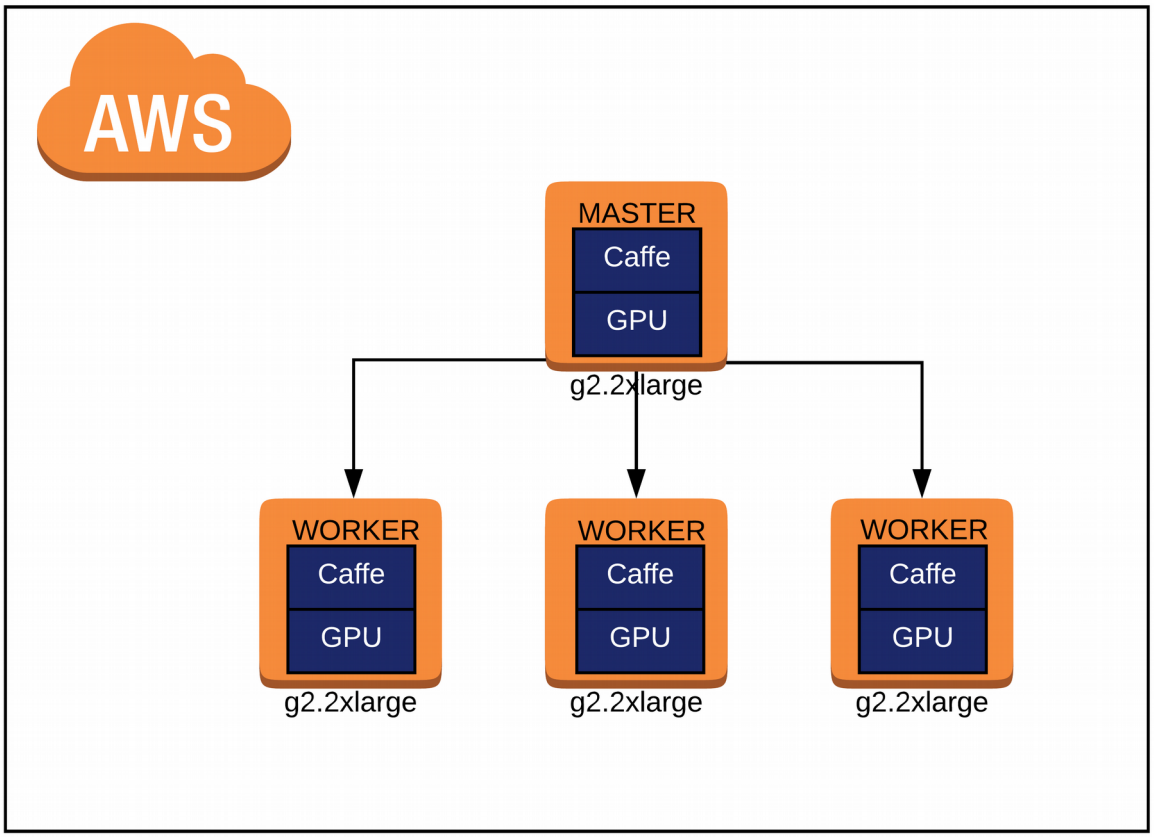
Yaklaşım	Metot	Doğruluk Ort.	Doğruluk K=1	Doğruluk K=2	Doğruluk K=3
Görsel Tabanlı	Alex-DVM	0,686	0,68	0,674	0,706
	GoogLeNet-DVM	**0,701**	0,698	0,696	0,709
İşitsel Tabanlı	MFCC-Std-DVM	0,52	0,517	0,549	0,495
	MFCC-Ort-DVM	0,43	0,426	0,423	0,443
	MFCC-Var-DVM	0,372	0,373	0,371	0,373
	MFCC-TBA-DVM	**0,568**	0,563	0,576	0,565
Çok Kipli	Önerilen Yöntem	**0,721**	0,71	0,72	0,733

değeri Doğruluk K=1, Doğruluk K=2 ve Doğruluk K=3 şeklinde gösterilmiştir. Doğruluk Ort ile gösterilen değerler, K=3 kat çapraz doğrulama tekniğinden elde edilen 3 doğruluk değerinin ortalamasıdır.

Elde edilen sonuçlar göstermektedir ki kavram sınıflandırma probleminde olduğu gibi GoogLeNetin son tam bağlı katmanından elde edilen öznitelikler AlexNetin son tam bağlı katmanından elde edilen özniteliklerden daha başarılı bir performans sergilemiştir. TBA metodunun bilgiyi koruyarak boyut indirgeme motivasyonu MFCC özniteliklerinin TBA metodu ile elde edilen temsilleri eğitilen DVM işitsel yöntemler arasında en başarılı performansı sergilemesini sağlamıştır. İşitsel yöntem için MFCC özniteliklerinin TBA ile elde edilen temsillerinin gösterdiği performansı standart sapma, ortalama ve varyans ile elde edilen temsillerinin performansları takip etmektedir. Önerilen yöntem olan GoogLeNet öznitelikleri ile MFCC özniteliklerinin TBA temsilinin füzyon işlemi sonucu oluşan çok kipli öznitelikler ile eğitilen DVM en başarılı performansı sergilemiştir. Çok kipli yöntem, en yakın performansı sergileyen GoogLeNet öznitelikleri ile eğitilen DVM performansına yaklaşık olarak %2'lik bir performans artışı göstererek %72,1'lik doğruluk performansı göstermiştir. Görsel kip ile işitsel kipin arasındaki tamamlayıcı etki çok kipli yöntemin en başarılı performansı göstermesini sağlamıştır.

5.5 Apache Spark Deneyleri

Büyük veri işleme platformlarının, uygulanan yöntemlere etkilerini incelemek amacıyla uygulanan yöntem Apache Spark üzerine taşınmıştır. Bu çalışmayı gerçekleştirebilmek için 1 master, 3 worker'dan oluşan Apache Spark bilgisayar kümesi oluşturulmuştur. Oluşturulan kümenin temsili Şekil 5.1'de verilmektedir. Bu küme üzerinde uygulanan yöntemin ana aşamaları olan video verisinin kiplerine ayrılması, görsel öznitelik çıkarımı, işitsel öznitelik çıkarımı ve temsillerinin sağlanması ve DVM eğitimi ve testi gerçekleştirilmiştir. Video içerik analizinde farklı yöntemler için bu aşamalar değişebileceğinden, her aşamanın hesaplama zamanı performansı ölçümlenmiştir. Hesaplama zamanı ölçümleri için Apache



Şekil 5.1 Geliştirilen Apache Spark kümesinin temsili gösterimi

Spark'ın sağlamış olduğu ekranlama (monitoring) [91] özelliği kullanılmıştır. Ekranlama özelliği küme üzerinde çalışan uygulamanın tamamlanması için geçen süreyi vermektedir. Apache Spark küme boyutunun hesaplama zamanı performansına etkisini gözlemlemek amacı ile 1 master ve 1 worker, 1 master 2 worker, 1 master 3 worker'dan oluşan Apache Spark kümeleri oluşturulmuştur.

Bu çalışmayı gerçekleştirebilmek için Amazon firmasının sunmuş olduğu Amazon Web Servislerinden Elastic Compute Cloud (EC2) servisi kullanılmıştır. EC2 bulut ortamında güvenli ve ölçeklendirilebilir hesaplama kapasitesi sağlayan bir web servisidir [92]. EC2 web servisi farklı amaçlar için kullanılabilecek, çok sayıda çeşidi olan, örnek dedikleri bilgisayarlar sağlamaktadır. Bu örneklerden g2.2xlarge örneği seçilmiştir. g2.2xlarge 8 tane merkezi işleme birimi, 1 tane grafik işleme birimi ve 60 GB büyüklüğünde SSD depolama birimi bulunmaktadır. g2.2xlarge

örneğinin seçilme nedeni, ESA mimarilerinin yapılan çalışmalarda [93] grafik işleme birimi üzerinde çalışırken merkezi işleme birimi üzerinden çalışmasına göre oldukça performanslı olmasıdır. Aynı zamanda 60 GB depolama biriminin olması ve bu depolamam birimine ekleme yapılarak büyüklüğünün artırılması gibi nedenlerden tercih edilmiştir.

Kullanılan kütüphaneler, sürücüler performansın ölçülmesinde etkili olmaktadır. Bilgisayarların işletim sistemi olarak Linux tabanlı Ubuntu 14.04 LTS işletim sistemi seçilmiştir. Yapılan bütün kodlamalar için Python programlama dili kullanılmıştır. Video verisinin görsel ve işitsel kiplerine ayrılması için FFMPEG [53] kütüphanesi tercih edilmiştir. ESA mimarilerinden görsel öznitelik çıkarımı için Caffe [93] derin öğrenme çatısı kullanılmıştır. Bulunan grafik işleme birimi NVIDIA firmasının ürünü olduğu için Cuda 7.5 sürücüsü yüklenmiştir. NVIDIA firmasının yaptığı çalışmalarda cuDNN (Cuda Deep Neural Network) kütüphanesi, ESA mimarilerinin grafik işleme birimi üzerinde çalışırken performansı artırdığı gözlemlenmiştir [94]. Bu sebeple ve grafik işleme birimi sürücüsü ile uyumlu olmasından cuDNN v5.1 tercih edilmiştir. ESA mimarilerinden en iyi performansı sergilediği için GoogLeNet seçilmiştir. İşitsel özniteliklerin çıkarılmasında python_speech_features [95] kütüphanesi kullanılmıştır. DVM ve TBA için Scikit-learn [96] kütüphanesi kullanılmıştır.

Apache Spark çalışması için video kavram sınıflandırma için kullanılan veri kümesi kullanılmıştır. Bu veri kümesi 5103 videodan oluşmakta ve video sürelerinin toplamı 8 gün 20 saat 10 dakika 3 saniyedir. Bu veri kümesi Apache Spark'ın erişebilmesi için kümede yer alan tüm bilgisayarlarda aynı dosya dizininde bulunmaktadır. Apache Spark'ta RDD oluşturmak için video isimleri ve anahtar çerçeve süreleri **video-ismi anahtar-çerçeve-süresi** olacak şekilde metin dosyası hazırlanmıştır. Hazırlanan metin dosyası Apache Spark uygulaması içerisinde RDD'ye çevrilmiştir. RDD'yi oluşturan elementler metin dosyasının satırlarıdır. Apache Spark kümesinin boyutuna ve bulunan merkezi işleme birimlerine göre RDD'yi parçalarına ayrılmaktadır. Video isimlerinden ve anahtar çerçeve sürelerinden elde edilen RDD görsel ve işitsel kiplerine ayrılmak için hazırlanan

fonksiyona haritalandırılır. RDD'yi oluşturan video isimleri ve anahtar çerçeve sürelerinden oluşan her elemente belirlenen dosya dizini eklenerek işitsel ve görsel kiplere ayırmak için hazırlanan fonksiyon uygulanır. Bu işlem sonunda işitsel ve görsel kipler kendileri için belirlenen dosya dizinine kaydedilerek elde edilir. Video verilerine uygulanan yöntem ile aynı işleyişte, çok kipli içerik analizi yönteminde belirtildiği gibi kendilerine özel öznitelik çıkarımı fonksiyonları olmakla şartıyla, imge ve ses dosyalarına uygulanır. Bu işlem sonucunda işitsel ve görsel öznitelikler belirlenen dosya dizinine kaydedilir. Elde edilen öznitelikler DVM aşaması için **öznitelik kavram** olacak şekilde tek bir metin dosyasında birleştirilir. Bu metin dosyası RDD'ye çevrilir ve DVM eğitmek ve test etmek amacı ile hazırlanmış fonksiyona haritalanır. DVM eğitim ve testi için hazırlanan fonksiyon diğer fonksiyonlar gibi RDD'yi oluşturan her elemente uygulanmaz. DVM eğitmek bir grup özniteliğe ihtiyaç vardır. Bu sebeple RDD gruplar halinde parçalarına ayrılmaktadır. Gruplar halinde parçalarına ayrılan öznitelik kavram çiftlerinin %67'si eğitim %33'ü test için ayrılır. Böylece büyük veri platformunun doğruluk performansına etkisi gözlemlenir. Aşamalarına göre hesaplama zamanı performansları Çizelge 5.9'da verilmiştir.

Çizelge 5.9 Küme boyutuna göre hesaplama zamanları

	Kiplere Ayrıştırma	İşitsel Öznitelik Çıkarımı	Görsel Öznitelik Çıkarımı	DVM Eğitimi &Test
1 Master 1 Worker	7,5 dk	5,8 dk	17 dk	23 s
1 Master 2 Worker	3,9 dk	2,6 dk	8,9 dk	15 s
1 Master 3 Worker	2,6 dk	2 dk	6,1 dk	13 s

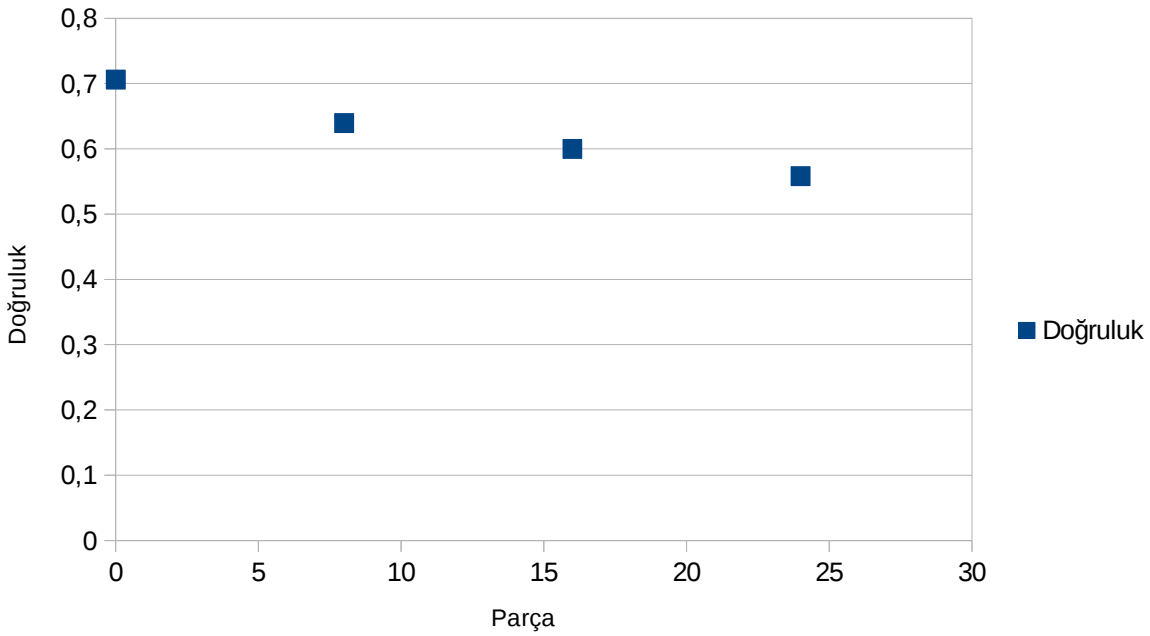
Çizelge 5.9'da görüldüğü gibi çok kipli içerik analizi için uygulanan yöntemin her aşaması için küme boyutu arttıkça hesaplama zamanı performansı düşmektedir. Video verisinin görsel ve işitsel kiplerine ayrılması, işitsel öznitelik çıkarımı, ve DVM eğitilmesi küme çapında var olan merkezi işleme birimleri üzerinde gerçekleştirmiştir. Görsel öznitelik çıkarımı ise küme çapında bulunan grafik işleme birimleri üzerinde gerçekleştirilmiştir. Görsel öznitelik çıkarımı, ESA mimarilerinin doğasından dolayı en çok zaman alan işlemdir. ESA mimarilerinden öznitelik çıkarımı işlemi için hesaplama zamanını 3 grafik işleme birimi üzerinde çalıştığında 1 grafik işleme birimi üzerinde çalışmasına göre yarıdan fazla

düşmüştür. Elde edilen bütün sonuçlara bakıldığında DVM hariç worker sayısı *n* ile hesaplama zamanı *t* arasında yaklaşık t/n gibi bir oran gözlemlemiştir. DVM ise hesaplama zamanı sergilememesinin nedeni saniyeler içinde eğitim ve test işlemlerini gerçekleşmesidir.

Ayrıca Apache Spark farklı küme boyutları için, bulanan kaynağa göre, veri kümesini RDD'ye çevirirken parçalarına ayırır. DVM eğitmek için hazırlanan veri kümesi RDD tarafından parçalarına ayrılması nedeniyle veri kümesindeki örnek sayısında düşüş yaşanmıştır. Bu örnek sayılarına göre eğitilen DVM doğruluk oranları düşmektedir. Örneğin 1 master 3 worker'dan oluşan bilgisayar kümesi için veri kümesi toplamda 24 parçaya ayrılacaktır. Her parçada bulunan toplam örnek sayısına göre % 67'si eğitim verisi, %33'ü test verisi olacak şekilde bölünür. 1 master 3 worker'dan oluşan bilgisayar kümesi için parçalara göre doğruluk oranları Çizelge 5.10'da verilmiştir. Bu doğruluk oranlarının ortalaması 0,558'dir. Veri kümesi 1 master 2 worker'dan oluşan bilgisayar kümesi için 16 parçaya bölünmüştür. Aynı şekilde 1 master 1 worker'dan oluşan bilgisayar kümesi için 8 parçaya bölünür. Bölünen parçalara göre doğruluk oranlarının ortalaması Şekil 5.2'de verilmiştir. İlk olarak Apache Spark kullanılmadan veri kümesi % 67'si eğitim verisi, %33'ü test verisi olacak şekilde bölünmüş ve 0,706'lık bir doğruluk oranı

Çizelge 5.10 Bir kümedeki parçalara göre doğruluk oranları

Parça	Doğruluk	Parça	Doğruluk
1	0,562	13	**0,617**
2	0,564	14	0,569
3	0,533	15	0,563
4	0,576	16	0,498
5	0,556	17	0,548
6	0,585	18	0,537
7	0,55	19	0,588
8	0,556	20	0,558
9	0,56	21	0,562
10	0,582	22	0,562
11	0,548	23	0,547
12	0,524	24	0,538



Şekil 5.2 Parça doğruluk grafiği

sergilemiştir. 1 master 1 worker'dan oluşan bilgisayar kümesinde 8 parçaya bölünen veri kümesi için doğruluk oranları ortalaması 0,639'dur. 1 master 2 worker'dan oluşan bilgisayar kümesinde 16 parçaya bölünen veri kümesi için doğruluk oranları ortalaması 0,599'dur. 1 master 3 worker'dan oluşan bilgisayar kümesinde 24 parçaya bölünen veri kümesi için doğruluk oranları ortalaması 0,558'dir. Bulunan kaynak sayısı arttıkça DVM doğruluk performansı düşmektedir.

6 SONUÇLAR VE TARTIŞMA

Bu tez kapsamında, video verisi için çok kipli içerik analizi yöntemi önerilmiş ve büyük veri platformlarından Apache Spark küme ortamında gerçeklenmiştir. Video verisinin görsel kipi için, AlexNet ve GoogLeNet gürbüz ESA mimarilerinin son tam bağlı katmanlarından görsel öznitelikler çıkarılmıştır. İşitsel kip için, MFCC öznitelikleri çıkarılmış ve standart sapma, ortama, varyans ve TBA kullanılarak temsilleri oluşturulmuştur. İşitsel öznitelik temsilleri ve görsel öznitelikler uç uca eklenerek füzyon işlemi uygulanmıştır. Füzyon işlemi sonucu elde edilen veriler ile DVM eğitilmiştir. DVM için, radyal tabanlı çekirdek fonksiyonu ve çok sınıflı sınıflandırma stratejisi olarak Bir Karşı Bir stratejisi kullanılmıştır. Önerilen yöntem, 5103 adet videodan oluşan, 8 gün 20 saat 10 dakika 3 saniye süresinde, TRECVID IACC.1.A-C veri kümesinden yaratılan, alt küme üzerinde değerlendirilmiştir. Oluşturulan veri kümesi bir birinden anlamsal olarak ayrılması zor kavramların bulunduğu, 30 kavramdan oluşmaktadır. Oluşturulan veri kümesinde önerilen yöntemi değerlendirmek için çapraz doğrulama tekniği kullanılmıştır. Sınıflandırmaların değerlendirilmesinde doğruluk performans kriteri kullanılmıştır. Önerilen yöntem, EC2 servisi kullanılarak, 4 bilgisayardan oluşturulan Spark küme ortamında çalıştırılmıştır. Büyük veri platformunun video verisinin çok kipli içerik analizin olan etkileri gözlemlenmiştir. Küme boyutu ve önerilen yöntemin aşamalarına göre elde edilen hesaplama zamanları sunulmuştur.

Yapılan deneysel çalışmalar gösteriyor ki video verisinin görsel kipinden öznitelik çıkarımı için kullanılan ESA mimarisi GoogLeNet'in son tam bağlı katmanından elde edilen öznitelikler, AlexNet'in son tam bağlı katmanından elde edilen özniteliklere göre başarılı performans sergilemiştir. GoogLenet AlexNet'e göre daha az parametresi bulunmasına rağmen daha fazla katmanı bulduğu için imge içerisinde belirlenen kavramı daha iyi temsil eden öznitelikler sergilemiştir. Gösterdiği başarılı performans ile video içerik analizinde kullanılması için tercihi edilmesi düşünülebilir. Aynı zamanda GoogLeNet'in elde edilen öznitelikler 1024

boyutunda olup AlexNet'in özniteliklerinin 1/4'ü kadardır. Bu boyut farkı, bellek karmaşıklığı göz önünde bulundurulduğunda, bir avantaj sağlamıştır. GoogLeNet'in daha fazla katmanı bulunmasında dolayı hesaplama maliyeti için dezavantajlıdır.

İşitsel kip için yapılan çalışmalarda, en iyi performansı MFCC özniteliklerinin TBA yöntemi kullanılarak boyut indirgenmesinden sonra elde edilen temsili sergilemiştir. TBA'nın bilgiyi koruyarak boyut indirgeme motivasyonu, MFCC özniteliklerinin temsileri arasında en başarılı performansı sergilemesini sağlamıştır. TBA temsilini standart sapma, ortalama ve varyans temsilleri takip etmektedir. MFCC özniteliklerinin standart sapma, ortalama ve varyans işlemleri kullanılarak elde edilen temsiller bilgi kaybına yol açmıştır. İşitsel özniteliklerinin temsilleri genel olarak görsel özniteliklerden başarısız bir performans sergilemiştir. Bunun sebebi belirlenen kavramlardan bazılarının kendine özgü sesi olmayışıdır. Aynı zamanda video verileri içerisinde belirlenen kavramlara özgü ses verisinin olup olamadığı bilinmemektedir.

Önerilen çok kipli yöntem için kullanılan, MFCC özniteliklerinin TBA temsili ve GoogLeNet özniteliklerinin uç uca eklenerek elde edilen çok kipli öznitelik en iyi performansı sergilemiştir. En yakın rakibi olan GoogLeNet özniteliklerini kullanan, görsel tabanlı yöntemden performansı daha yüksektir. GoogLeNet öznitelikleri gösterdiği performans ile belirlenen kavramların statik görünüm özelliğini temsil etmiştir. Fakat göl, okyanus gibi statik görünüm özelliği belirlenmesi zor kavramlar için genel olarak sergilediği performanstan düşük performans sergilemiştir. GoogLeNet'in özniteliklerinin sınıflandırma performansında düşüş yaşadığı kavramlar için, MFCC özniteliklerinin TBA temsili tamamlayıcı etki göstererek, önerilen yöntemin performansını katkı sağlamıştır. GoogLeNet özniteliklerinin 1x1024 boyutunda ve MFCC özniteliklerinin TBA temsilinin 1x20 boyutunda olduğu göz önünde bulundurulduğunda, görsel özniteliklerin işitsel özniteliklere göre sınıflandırma sistemleri için daha baskındır. Bu boyut farkı, GoogLeNet özniteliklerinin performansının çok kipli içerik analizi performansındaki etkisini artırarak avantaj sağlamıştır. Fakat işitsel özniteliklerin, sınıflandırma için tam

olarak temsil edilememesini sağlamıştır.

Apache Spark platformu üzerinde gerçekleştirilen çalışmadan elde edilen sonuçlardan, çok kipli video içerik analizinin hesap zamanı performansı için önemli gelişmeler gözlemlenmiştir. Apache Spark gibi büyük veri platformlarının ölçeklendirilebilir olması, büyük boyuttaki veri kümeleri kullanarak, yöntemlerin eğitilmesine ve değerlendirilmesine olanak sağlar. Apache Spark kullanılarak eğitilen DVM doğruluk performansından düşüş gözlemlenmiştir.

Gelecek çalışmalar arasında farklı füzyon tekniklerinin öznitelik ve/veya model düzeyinde probleme uygulanması yer almaktadır. Video verisinin ses kipi için derin öğrenme mimarilerinin kullanılması planlanmaktadır. Video verisinin ses kipinin kavrama özgü ses barındırıp barındırmadıkları değerlendirilecek ve değerlendirme sonucu elde edilen bilgiler ile yöntem geliştirilecektir. DVM eğitiminde kullanılan öznitelikler için öznitelik seçimi tekniklerinin kullanılması planlanmaktadır. Apache Spark platformu için, video verilerinin Hadoop dağıtık dosyalama sisteminde saklanması sağlanarak, video verilerinin kümeyi oluşturan her bilgisayarda kopyasının saklanmasının önüne geçilebilir.

KAYNAKLAR LİSTESİ

[1] K. David, “Latest content ID tool for YouTube,” 2007. [Online]. Available: https://googleblog.blogspot.com.tr/2007/10/latest-content-id-tool-for-youtube.html. [Accessed: 08-Jan-2018].

[2] C. V. M. L. Team, “An On-device Deep Neural Network for Face Detection,” 2017. [Online]. Available: https://machinelearning.apple.com/2017/11/16/face-detection.html. [Accessed: 08-Jan-2018].

[3] “License Plate/Number Plate Detection and Recognition.” [Online]. Available: https://www.intelli-vision.com/license-plate-recognition/. [Accessed: 01-Jan-2018].

[4] J. P. Daly, C. A. Davis, and J. D. Tuton, “Traffic violation processing system,” 1999.

[5] M. L. Eichner and T. P. Breckon, “Integrated speed limit detection and recognition from real-time video,” in 2008 IEEE Intelligent Vehicles Symposium, 2008, pp. 626–631.

[6] J. Sang and C. Xu, “On Analyzing the ‘Variety’ of Big Social Multimedia,” in 2015 IEEE International Conference on Multimedia Big Data, 2015, pp. 5–8.

[7] J. R. Smith, “History Made Every Day,” IEEE Multimed., vol. 18, no. 3, pp. 2–3, Mar. 2011.

[8] Y.-G. Jiang, “Categorizing Big Video Data on the Web: Challenges and Opportunities,” in 2015 IEEE International Conference on Multimedia Big Data, 2015, pp. 13–15.

[9] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., vol. 28, no. 4, pp. 357–366, Aug. 1980.

[10] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-Up Robust Features (SURF),” Comput. Vis. Image Underst., vol. 110, no. 3, pp. 346–359, 2008.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in 2011 International Conference on Computer

Vision, 2011, pp. 2564–2571.

[13] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, vol. 1, pp. 886–893.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[15] C. Szegedy et al., “Going deeper with convolutions,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[16] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[17] B. V. Dasarathy, “Sensor fusion potential exploitation-innovative architectures and illustrative applications,” Proc. IEEE, vol. 85, no. 1, pp. 24–38, 1997.

[18] T. M. Mitchell, Machine Learning, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[19] C. Cortes and V. Vapnik, “Support-Vector Networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[20] D. Oneata, J. Verbeek, and C. Schmid, “Action and Event Recognition with Fisher Vectors on a Compact Feature Set,” in 2013 IEEE International Conference on Computer Vision, 2013, pp. 1817–1824.

[21] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” Work. Stat. Learn. Comput. Vision, ECCV, pp. 1--22, 2004.

[22] M. Zaharia et al., “Apache Spark,” Commun. ACM, vol. 59, no. 11, pp. 56–65, Oct. 2016.

[23] A. Spark, “Spark Overview.” [Online]. Available: https://spark.apache.org/docs/latest/. [Accessed: 03-Jan-2017].

[24] B. Selbes and M. Sert, “Multimodal video concept classification based on convolutional neural network and audio feature combination,” in 2017 25th

Signal Processing and Communications Applications Conference (SIU), 2017, pp. 1–4.

[25] B. Selbes and M. Sert, “Multimodal vehicle type classification using convolutional neural network and statistical representations of MFCC,” in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–6.

[26] A. Oliva and A. Torralba, “Chapter 2 Building the gist of a scene: the role of global image features in recognition,” 2006, pp. 23–36.

[27] D. Navon, “Forest before trees: The precedence of global features in visual perception,” Cogn. Psychol., vol. 9, no. 3, pp. 353–383, Jul. 1977.

[28] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, “Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study,” Int. J. Comput. Vis., vol. 73, no. 2, pp. 213–238, 2007.

[29] A. E. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3D scenes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 5, pp. 433–449, May 1999.

[30] Jingen Liu, Jiebo Luo, and M. Shah, “Recognizing realistic actions from videos “in the wild”,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1996–2003.

[31] M. Leordeanu and M. Hebert, “A spectral technique for correspondence problems using pairwise constraints,” in Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, 2005, p. 1482–1489 Vol. 2.

[32] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior Recognition via Sparse Spatio-Temporal Features,” in 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72.

[33] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2929–2936.

[34] M. Sert and H. Ergiin, “Video scene classification using spatial pyramid based features,” in 2014 22nd Signal Processing and Communications Applications Conference (SIU), 2014, pp. 1946–1949.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[36] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective Search for Object Recognition,” Int. J. Comput. Vis., vol. 104, no. 2, pp. 154–171, 2013.

[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[38] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[39] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” Dec. 2012.

[40] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN Features off-the-shelf: an Astounding Baseline for Recognition,” Mar. 2014.

[41] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,” Dec. 2013.

[42] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, “Audio keywords generation for sports video analysis,” ACM Trans. Multimed. Comput. Commun. Appl., vol. 4, no. 2, pp. 1–23, May 2008.

[43] K. Lee and D. P. W. Ellis, “Audio-Based Semantic Concept Classification for Consumer Video,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 18, no. 6, pp. 1406–1416, Aug. 2010.

[44] K. Lee, D. P. W. Ellis, and A. C. Loui, “Detecting local semantic concepts in environmental sounds using Markov model based clustering,” in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 2278–2281.

[45] C. Okuyucu, M. Sert, and A. Yazici, “Audio Feature and Classifier Analysis for Efficient Recognition of Environmental Sounds,” in 2013 IEEE International Symposium on Multimedia, 2013, pp. 125–132.

[46] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, “Dense Trajectories and Motion Boundary Descriptors for Action Recognition,” Int. J. Comput. Vis., vol. 103, no. 1, pp. 60–79, May 2013.

[47] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, “Image Classification with the Fisher Vector: Theory and Practice,” Int. J. Comput. Vis., vol. 105, no. 3, pp. 222–245, Dec. 2013.

[48] Y.-G. Jiang et al., “Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching,” NIST TRECVID Work., 2010.

[49] I. Laptev, “On Space-Time Interest Points,” Int. J. Comput. Vis., vol. 64, no. 2–3, pp. 107–123, Sep. 2005.

[50] H. Tan and L. Chen, “An approach for fast and parallel video processing on Apache Hadoop clusters,” in 2014 IEEE International Conference on Multimedia and Expo (ICME), 2014, pp. 1–6.

[51] “What Is Apache Hadoop?” [Online]. Available: http://hadoop.apache.org/. [Accessed: 25-Nov-2017].

[52] G. Bradski, “The OpenCV Library,” Dr. Dobb’s J. Softw. Tools, 2000.

[53] “FFmpeg.” [Online]. Available: https://www.ffmpeg.org/. [Accessed: 29-Nov-2017].

[54] S. Yang and B. Wu, “Large Scale Video Data Analysis Based on Spark,” in 2015 International Conference on Cloud Computing and Big Data (CCBD), 2015, pp. 209–212.

[55] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[56] H. Wang, Xiaobin Zheng, and Bo Xiao, “Large-scale human action recognition with spark,” in 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP), 2015, pp. 1–6.

[57] G. Lu, Multimedia Database Management Systems. Norwood, MA, USA: Artech House, Inc., 1999.

[58] I. Jolliffe, Principal Component Analysis. New York: Springer-Verlag, 2002.

[59] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex.,” J. Physiol., vol. 195, no. 1, pp. 215–43, Mar. 1968.

[60] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[61] “Convolutional Neural Network.” [Online]. Available: http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/. [Accessed: 16-Jul-2017].

[62] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[63] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[64] L. Torrey and J. Shavlik, “Transfer learning,” Handb. Res. Mach. Learn. Appl. Trends Algorithms, Methods, Tech., vol. 1, p. 242, 2009.

[65] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in Proceedings of the fifth annual workshop on Computational learning theory - COLT ’92, 1992, pp. 144–152.

[66] S. Shahrivari, “Beyond Batch Processing: Towards Real-Time and Streaming Big Data,” Computers, vol. 3, no. 4, pp. 117–129, 2014.

[67] J. Bloomberg, “The Big Data Long Tail.” [Online]. Available: http://www.devx.com/blog/the-big-data-long-tail.html. [Accessed: 23-Oct-2017].

[68] A. Jacobs, “The Pathologies of Big Data,” Commun. ACM, vol. 52, no. 8, pp. 36–44, 2009.

[69] “Demystifying big data: A practical guide to transforming the business of Government,” 2012. [Online]. Available: https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf.

[70] “What is big data?” [Online]. Available: https://www.ibm.com/analytics/us/en/big-data/.

[71] “What is big data?” [Online]. Available: https://msdn.microsoft.com/en-us/library/dn749868.aspx.

[72] “What is Big Data?” [Online]. Available: https://www.oracle.com/big-data/index.html.

[73] “What is big data?” [Online]. Available: https://cloud.google.com/what-is-big-data/.

[74] D. Laney, “3D Data Management Controlling Data Volume Velocity and Variety,” 2001. [Online]. Available: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[75] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, “Addressing big data issues in Scientific Data Infrastructure,” in 2013 International Conference on Collaboration Technologies and Systems (CTS), 2013, pp. 48–55.

[76] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster Computing with Working Sets,” HotCloud, vol. 10, p. 95, 2010.

[77] J. Dean and S. Ghemawat, “MapReduce,” Commun. ACM, vol. 51, no. 1, p. 107, Jan. 2008.

[78] “Apache Hadoop YARN.” [Online]. Available: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html. [Accessed: 21-Nov-2017].

[79] “What is Mesos? A distributed systems kernel.” [Online]. Available: http://mesos.apache.org/. [Accessed: 21-Nov-2017].

[80] E. Boyaci and M. Sert, “Video classification based on ConvNet collaboration and feature selection,” in 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017, pp. 1–4.

[81] E. Boyaci and M. Sert, “Feature-level fusion of deep convolutional neural networks for sketch recognition on smartphones,” in 2017 IEEE International Conference on Consumer Electronics (ICCE), 2017, pp. 466–467.

[82] H. Ergun and M. Sert, “Fusing Deep Convolutional Networks for Large Scale Visual Concept Classification,” in 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), 2016, pp. 210–213.

[83] H. Ergun, Y. C. Akyuz, M. Sert, and J. Liu, “Early and Late Level Fusion of Deep Convolutional Neural Networks for Visual Concept Recognition,” Int. J. Semant. Comput., vol. 10, no. 3, pp. 379–397, Sep. 2016.

[84] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014.

[85] “Data Preprocessing.” [Online]. Available: http://ufldl.stanford.edu/wiki/index.php/Data_Preprocessing#Feature_Standardization. [Accessed: 01-Nov-2017].

[86] F. White, “Data Fusion Lexicon,” in Joint Directors of Laboratories, Technical Panel for C3, 1987.

[87] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion,” Proc. IEEE, vol. 85, no. 1, pp. 6–23, 1997.

[88] H. F. Durrant-Whyte, “Sensor Models and Multisensor Integration,” Int. J. Rob. Res., vol. 7, no. 6, pp. 97–113, Dec. 1988.

[89] G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Quénot, “[Invited Paper] TRECVid Semantic Indexing of Video: A 6-Year Retrospective,” ITE Trans. Media Technol. Appl., vol. 4, no. 3, pp. 187–208, 2016.

[90] S. Ayache and G. Quénot, “Video Corpus Annotation Using Active Learning,” in Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 187–198.

[91] “Monitoring and Instrumentation.” [Online]. Available: https://spark.apache.org/docs/latest/monitoring.html. [Accessed: 08-Jan-2018].

[92] “Amazon EC2.” [Online]. Available: https://aws.amazon.com/ec2/. [Accessed: 08-Jan-2018].

[93] Y. Jia et al., “Caffe,” in Proceedings of the ACM International Conference on Multimedia - MM ’14, 2014, pp. 675–678.

[94] “NVIDIA cuDNN.” [Online]. Available: https://developer.nvidia.com/cudnn. [Accessed: 08-Jan-2018].

[95] J. Lyons, “python_speech_features.” [Online]. Available: https://github.com/jameslyons/python_speech_features. [Accessed: 08-Jan-2018].

[96] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2012.

[97] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, Learning Spark Lightning-Fast Big Data Analysis, 1st ed. O’Reilly Media, 2015.